BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–7

Zerone: a ChIP-seq discretizer for multiple replicates with built-in quality control

Pol Cuscó [1,2] and Guillaume Filion [1,2*]

[1]Genome Architecture, Gene Regulation, Stem Cells and Cancer Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is the standard method to investigate chromatin protein composition. As the number of community-available ChIP-seq profiles increases, it becomes more common to use data from different sources, which makes joint analysis challenging. Issues such as lack of reproducibility, heterogeneous quality and conflicts between replicates become evident when comparing data sets, especially when they are produced by different laboratories.

Results: Here we present Zerone, a ChIP-seq discretizer with built-in quality control. Zerone is powered by a Hidden Markov Model with zero-inflated negative multinomial emissions, which allows it to merge several replicates into a single discretized profile. To identify low quality or irreproducible data, we trained a Support Vector Machine and integrated it as part of the discretization process. Zerone detects low quality data with 98% accuracy. Zerone is also flexible, fast and accurate: it discretizes broad or peaky profiles, it runs in a few minutes on modern hardware while using under 800 MB of memory and produces high quality discretizations.

Availability: Zerone is available as a command line tool and as an R package. The C source code and R scripts can be downloaded from https://github.com/gui11aume/zerone.

Contact: guillaume.filion@gmail.com

1 INTRODUCTION

One of the major challenges of biology is to understand how transcription factors and chromatin proteins coordinate transcription, replication and repair. In front of this colossal task, the community invests massive research efforts into collecting protein-genome interaction data. Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-seq) has become the standard method to identify the targets of a transcription factor or a histone modification in a cell population. However, ChIP is not fully understood and artifacts are still discovered more than 10 years after its adoption as a standard (Park *et al.*, 2013; Teytelman *et al.*, 2013). Besides, the constant improvement of sequencing technologies makes analysis of ChIP-seq profiles difficult to standardize. There is thus a continuous need to develop and improve computational tools to analyse ChIP-seq data.

One of the most common analyses performed on ChIP-seq profiles is to discretize the signal, *i.e.* tell for each locus of the genome whether the transcription factor (or other feature) is present or absent. This makes the signal simpler to interpret from the human perspective, it removes part of the experimental noise, it simplifies downstream analyses and it allows comparing or combining profiles of different nature. This raises a challenge at the computational level because discretization has to be carried out uniformly for signals that may be very different. For instance, lamins bind in megabase-scale domains covering 40% of the genome (Guelen *et al.*, 2008), whereas transcription factors may bind as few as 6 bp with a coverage below 1%.

Large consortia such as ENCODE have brought to light a more severe type of issue related to the quality of ChIP-seq data. Conflicts between replicates are common, and sometimes laboratory effects are clearly detectable in the data, even when experimentalists use the same material and follow the same protocol (our unpublished observations). The most popular remedy is to use a metric called IDR (Irreproducible Discovery Rate, Li *et al.*, 2011), which allows to weed out poorly reproducible signal. This approach is a significant step forward, but the IDR is undefined when more than two replicates are available. Besides, keeping only the reproducible ChIP peaks is not always the best option. If one of the replicates is mislabelled, for instance, it is more appropriate to reject the data set than to keep the common ChIP peaks. In summary, how to integrate ChIP-seq data from different sources and with variable qualities is still an open problem.

Here we propose an approach to discretize ChIP-seq data where conflict resolution and quality control are integrated in a tool that we called Zerone. The key idea of Zerone is to combine an arbitrary number of ChIP-seq replicates in a single discretized profile, where conflicts are resolved by maximizing the likelihood of the underlying statistical model. Following discretization, Zerone controls the quality of its output in order to detect potential anomalies, and when applicable rejects the output as a whole. Internally, the first step implements a Hidden Markov Model (HMM) with zero-inflated negative multinomial (ZINM) emissions, and the second implements a Support Vector Machine (SVM) trained using ENCODE ChIP-seq data. HMM-based discretization

[*] to whom correspondence should be addressed

is agnostic about the shape of the signal (broad or peaky) and the ZINM distribution captures the essential features of the read count distribution in ChIP-seq data.

Zerone is designed for large volume pipelines aiming to combine many ChIP-seq profiles with little human intervention. To this end, it is compatible with the standard BED, SAM/BAM, and GEM formats, it produces congruent window-based outputs, and it can process hundreds of experiments per day on average hardware. We benchmarked Zerone against MACS (Zhang *et al.*, 2008), BayesPeak (Spyrou *et al.*, 2009) and JAMM (Ibrahim *et al.*, 2015) on the core task of discretizing ChIP-seq profiles of CTCF, Pol2 and H3K36me3. Our results show that Zerone is competitive in terms of speed and accuracy.

