Pyicos: A versatile toolkit for the analysis of high-throughput sequencing data

Sonja Althammer1,4, Juan González-Vallinas1,4, Cecilia Ballaré2, Miguel Beato1,2, Eduardo Eyras1,3,*

1Universitat Pompeu Fabra. Dr. Aiguader 88, E08003, Barcelona, Spain
2Centre for Genomic Regulation (CRG). Dr. Aiguader 88, E08003 Barcelona, Spain
3Catalan Institution for Research and Advanced Studies (ICREA). Passeig Lluís Companys 23, E08010, Barcelona, Spain

† These authors contributed equally
* Corresponding author

ABSTRACT

Motivation: High-Throughput Sequencing (HTS) has revolutionized gene regulation studies and is now fundamental for the detection of protein-DNA and protein-RNA binding, as well as for measuring RNA expression. With increasing variety and sequencing-depth of HTS datasets, the need for more flexible and memory-efficient tools to analyse them is growing.

Results: We describe Pyicos, a powerful toolkit for the analysis of mapped reads from diverse HTS experiments: ChIP-Seq, either punctuated or broad signals, CLIP-Seq, and RNA-Seq. We prove the effectiveness of Pyicos to select for significant signals and show that its accuracy is comparable and sometimes superior to that of methods specifically designed for each particular type of experiment. Pyicos facilitates the analysis of a variety of HTS datatypes through its flexibility and memory efficiency, providing a useful framework for data integration into models of regulatory genomics.

Availability: Open source software, with tutorials and protocol-files, is available at http://regulatorygenomics.upf.edu/pyicos or as a Galaxy server at http://regulatorygenomics.upf.edu/galaxy

Contact: eduardo.eyras@upf.edu

Supplementary Information: Supplemental data are available at Bioinformatics online.

1 INTRODUCTION

Gene expression is regulated through a complex network of protein interactions with RNA and DNA. Recent advances in high-throughput sequencing (HTS) technologies provide a very effective way to obtain information about these interactions at genome wide level at reasonable cost. ChIP-Seq (Chromatin Immunoprecipitation followed by HTS) has become the preferred method for current analysis of transcriptional regulation in vivo, since it provides genome-wide coverage at a high sensitivity and signal-to-noise ratio (Robertson et al., 2007; Johnson et al., 2007). However, to unravel the complete gene expression network, it is also necessary to systematically identify protein–RNA interactions in cells. HITS-CLIP or CLIP-Seq (Cross-Linking ImmunoPrecipitation followed by HTS) provides a powerful mean to obtain global information about direct protein-RNA interactions in cells (Licatalsi et al., 2008; Xue et al., 2009). Additionally, the RNA-Seq experiments provide a way to measure RNA levels at a higher sensitivity and coverage than microarray experiments. Despite the inherent biases, RNA-Seq can overcome some of the array limitations and provides the possibility to measure alternative splicing events (Pan et al., 2008; E. T. Wang, Sandberg, et al., 2008) or expression levels (Pepke et al., 2009; Trapnell et al., 2010), to detect single nucleotide polymorphisms (Xu Wang, Q. Sun, et al., 2008) and to discover novel genes (Khalil et al., 2009). Short read sequencing platforms such as Illumina are very suitable for the study of protein-DNA and protein-RNA interactions, since they can produce enough data to perform accurate quantitative analysis. Furthermore the resulting reads are of sufficient length to be mapped accurately to a reference sequence but short enough to define the interaction sites with enough precision. However, not all interactions can be detected at the same coverage. For instance, transcription factors produce generally a clearly localized, or punctuated, signal (Park, 2009), whereas histone marks or RNA polymerase II (RNAPII) produce broad signals, covering more extended regions (Park, 2009). Most of the tools published for the analysis of ChIP-Seq data focus on punctuated signals (Zhang et al., 2008; Fejes et al., 2008; Nix et al., 2008; Park, 2009; Pepke et al., 2009), with only few methods devoted to analyse broad signals (Zang et al., 2009; Shin et al., 2009). However, these methods are not readily applicable to other HTS datatypes, like CLIP-Seq or RNA-Seq, making it necessary to incorporate additional tools to perform gene regulation studies, which generally involve multiple HTS experiments. The increase in variety and sequencing-depth of such experiments, as well as the growing need for memory-efficient tools that can be applied to various datatypes, motivated us to develop Pyicos.