2 METHODS

2.1 Emission model

It is natural to model read counts per genomic window by an unbounded discrete distribution. The Poisson distribution is an obvious candidate, but it is a poor choice because in ChIP-seq data the variance of read counts is usually higher than the mean. The reason is that windows are non homogeneous, which increases the dispersion. More specifically, windows are not equally PCR-prone and not equally mappable. The negative binomial (NB) distribution is thus a better choice because it allows some variation between windows. However, genomes are fraught with repeats, which creates an excess of windows where reads cannot be mapped. Since such windows will always have 0 read count, a natural choice for this distribution is the zero-inflated negative binomial (ZINB), *i.e.* the mixture of a negative binomial distribution and a distribution concentrated at 0.

The ZINB distribution has 3 parameters that can be fitted by maximum likelihood. Zerone uses a custom solver based on the Newton-Raphson method, which converges much faster than the popular routine `zeroinfl` (Zeileis *et al.*, 2008) from the R (R Core Team, 2014) package `pscl` (Jackman, 2015). Fig. 1 shows that the ZINB distribution gives a better fit to ChIP-seq data than Poisson and NB distributions.

The NB distribution can be interpreted as a Gamma-Poisson process, which gives a straightforward extension to a multivariate distribution called the Negative Multinomial (NM) and to the zero-inflated version called Zero-Inflated Negative Multinomial distribution (ZINM, see supplementary material for detail). In this model, windows have an intrinsic ChIP-seq bias due to their sequence composition, mappability and other inherent properties, which gives a baseline variation present in all ChIP-seq experiments performed in the same conditions. The dependency between replicates is explicitly modeled by the ZINM distribution, which gives more accurate probabilities than assuming independence.

2.2 Discretization

Discretization is performed by fitting an HMM with ZINM emissions. The HMM has three states corresponding to “low”, “medium” and “high” abundance of the given chromatin feature. We have observed that in many ChIP-seq profiles, the baseline signal shows piece-wise variations of low amplitude but large size (typically 10-100 Kb). This will sometimes be the dominant signal and a two-state HMM will identify these blocks instead of the targets. Dedicating two states to fit the baseline is a way to make sure that the “high” state corresponds to the targets of the chromatin feature. In what follows, targets are always considered windows with “high” abundance of ChIP-seq reads.

Fitting is performed with the Baum-Welch algorithm (Baum and Petrie, 1966), which is a special case of the EM algorithm (Dempster *et al.*, 1977). Discrete variables take only a small number of distinct values, which allows to save computation time by hashing the observations. With this technique, we need to compute each value of the emission probabilities only once per cycle of the Baum-Welch algorithm. Transition parameters are updated through the forward-backward algorithm, and emission parameters are updated by solving maximum likelihood equations with the Newton-Raphson method (see supplementary material for detail). Approximately 3/4 of the computation time is spent in foward-backward cycles, and 1/4 in updating emission parameters (the time spent computing emission probabilities is insignificant). The algorithm stops when parameters reach a stable value, or after a limit number of cycles (100 by default). The state calls are computed by finding the most likely segmentation given the value of the parameters through the Viterbi algorithm (Viterbi, 1967).