In this paper, we describe Pyicos, a versatile toolkit operating on short HTS reads that have already been mapped to a reference
Pyicos provides different protocols to define significant signals from ChIP-Seq (both punctuated and broad), CLIP-Seq and RNA-Seq data; and all operations can be applied independently or combined in a protocol file, which makes Pyicos useful for the integrative analysis of a wide variety of HTS data. Moreover, the software has a modular architecture that allows re-usability in other Python applications. Pyicos thus offers a useful framework to facilitate the analysis of a variety of HTS datatypes, providing the basis for data integration into gene regulation studies.

2 METHODS

Pyicos is written in Python and is available under the GNU General Public License. It can operate from the command line or as a software library within Python programs (tested on GNU/Linux Fedora 12 and MAC OSX). All operations can be used independently and the parameters can be set by the user; thus providing as much flexibility as possible. Additionally, Pyicos can be run using protocols files (Supplementary Figure S1 and Supplementary Material), which allows the user to apply them directly for standard analyses or to design customized workflows.

Pyicos has its own compressed format for peaks as explained in the Supplementary Material. This is also used as an internal representation for read clusters, and can be used as input and output formats. Furthermore, Pyicos can handle other various formats, providing converter capabilities. It can read eland, SAM, BED, and wiggle (fixed and variable step); and write to SAM, BED, wiggle (fixed and variable step). Some formats are obligatory for some of the operations, which is specified in the corresponding command help.

2.1 Callpeaks protocol

The callpeaks protocol is applied to punctuated ChIP-Seq data and starts by optionally removing duplicated reads, i.e. redundant reads in terms of position and strand in a genome. Next, all reads that fall into a set of regions specified by the user, such as satellites and pericentromeric regions, are removed, as they generally lead to false positives. Reads are extended to the expected initial fragment size, given as input. Alternatively, Pyicos can also estimate this extension using a shift correlation analysis (Zhang et al., 2008). Next, reads are grouped into read clusters according to genomic overlap. To follow standard nomenclature, for punctuated data we will refer to read clusters as peaks. When a control experiment is available, callpeaks normalizes the sample with respect to control and subtracts the profile of the control from that of the sample with nucleotide (nt) resolution (Supplementary Material). At this stage, callpeaks removes two possible artefacts: peaks that are shorter than a certain length (100nt by default), which may be produced by the subtraction step; and block-like peaks, which may occur when duplicates are kept. This removal is optional to the user. Pyicos also includes a split operation, which allows exploring the possibility that some peaks may describe multiple binding sites. Finally, callpeaks calculates the significance for a peak to correspond to a real binding site using Poisson analysis applied to three different parameters: the peak height, the number of reads per peak, and the number of reads per nucleotide. For each chromosome, a p-value is calculated independently for each of these parameters. We also define a peak-score as a combination of the Poisson p-value and the peak parameter P: Peak-score = (P - log_{10}(p-value))/2, for P equals to either the peak height or the peak read-count.

2.2 Benchmarking of callpeaks

We compared Pyicos with MACS (Zhang et al., 2008) version 1.3.7.1, FindPeaks (Fejes et al., 2008) version 3.3 and USeq (Nix et al., 2008) version 7.2, using ChIP-Seq datasets for four different factors: the insulator (CTCF) in K562 cells (ENCODE Consortium, 2011), the CCAAT/enhancer-binding protein alpha (CEBPA) in liver cells (Schmidt et al., 2010), the neuron-restrictive silencer factor (NRSF) in Jurkat T cells (Johnson, Mortazavi, et al., 2007) and the progesterone receptor (PR) in T47D cells (Vicent et al., 2011). We have used various measures to evaluate the quality of peak definition on the punctuated ChIP-Seq data. Two of the descriptors of the peak definition used are the peak-length and the distance between the peak centre and summit, i.e. the position with the highest read pileup. We further compared the rankings of the peaks provided by each method: the peak-score (Pyicos), the peak-height (FindPeaks), the -10-log_{10}(p-value) (MACS), and the ranking provided by USeq.

For the accuracy evaluation, we also used the 83 positive and 30 negative regions validated by ChIP-qPCR for NRSF (Mortazavi et al., 2006) as done in (Johnson, Mortazavi, et al., 2007). Since peaks differ a lot in their extensions (Supplementary Figure S2), we took a region of 100nt centred on the summit as the predicted binding region. We classified the peaks from each method as True Positives (TP) and False Positives depending on whether the 100 nt long region centred on the summit overlapped a positive or negative validated region, respectively. The True Positive Rate (TPR) was calculated as TP over the total number of positive regions and the False Positive Rate (FPR) was calculated as FP over the total number of negative regions. Receiver Operating Characteristic (ROC) curves were calculated using incremental subsets of peaks along the ranked peaks from each method, with increasing subset sizes of 500 peaks.

An additional measure of the quality of the predictions is the fraction of peaks associated to the expected motif (Ji et al., 2008; Zhang et al., 2008). We thus measured the fraction of peaks with motif occurrence in the sequence underlying the summit, considering 100 nt for NRSF and PR peaks and 200 nt for CEBPA and CTCF peaks. For the motif analysis, we scanned the selected sequences from the peaks using the matrices for NRSF, CTCF, CEBPA and the three available matrices for PR from TRANSFAC (Knüppel et al., 1994) (Accession numbers: M00256, M01200, M00116, M00954, M00957, M00960) using a custom script. We used as core similarity cut-off 0.99 and as total similarity cut-off 0.85. As the summit of a peak is associated more strongly with the binding motif than the peak-centre (Supplementary Figure S3), we considered it a suitable reference position. Additionally, we used as fourth descriptor the spatial resolution, i.e. the distance from the summit position to the centre of the detected motif.

Finally, for all four ChIP-Seq datasets we considered the Pyicos rankings using both, the peak height and read-count scores. For further analyses we selected the score that gives the best motif content in top scoring peaks: the peak height score for PR and CEBPA, and the read-count score for CTCF and NRSF.

2.3 Enrichment protocol

Pyicos incorporates a method to calculate the significance of the enrichment of the signal between two samples based on the comparison with the distribution of enrichment values on the same regions for experimental or theoretical replicates, similarly to the methods MATR and MARS from DESeq (L. Wang et al., 2010). Using subsets of about 5% of neighbouring data points along the axis of average densities A, the enrichment values M calculated for genes between RNA-Seq samples for two liver replicates from (Marioni et al., 2008) follow a normal distribution (Supplementary Figure

Downloaded from Bioinformatics.oxfordjournals.org by guest on October 18, 2011
Accordingly, Pyicos uses a sliding window along the A axis and calculates a Z-Score for the enrichment M of each region from the comparison of the replicates (details in Supplementary Material). The region is assigned to a background window according to proximity: the region is compared to the values of the window which centre is closest to the region in terms of the value of A. If no replicates are provided, theoretical replicates are created by randomly rearranging the reads into two subsets, taking into account the relative sizes of the original samples. We integrated the TMM normalization (Robinson and Oshlack, 2010) in Pyicos EA. The calculation of the TMM factor on our read-count lists yielded a value of 0.68 for liver versus kidney and a value of 1.01 for the comparison of the two liver replicates, which agrees with the results reported in (Robinson and Oshlack, 2010).

Pyicos EA can have as input read-counts or Reads Per Kilobase per Million reads (RPKM) (Mortazavi et al., 2008); or alternatively, directly the BED files with regions and read coordinates, overcoming the need of an additional tool to count reads. Additionally, Pyicos can calculate enriched regions de novo by scanning the samples provided, using two configurable parameters, the proximity between two reads to be taken as part of the same region, and the minimum number of reads in the region to be considered.