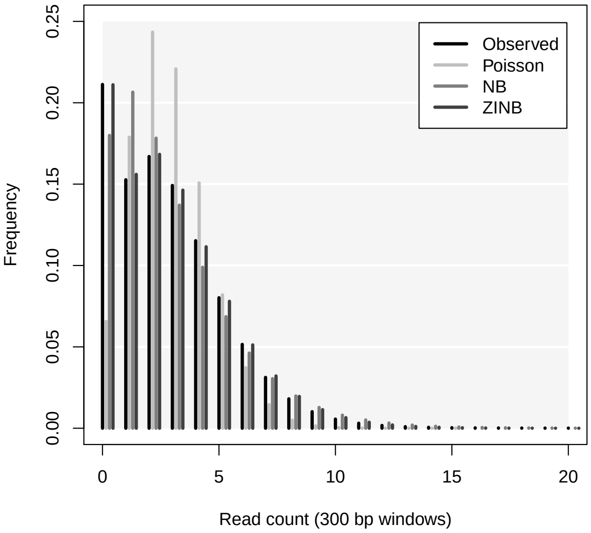

Fig. 1. Using the ZINB distribution to model ChIP-seq data. Reads from the negative control data set XX were mapped on the human genome and pooled in 300 bp windows after removing duplicates. The histogram of the read counts is shown in black (no immunoprecipitation was performed in this experiment, so this variation corresponds to the ‘baseline’). The histograms in gray scales show the maximum likelihood fit of the Poisson, Negative Binomial (NB) and Zero-Inflated Negative Binomial (ZINB) distributions. The fit of the Poisson distribution (light gray) is poor. The NB distribution (medium gray) gives a good fit at the tail, but not for windows with 0 and 1 read. The ZINB distribution (dark gray) gives a good fit over the whole range.

The shape parameter and the mixture ratio of the ZINM distribution are fitted directly from the negative control profiles and they are considered constant throughout the Baum-Welch cycles. Overall, the total number of estimated parameters is $3r + 9$, where r is the number of replicate experiments (excluding mock profiles).

2.3 Classification and training

We used a machine learning strategy to identify discretization failures. We first prepared a high confidence data set where the output of Zerone was labelled positive (success) or negative (failure). We discretized 144 replicated ChIP-seq experiments, together with their respective input control. We labelled the output of the discretization as positive (91 cases) or negative (53 cases), based on visual inspection and on the available literature about the chromatin features. The most common cases of poor data quality in ChIP-seq correspond to low signal-to-noise ratio (*e.g.* when the antibody

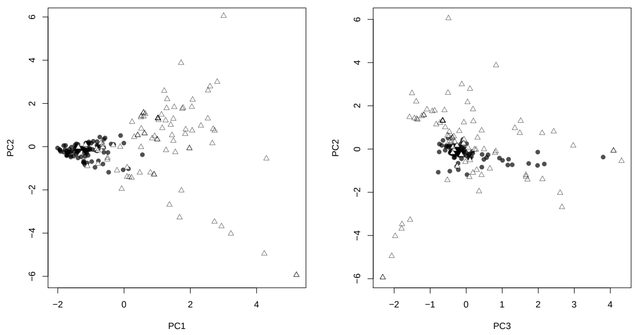

Fig. 2. Principal Component Analysis of the training data set. Each symbol represents a discretization performed by Zerone. The 8 features extracted from each discretization are projected on the first two principal components. Positive examples (black circles) are similar to each other, while negative examples (grey crosses) are different from each other and from positive examples. The two groups overlap, which creates an ambiguous zone where failures and successes are indistinguishable.

is not specific enough), and lack of reproducibility between replicates (*e.g.* when samples are swapped). To cover these cases, we included in the trusted set 38 other cases obtained by discretizing non replicate profiles (*e.g.* CTCF and Pol2) and controls without immuno-precipitation. Thus, a total of 182 cases (91 positive and 91 negative) were used to build a balanced data set.

To identify discretization failures, we used the paremeters of the fitted HMM. Based on the transition matrix, the emission paramaters, the Viterbi Path and the posterior probabilities, we computed 8 features expected to vary with the quality of the discretization, and we used them to train a classifier.

Our first attempts with logistic regression suggested that linear classifiers are unable to properly separate these two classes because they overlap in the feature space (Fig. 2). To obtain more complex, nonlinear separation, we used a Support Vector Machine (SVM, Chang and Lin, 2011; Meyer *et al.*, 2014), as this approach allows nonlinear classification by mapping the training data to a kernel space. Also, SVMs are fast to train and they require only one hyperparameter to be fitted, which were advantages over other approaches such as neural networks.

We trained the SVM with a radial basis function kernel and selected the hyperparameters that maximized the prediction performance on test sets using a 10-fold cross-validation scheme. The prediction accuracy on the trusted set was 98.3%. We then implemented a prediction function in Zerone that used the model trained by the SVM to classify the discretizations and suggest whether they should be accepted or discarded.

2.4 Data sets and preprocessing

We used all the ChIP-seq fastq files produced by the ENCODE Consortium on the human leukemia cell line K562. The data was downloaded from repository http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC in November 2013. We did not use mapped files because they were mapped in different genome assemblies (Szalkowski and Schmid, 2011). Instead, we mapped all the raw reads onto the hg19 assembly of the human genome with GEM (Marco-Sola *et al.*, 2012), using the options `--unique-mapping` and `-q ignore` of gem-mapper version 1.376 beta (and gem-indexer version 1.423 beta). For convenience, we converted the mapped files to SAM format with gem-2-sam version 1.423 beta.