2.4 Benchmarking of enrichment analysis

To assess the accuracy of Pyicos in the prediction of differentially expressed (DE) genes, we compared it to three other methods, DESeq (L. Wang et al., 2010), DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010); using published datasets for RNA-Seq and equivalent microarray experiments for liver and kidney samples (Marioni et al., 2008). First, we mapped the microarray results to the Ensembl annotation (Flicek et al., 2011). Then, using the microarray data, we defined a benchmarking set composed of DE and non-DE genes. DE genes were defined to have an FDR of less than 0.001 and an absolute log2-fold change of more than 0.5 (Marioni et al., 2008), regardless of whether they are up- or down-regulated. Non-DE genes were defined to have a FDR of more than 0.01 and an absolute log2-fold change of at most 0.5. This benchmarking set is composed of 6700 DE genes and 7060 non-DE genes. To predict DE genes we considered for each Ensembl locus the mean of the values for read-counts or RPKM for the corresponding Ensembl transcripts in the locus. The RPKM and read-count per transcript were calculated with reads falling within the exons of each transcript. A pseudocount of one read per transcript was added to be able to operate with transcripts whose exons did not have any reads in any of the samples. We estimated the True Positives (TPs) and False Positives (FPs) as the number of predicted DE (enriched or depleted) genes that were annotated as DE or non-DE genes in the benchmarking set, respectively. ROC curves were calculated by considering increasing absolute values of the Z-score or -log10(p-value) for the predictions for each method. We also calculated precision-recall curves along Z-score thresholds, where precision is the ratio of TP over the number of predicted cases, and recall is the same as the TPR.

2.5 Enrichment on ChIP-Seq data

We performed enrichment analysis on broad ChIP-Seq signals using ENCODE data for H3K36me3 and RNAPII, and correlated the results to the enrichment of RNA-Seq data, for the K562 and NHEK cells (ENCODE Consortium, 2011). The calculation of significantly enriched or depleted regions was based on RPKM values. Replicates 1 and 2 from K562 and replica 1 from NHEK were used to calculate enrichment on three different types of regions from the Ensembl annotation (Flicek et al., 2011): for H3K36me3 we used the transcript-body, spanning from the transcription start site (TSS) to the transcription termination site (TTS); for RNAPII we used a window of 4kb around the TSS; and the transcript exons for RNA-Seq, as before. Significant regions were defined as those with an absolute Z-score higher than 10. We calculated the Pearson correlation coefficient of Z-Scores from significant regions between H3K36me3 and RNA-Seq, as well as between RNAPII and RNA-Seq.

2.6 Memory usage and running time benchmarking

To test the memory usage and running time of the callpeaks protocol, we used ChIP-Seq data for the human transcription factor CEBPA (Schmidt et al., 2010). To simulate files of different sizes, we first pooled together all available reads and took random subsets with an increasing size step of 3 million reads, up to 30 million reads, separately for sample and control. In each run we sub-sampled two random sets of equal size from the ChIP-Seq reads and from control, using steps of 3 million reads.

In order to test the performance of Pyicos enrichment analysis we compared it to DESeq and to a combination of BEDTools (Quinlan and Hall, 2010) with edgeR and DESeq, as these two programs do not accept as input BED files with the regions of interest and the positions of the mapped reads from the samples to be compared. In this way, all four calculations start with the same input: BED files for the samples and the region file. We used the same CEBPA data mentioned before and we calculated the enrichment in the region of 2 kilobases (kb) upstream of the TSS for RefSeq annotated genes. This time we sampled twice per run on the experiment, in order to get a simulated replica, and once for the control. We sampled reads increasing by 3 million reads in each run.

3 RESULTS

3.1 Analysis of punctuated ChIP-Seq data

The problem of peak detection entails two challenges: first, to determine significant peaks by distinguishing between real and false binding signal; and second, once we have selected the candidate peaks, we need to properly determine the site of interaction. In what follows, we use various methods to establish the accuracy and quality of peak prediction for Pyicos and three other published methods for punctuated ChIP-Seq analysis. But first, we assess the effectiveness of Pyicos operations for peak definition using ChIP-Seq datasets for NRSF, PR, CTCF and CEBPA.