2.5 Benchmark conditions

To compare Zerone to other discretizers, we analysed three different ChIP-seq data sets: CCCTC-binding factor (CTCF), RNA polymerase II (Pol2) and tri-methylated histone H3 at lysine 36 (H3K36me3) — that represent peaky, mixed and broad signals, respectively. Each data set consisted of a mock profile without immunoprecipitation and two replicate ChIP-seq profiles. The ENCODE accession numbers for these three data sets are ENCSR000DWE, ENCSR000DWB and ENCSR000EHP respectively.

We tested MACS `callpeak` version 2.1.0.20140616, BayesPeak version 1.20.0 and JAMM version 1.0.7rev1. All tests were performed on an 8-core Intel Xeon E5606 machine with 48 GB of DDR3-RAM at 1333 MHz. All programs were run on a single core with the default options.

CTCF motifs were obtained from the JASPAR database version 5.0_ALPHA (Mathelier *et al.*, 2014).

3 RESULTS

We benchmarked Zerone against three other ChIP-seq discretizers. We included MACS as the standard method for ChIP-seq peak calling, BayesPeak because it is powered by an HMM with ZINB emissions similar to the model implemented in Zerone, and JAMM because, as Zerone, it can perform joint discretization of experimental replicates.

3.1 Speed and memory footprint

We compared the running times of the different programs on discretizing three data sets of similar size that represent the three major types of ChIP-seq signal usually observed. The CTCF signal consists of sharp peaks at the transcription factor binding site, the H3K36me3 signal consists of broad domains, and the Pol2 signal consists of peaks at promoters and potentially broad domains on transcribed genes.

The results were similar between experiments, and Zerone was consistenty the fastest tool, with a running time around 5 minutes (Fig. 3, top row). The advantage is only marginal over MACS, which ran for around 10 minutes, but it is substantial over BayesPeak and JAMM, which ran in over 9 hours. The results for peak memory usage were more variable between experiments (Fig. 3, bottom row). MACS achieved the best performance with a memory footprint around 0.5 GB, followed by Zerone around 0.7 GB. BayesPeak and JAMM each used more than 1.5 GB.

The benchmark is partly confounded by the fact that BayesPeak and MACS discretize a single input per run, whereas Zerone and JAMM discretize multiple inputs simultaneously. This makes a difference for pipelines where all files have to be processed in parallel with the minimum amount of resources. In our benchmark, Zerone used around 40% more memory than MACS, but it processed twice as many files. For the same amount of available memory, a Zerone pipeline would run two times faster than a MACS pipeline, and with only half the computer power.

3.2 Accuracy

The purpose of discretizers is to identify the targets of a transcription factor or a histone mark, *i.e.* the sites of the genome where it is present. Intuitively, good discretizers capture a large fraction of the ChIP-seq signal within few targets. The number of targets and the amount of reads they represent are therefore critical characteristics of a discretization. Unfortunately, there is no gold standard to estimate the trade-off between false positives and false negatives in ChIP-seq experiments, and thus there is no objective way to rank discretizers. However, we can compare them with a partial order, as explained below.

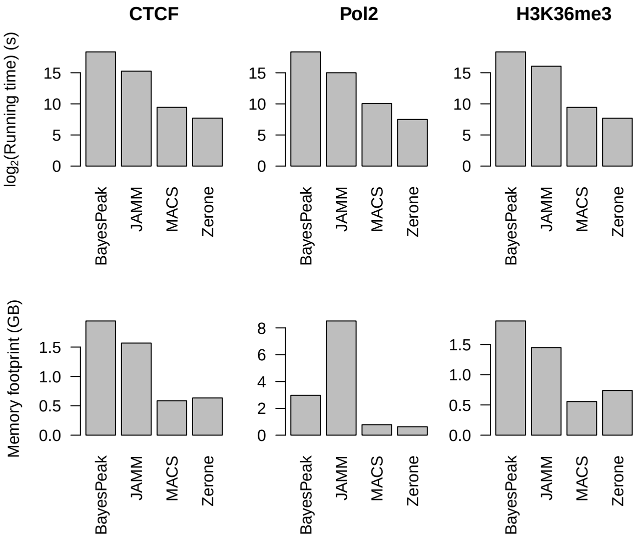

Fig. 3. Running times and peak memory footprint of the discretizers on the three ChIP-seq data sets. For programs that only allow single-profile discretization (*i.e.* BayesPeak and MACS), mean values (not the sum) are shown. Note the logarithmic scale in the running times.

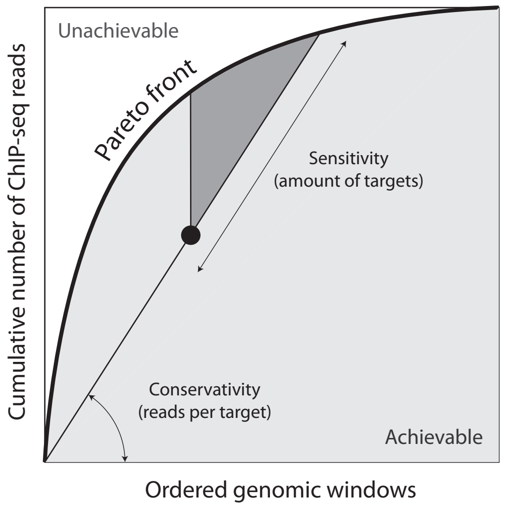

Fig. 4. Graphical representation of discretizations. Genomic windows of the ChIP-seq profile are ordered by decreasing amount of reads on the x axis, and the cumulative amount of reads is plotted on the y axis. This line forms a Pareto front representing the maximum number of reads for a given number of windows. Discretizations are represented as a single point on this plane (black disc) whose coordinates are the number of targets and the total number of reads in the targets. The dark triangle represents discretizations that are more conservative and discover more targets.

Table 1. Performance on the CTCF motifs data set. The table lists the total number of peaks found by the different programs, how many of those peaks contain at least one CTCF motif, and the associated precision, recall and F_1 score relative to the CTCF motif data set.

Software	Total	Motif	Precision	Recall	F_1 score
BayesPeak$^{(1)}$	45,316	25,228	0.56	0.29	0.39
BayesPeak$^{(2)}$	45,154	23,428	0.52	0.27	0.36
JAMM	264,410	31,709	0.12	0.37	0.18
MACS$^{(1)}$	48,358	26,449	0.55	0.31	0.39
MACS$^{(2)}$	41,030	23,542	0.57	0.27	0.37
Zerone	50,792	30,972	0.61	0.36	0.45

The numbers in parentheses indicate the results on the two replicates by separate.

When arranging genomic windows from high to low amount of ChIP-seq reads, the cumulative number of reads forms a Pareto front. It represents the largest amount of reads that can be captured by the given amount of targets, or alternatively the smallest number of targets that can catpure the given amount of reads (Fig. 4). A discretization can be represented as a point on the xy plane. By construction, no discretization can lie on the left of the Pareto front, and those on the front represent an optimum. Others are suboptimal, since fewer windows can capture more targets.

This representation reveals that discretizing the CTCF profile yields similar outputs regardless of the software (Fig. 5, left panel). On the other hand, discretizing the Pol2 or the H3K36me3 profile yields very distinct outputs (middle and right panels). In all the cases, Zerone produces the discretization capturing the most reads. For CTCF and Pol2, it lies on the Pareto front. For H3K36me3, it lies somewhat off the Pareto front, but at a sensible location. As detailed below, H3K36me3 is deposited on transcribed genes (Pokholok *et al.*, 2005; Kimura, 2013) so the coverage is expected to be higher than for transcription factors.

This representation shows that Zerone produces discretizations that are sensitive and adapted to the profile being discretized.

3.2.1 Identification of CTCF binding sites. CTCF binds a 20 bp consensus sequence that is highly conserved in vertebrates. In humans, nearly 80% of the CTCF binding sites contain the consensus motif (Kim *et al.*, 2007). In order to determine the capacity of the different programs to call peaks of CTCF binding, we compared the discretized profiles against a data set of CTCF binding motifs. We used FIMO (Grant *et al.*, 2011) from the MEME suite version 4.10.1 (Bailey *et al.*, 2009) to identify and map CTCF motifs in the human genome. We obtained a reference data set containing 85,690 CTCF motifs.

Table 1 shows that Zerone outperforms the other tools in terms of the F_1 score (the harmonic mean between precision and recall). It has the second higest recall score, close to that of JAMM, and it achieves the highest precision.

In summary, Zerone achieves a good balance between sensitivity and specificity for transcription factor profiles, as was suggested by the left panel of Fig. 5.