ChIP-Seq analysis often starts by removing duplicated reads, i.e. redundant reads in terms of position and strand in a genome. This has been suggested to eliminate amplification biases (Zhang et al., 2008). We observe that keeping duplicates does not produce any improvement on the four sets and that, in fact, the highest fraction of peaks with motifs is achieved when all duplicates are removed (Supplementary Figure S5). An additional way to remove experimental biases, which may introduce false positives, is the use of a control. When this is available, callpeaks subtracts the control from the sample after normalization (see Supplementary Methods). To test whether Pyicos subtraction improves peak detection we generated peaks with and without the subtraction of the control. We observe a great improvement in peak detection for CEBPA (Figure 1a), CTCF and PR (Supplementary Figures S6), while we only saw a slight increase for the top peaks for NRSF (Figure 1b). Interestingly, the subtraction of the control results into almost no difference in the spatial resolution (Supplementary Figure S7).

To explore the possibility that some regions of ChIP-Seq signal may describe multiple binding sites, Pyicos includes the option to split peaks (see Supplementary Material). We found that splitting peaks showed improvement only on the CTCF dataset, while spatial resolution was maintained (Supplementary Figure S8). The protocol callpeaks incorporates a Poisson test on the retained peaks to select those that are more likely to be functional (see Methods). Although the Poisson distribution seems not to fit well the background biases of ChIP-Seq experiments (Zhang et al., 2008), our analyses indicate that by subtracting the control we remove these
biases and Poisson analysis can be applied. Each predicted peak is given a peak-score and a p-value (see Methods). The peak-score gives a ranking for all peaks, whereas the p-value is used to select a subset of candidate peaks with certain significance. In order to show that the peak-score produces an appropriate ranking, we calculated the fraction of peaks containing the expected motif along ranked peaks. We observe that for increasing cut-offs of the p-value, the fraction of peaks with motifs decreases with the peak-score (Supplementary Figure 9). The selection of candidate peaks by p-value is therefore consistent with the expected association of significant peaks to functional motifs. The peak-score can be calculated using the peak height or read-count. On the one hand, we found no differences in peak quality on NRSF and slight differences on CTCF and CEBPA, using peak-height or read-count. On the other hand, the peak-height score for PR results into an increased fraction of peaks with motifs relative to the read-count score (Supplementary Figure S10).

In order to further assess the quality of Pyicos peak calling, we compared Pyicos with three other methods: MACS, FindPeaks and USeq. We evaluated the peak definition of the four methods in terms of motif content (Methods). They all show a similar trend along the ranking, with Pyicos and MACS showing higher densities for all ranking positions (Figure 1c and 1d and Supplementary Figure S11). Moreover, calculating the spatial resolution (Methods), the four methods show the largest agreement for the PR dataset, whereas USeq shows the lowest spatial resolution on NRSF, CEBPA and CTCF (Supplementary Figure S12).

We also evaluated the accuracy of peak definition measuring the overlap of the selected peaks with a benchmarking set of positive and negative NRSF binding regions validated by ChIP-qPCR (Methods). All four methods show high agreement for their top 3,000 NRSF peaks in terms of overlap (Supplementary Table S1) and in terms of the pairwise correlations (Supplementary Table S2); hence we expect they would achieve similar accuracies. Indeed, the four methods perform similarly well with an area under the ROC curve (AUC) between 0.90 and 0.91 (Table 1). We also compared the accuracy of peak definition of the selected peaks by calculating the distances between summit and peak-centre as well as the lengths of the peaks. We found that USeq, which produces the shortest peaks (Supplementary Figure S2), achieves the shortest distance between summit and peak-centre, closely followed by Pyicos (Supplementary Figure S13).