3.2.2 Pol2 binding around transcription start sites. To compare the behavior of the discretizers on the Pol2 profile, we determined the extent to which the enriched regions contained Transcription Start Sites (TSS), as Pol2 is expected to bind near annotated TSS. Data about TSS positioning was extracted from the knownGene

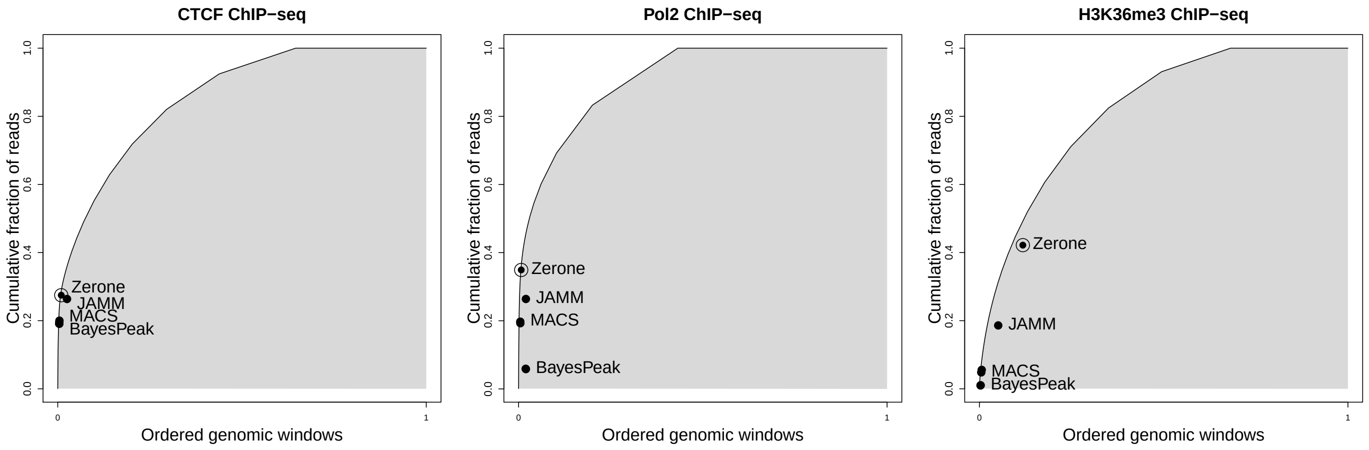

Fig. 5. Characteristics of the discretizations for different programs. The representation is obtained as shown on Fig. 4. When discretizing the CTCF and the Pol2 profile (left and middle), Zerone produces discretizations that are very close to the Pareto front. In the case of H3K36me3, the discretization is off the Pareto front, but it is more sensitive and more sensible than the others.

Table 2. Performance on the Pol2 data set. The table lists the total number of peaks found by the programs, how many of those peaks contain at least one Transription Start Site (TSS), and the associated precision, recall and F_1 score.

Software	Total	TSS	Precision	Recall	F_1 score
BayesPeak$^{(1)}$	208,809	2,722	0.01	0.06	0.03
BayesPeak$^{(2)}$	203,438	2,603	0.01	0.05	0.03
JAMM	209,882	6,761	0.03	0.21	0.11
MACS$^{(1)}$	51,313	6,845	0.13	0.22	0.27
MACS$^{(2)}$	49,034	6,546	0.13	0.21	0.26
Zerone	23,976	6,926	0.29	0.24	0.36

The numbers in parentheses indicate the results on the two replicates by separate.

table of the UCSC Genes annotation for the hg19 genome (Karolchik *et al.*, 2004).

Table 2 shows that, as in the previous case, Zerone achieves the highest F_1 score. Here it achieves both better precision and better recall than the other discretizers, with a very neat advantage in precision. These results confirm that the performance of Zerone is outstanding on this data set, as suggested by the middle panel of Fig. 5.

3.2.3 H3K36me3-enriched domains. There is no consensus sequence to determine the location of histone modifications. However, it is known that the bodies of active genes are enriched in H3K36me3 (Pokholok *et al.*, 2005; Kimura, 2013). Therefore, the genes that contain peaks or windows determined as enriched in H3K36me3 by the different discretizers should be more expressed than the background.

We benchmarked the quality of the discretization with expression data obtained in the same cell. We used the number of RNA reads as a response variable and computed the amount of variance explained by the binarized profile of H3K36me3. The discretization produced by Zerone is the best predictor of expression (Fig. 6). It is also the one with highest coverage (Fig. 5, right panel), which shows that the increased number of targets does not come at the cost of accuracy. On the contrary, the high coverage of H3K36me3 is confirmed by expression data.