Finally, in order to test our software performance, we used ChIP-Seq data for the human transcription factor CEBPA (Schmidt et al., 2010) (Methods). Running all four methods with conditions as similar as possible (see Supplementary Material), we observed that Pyicos is more efficient in memory usage than the other three methods (Figure 1e) and stays competitive in running time (Supplementary Figure S14).

### 3.3 Enrichment analysis on RNA-Seq data

Methods to measure differential gene expression from RNA-Seq are generally based on enrichment analysis (EA) between samples (L. Wang et al., 2010; Anders and Huber, 2010; Oshlack et al., 2010; Robinson and Oshlack, 2010; Bullard et al., 2010). Pyicos incorporates a method to detect regions of significant enrichment between two samples based on the comparison of the observed enrichment values with those measured between two replicas (Methods). In order to establish the accuracy of Pyicos for detecting DE genes, we compared it first to the methods MATR (with replicates) and MARS (without replicates) from DEGseq (L. Wang et al., 2010), using RNA-Seq data from liver and kidney (Marioni et al., 2008). Both methods, DEGseq and Pyicos, can accept as input read counts or RPKM values for each gene; and they also can estimate a theoretical replica when an experimental one is not provided. We measured the significance of the enrichment of liver over kidney in all genes using four different combinations of input

![Figure 1](image_url)
data: replicated count, non-replicated count, replicated RPKM and non-replicated RPKM. Comparing the Z-scores calculated by Pyicos and DEGseq, we observe a high correlation between both methods for all the inputs (Supplementary Table 3).

![Figure 2](https://example.com/figure2.png)

**Figure 2. Prediction of differentially expressed (DE) genes.**

a) Receiving Operating Characteristic (ROC) curves for the benchmarking against the microarray data (Marioni et al., 2008) for DESeq, Pyicos, edgeR and DEGseq using read-counts and replicated data. b) Precision-Recall curves for the benchmarking against microarray data for the same four methods using read-counts and replicated data. c) Precision-Recall curves for the benchmarking against microarray data for the same four methods using read-counts and replicated data. d) ROC curves of the different normalization methods: read-counts (Counts), TMM-normalized counts (TMM counts), RPKMs and TRPKs, for the microarray benchmarking. e) Absolute differences of the medians from the length distributions of DE and non-DE genes when not normalized counts are used. However, they have been observed to produce length-biases in the determination of DE genes (Oshlack and Wakefield, 2009). To explore the correction effect of the various normalizations on the length-biases, we calculated the lengths for all DE and non-DE genes predicted by Pyicos EA using counts, TMM normalized counts, RPKM and TRPK; and compared them to the lengths of the DE and non-DE genes from the microarray results (Supplementary Figures S18). We observe a significant difference between the absolute length medians of DE and non-DE genes when not normalized counts are used. However,

To further estimate the accuracy of Pyicos EA, we also ran DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010), on the same benchmarking sets, using again the four different input types. Using the microarray experiments from (Marioni et al., 2008), we created a benchmarking set (Methods). We found an overall good performance for Pyicos, with AUCs between 0.791 and 0.813 and a behaviour similar to DEGseq. All four methods agree similarly well with the benchmarking set when replicated read-count data is used (Figure 2a). However, we find greater disagreements between methods and lower accuracy for the other three input combinations (Supplementary Figure S15 a-c). Moreover, the precision-recall curves show a similar level of agreement for replicated read-count data (Figure 2b) and confirm the good performance of Pyicos for all combination of inputs (Supplementary Figure S15 d-f). For DESeq it was not possible to make calculations on non-integer values, so we restricted the calculations to counts, which resulted in a high AUC when replicas were used. Although edgeR was developed for replicated count data, it was also run with non-replicated count data and performed similarly well.

In order to test the memory usage and processing time of Pyicos EA calculation, we applied EA from all four methods on the CEBPA dataset to compare the ChIP-Seq sample to the control sample. While Pyicos and DEGseq can be run directly on BED files, edgeR and DESeq require count data as input. Hence, we combined these last two methods with BEDTools for the comparison. Since the most time consuming and memory intensive operation is the counting of reads per region, which is done with BEDTools, edgeR and DESeq show the same behaviour. The performance of DESeq suffers in terms CPU time and especially memory usage. Although Pyicos is not as fast as the combination of BEDTools with edgeR or DESeq, it still can handle very large datasets at much lower memory usage (Supplementary Figure S16 and Figure 2c). BEDTools algorithm compromises memory usage in exchange of execution time, whereas Pyicos algorithm focuses on memory efficiency.

### 3.4 Normalization of the RNA-Seq data

As genes have different lengths and RNA-Seq samples may vary considerably in size, proper normalization of the input data is essential. The trimmed mean of M values (TMM) normalization (Robinson and Oshlack, 2010) aims to correct biases due to differences in samples sizes and expression patterns. TMM normalization on count data correctly places the M median on zero on our gene set (Supplementary Figures S17a and S17b). We considered the combination of TMM normalization with RPKM densities, which we hypothesize would improve results as it takes into account both gene-lengths and sample sizes. We thus define a TRPK density as the Tmm-normalized Read Per Kilobase density (Supplementary Material). First, as expected, using TRPK we can achieve a correction of the M-median for RPKM (Supplementary Figures S17c and S17d). Next, we assessed whether the TMM normalization also results in an improvement of the accuracy calculation using the array results. However, for this particular benchmarking set, we could not observe such an improvement (Figure 2d).

Intriguingly, read-counts seem to perform slightly better in our benchmarking analysis, as shown above. However, they have been observed to produce length-biases in the determination of DE genes (Oshlack and Wakefield, 2009). To explore the correction effect of the various normalizations on the length-biases, we calculated the lengths for all DE and non-DE genes predicted by Pyicos EA using counts, TMM normalized counts, RPKM and TRPK; and compared them to the lengths of the DE and non-DE genes from the microarray results (Supplementary Figures S18). We observe a significant difference between the absolute length medians of DE and non-DE genes when not normalized counts are used. However,
this difference decreases when we use the RPKM and TMM-normalized read-counts, achieving the smallest difference and reaching a similar value to the one found for the microarray when TRPK is used (Figure 2e).

3.5 Enrichment analysis on broad ChIP-Seq data

We next provide evidence that Pyicos EA approach is also suitable for the analysis of broad ChIP-Seq data when comparing two conditions. Broad ChIP-Seq data does not generally produce clearly delimited regions; hence, one cannot speak of peaks. However, it is interesting to be able to measure how the signal changes between two conditions or cell types in specific regions. Pyicos allows the calculation of significant enrichment in predefined regions, e.g., gene-body, promoter regions, etc., which are given as input in a BED file. To show the applicability of Pyicos on broad ChIP-Seq data, we tried to reproduce the relation between RNAPII activity and transcription (Sultan et al., 2008) and between the H3K36me3 chromatin signal and transcription (Pokholok et al., 2005; Joshi and Struhl, 2005). For this, we calculated the correlation of the enrichment of H3K36me3 and RNAPII, with that of RNA-Seq using ENCODE datasets (ENCODE Consortium, 2011). For each datatype we calculated the EA values of K562 over NHEK, considering the variation between two K562 replicas (Methods). We selected a cut-off of \( \text{Z-score} > 10 \), obtaining 1238 genes with significant changes in RNAPII and RNA-Seq, and 377 genes with significant changes in H3K36me3 and RNA-Seq. For these genes, we measured a Pearson correlation between the enrichment Z-scores of 0.87 and 0.90, for RNA-Seq vs. RNA-Seq and H3K36me3 vs. RNA-Seq, respectively.