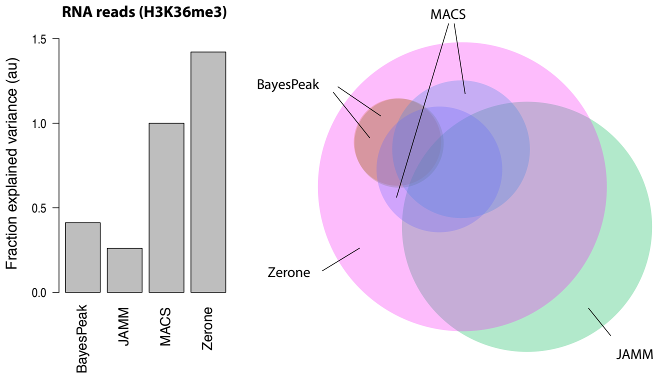

Fig. 6. Left panel: quality of the H3K36me3 discretization. Each bar represents the relative fraction of variance in RNA reads explained by the discretization of H3K36me3. The value for MACS was set to 1.0 and all other values were scaled accordingly. Right panel: Overlap between H3K36me3 targets. Each circle represents the H3K36me3 targets identified by a program. The size of the circles is proportional to the coverage of the targets, and their overlap approximates the amount of targets shared by the programs. Note the almost complete overlap between the BayesPeak discretizations on the two replicates.

A Venn diagram gives a graphical overview of the relationships between the discretizations (Fig. 6). Zerone finds most of the targets detected by the other discretizers (except JAMM), while discovering enriched windows not found by the others.

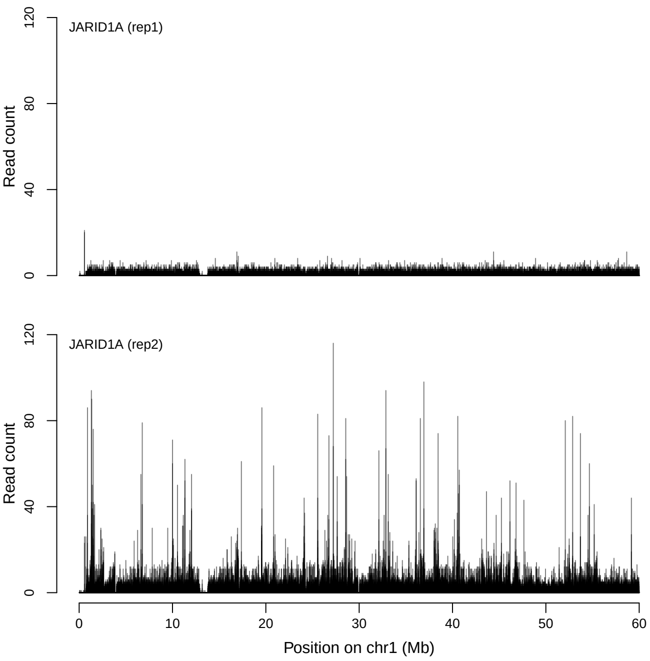

Fig. 7. ChIP-seq profiles of JARID1A in H1 ES cells (300 bp windows). The first replicate is not similar to the second, and it does not contain any target.

3.3 Automatic quality control

The most novel feature of Zerone is an embedded automatic quality control step taking place after the discretization. This not only ensures that the discretization is sensible, but also that the replicates are similar to each other and that ChIP-seq profiles are not too similar to the mock controls. Our approach is based on the idea that estimating distributional parameters from overly noisy or divergent profiles should give a signature that can be picked up by a specially trained classifier.

We identified 8 summary statistics that characterize the discretization (including the transition parameters of the HMM, the mean number of targets and their mean posterior probability) and trained an SVM to recognize failures. We thus obtained a classifier able to identify failed discretization attempts with 98% accuracy (*c.f.* section 2.3).

Computer-assisted quality control is essential for high throughput pipelines and for cases with no prior knowledge. To showcase this feature, we decided to discretize ENCODE ChIP-seq profiles obtained in human H1 ES cells. Upon discretizing the lysine-demethylase JARID1A, the quality control module of Zerone reported a failure.

Further investigation immediately revealed the nature of the issue. In one of the replicates, the signal is lacking entirely, as if the immuno-precipitation was non specific (Fig. 7). Once aware of the issue, users can handle it properly (for instance by discarding the protein or by working with a single replicate). Such cases are easy to miss in automatic pipelines, and automatic quality control can be of great help.

4 DISCUSSION AND CONCLUSIONS

Zerone was developed ground up for scalibility and throughput. Part of the speed is due to hashing methods that dramatically cut down the computation time during the Baum-Welch cycles. Zerone also rests on sound statistical bases. Theoretical arguments and experimental observations suggest that the Zero-Inflated Negative Multinomial distribution is appropriate to model ChIP-seq data (Fig. 1). This gives Zerone good specificity and sensitivity for very different profiles.

Here we introduced a way to compare discretizers with an intuitive graphical representation (Fig. 4). On this plane, the coordinates of a discretization indicate the number of targets (or occupancy) and the amount of reads captured by these targets. The Pareto front captures the inherent trade-off between sensitivity and specificity in the problem of discretizing ChIP-seq profiles. Points on this line that are close to the bottom-left corner represent discretizations with high amount of reads per target (most specific) and points that are close to the top-right corner represent discretizations with many targets (most sensitive). The Pareto front also highlights an unachievable region whose shape depends on the structure of the signal, *i.e.* on the feature being disretized (Fig. 4). One of the challenges of discretizing ChIP-seq profiles is to find algorithms that perform well in all the cases.

In practice, the specificity of a discretizer is unknown because the biological truth remains hidden. However, we can decide whether a discretizer is more or less conservative than another by measuring the amount of reads per target. This characteristic may be a matter of choice, and is usually tacit in the case of ChIP-seq discretizers. Out of two equally conservative discretizers, one may be more sensitive, *i.e.* discover more targets. The merit of the representation introduced here is to highlight these characteristics and to guide users when choosing the most appropriate tool for their need.

This representation naturally suggests a naive approach to discretize ChIP-seq profiles. Indeed, one could sort the genomic windows by decreasing amount of ChIP-seq reads and declare ‘target’ any window above a chosen threshold. While this method would only produce discretizations on the Pareto front, adjusting the threshold to the conditions would be challenging for lack of an underlying model. This is one of the major strengths of Zerone: the statistical model automatically adjusts conservativity and sensitivity in a sensible way.

Finally, Zerone proposes an original solution to the problem of data heterogeneity. First, the statistical model will be fitted in order to harmonize the replicates and solve conflicts by maximum likelihood. Second, an automatic quality control is performed after the discretization. The principle of this step is somewhat similar to anomaly detection. As examples of trusted ChIP-seq discretizations will increase, the classifier can be refined to reach greater accuracy.

ACKNOWLEDGEMENT

Funding: The fellowship of P.C. is partly financed by the Spanish Ministry of Economy and Competitiveness (State Training Subprogram: predoctoral fellowships for the training of PhD students (FPI) 2013). We acknowledge support of the Spanish Ministry of Economy and Competitiveness, ‘Centro de Excelencia Severo Ochoa 2013-2017’, SEV-2012-0208.

REFERENCES

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**(Web Server issue), W202–8.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, **37**(6), 1554–1563.

Chang, C.-c. and Lin, C.-j. (2011). LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2**, 1–39.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, **39**(1), 1–38.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7), 1017–8.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, **453**(7197), 948–51.

Ibrahim, M. M., Lacadie, S. A., and Ohler, U. (2015). JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics*, **31**(1), 48–55.

Jackman, S. (2015). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.4.9.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic acids research*, **32**(Database issue), D493—-6.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**(6), 1231–45.

Kimura, H. (2013). Histone modifications for human epigenome analysis. *J. Hum. Genet.*, **58**(7), 439–45.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**(3), 1752–1779.

Marco-Sola, S., Sammeth, M., Guig, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**(12), 1185–8.

Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., yu Chen, C., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**(Database issue), D142–7.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-4.

Park, D., Lee, Y., Bhupindersingh, G., and Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE*, **8**(12), e83506.

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K., and Young, R. A. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, **122**(4), 517–27.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Spyrou, C., Stark, R., Lynch, A. G., and Tavar, S. (2009). BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics*, **10**, 299.

Szalkowski, A. M. and Schmid, C. D. (2011). Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *Brief. Bioinformatics*, **12**(6), 626–33.

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**(46), 18602–7.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, **13**(2), 260–269.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, **27**(8).

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**(9), R137.