3.6 CLIP-Seq data analysis

CLIP-Seq reads can be mapped to the genome or to the transcriptome in order to detect RNA binding activity. CLIP experiments normally use a control with an unspecific antibody to check the presence of unspecific binding. However, such a control is not always sequenced; hence, a method to calculate enrichment without a control was proposed (Yeo et al., 2009; Xue et al., 2009). This method, called modified False Discovery Rate (modFDR), is included in Pyicos (see Supplementary Material). In order to show Pyicos applicability to CLIP-Seq experiments, we reproduced the results from (Xue et al., 2009), where the interacting RNA sites for the polypyrimidine tract binding protein (PTB) were mapped to the reference human genome. We used all the mapped reads from this experiment and ran Pyicos modFDR operation considering the gene-body of the RefSeq genes as regions to calculate the significant read clusters. Pyicos retrieved 92% of the 32,298 regions defined in (Xue et al., 2009) on the same gene set, which means 91.5% of the reported 10,515 genes having at least one significant cluster (Figure 3a). Moreover, the genes missed by Pyicos modFDR were those with a low number of reads: from the 467 genes missed by Pyicos, 89.7% contained clusters with heights of at most 3 reads (Figure 3b), suggesting that they are likely false positives. Moreover, there were 898 genes selected by Pyicos that were not in the predictions from (Xue et al., 2009), from which 11.1% had clusters that are higher than 3, indicating that Pyicos may still recover real sites that are borderline. Nonetheless, the clusters missed by either case had p-values closely above the used 0.001 thresholds, indicating that the observed variation was mostly due to the random nature of the background model calculation.

Figure 3. Detecting significant clusters in CLIP-Seq. a) Genes with at least one significant cluster b) Heights of cluster

4 DISCUSSION

We have described Pyicos, a powerful tool for the analysis of mapped High-Throughput Sequencing (HTS) reads. Pyicos framework facilitates the analysis of different HTS datatypes. We have described its application to ChIP-Seq, RNA-Seq and CLIP-Seq data using the three corresponding protocols: callpeaks, enrichment and clipseq. In callpeaks we define a peak-score in terms of a Poisson p-value, which is calculated independently for each chromosome, and in terms of one peak property, either the peak height or read-count. The peak-score therefore takes into account the differences across chromosomes and the fact that peaks with the same height or read count may have different p-values depending on the chromosome in which they are located. Using ChIP-Seq data for PR, CTCF, CEBPA and NRSF, we have shown that the peak-score provides an appropriate ranking for peaks. Interestingly, using either the peak height or read-count can make a difference depending on the dataset, as we found an increase of about 7% in the fraction of the top 500 PR peaks with motifs, leading to a better peak definition compared to the other methods. We have further shown that the subtraction of the control is effective to increase the fraction of peaks with motif, indicating that it eliminates potential false positives.

Pyicos has the advantage that all operations described are configurable by the user to keep as much flexibility as possible, since not all of them may be applicable to a dataset. For instance, splitting peaks seems to result in improvements only for CTCF. Similarly, although Pyicos allows the user to choose the number of tolerated duplicated reads, we found that in all datasets the best peak definition is achieved by removing all duplicates.

Furthermore, using various measures, we have compared callpeaks with three other methods specifically developed for punctuated ChIP-Seq data: MACS, USeq and FindPeaks. Regarding peak definition, we have found that FindPeaks, Pyicos and MACS show very similar spatial resolutions, which are also higher than those for USeq peaks. Furthermore, we observed that peaks ranked by Pyicos and MACS show a slightly higher fraction of motif-containing peaks than those ranked by FindPeaks and USeq.

Peak detection has been assessed using ChIP-qPCR validated positive and negative regions for NRSF. We found that all methods
perform similarly, probably due to high agreement between their generated peaks. The methods with the highest pairwise correlations are Pyicos, MACS and USeq. FindPeaks showed a lower correlation, probably due to a different handling of the control sample: whereas FindPeaks compares a peak-height with the distribution of peak-heights from the control sample, the other three methods compare signal to control locally, using windows (MACS and USeq) or at base-pair resolution (Pyicos). In summary, our results lead us to conclude that callpeaks provides an accurate protocol for peak detection for punctuated ChIP-Seq data. Moreover, due to the more flexible usage of the various operations compared to other methods, it allows to design a customized analysis of the ChIP-Seq data.

We have further illustrated Pyicos flexibility by applying it to other datatypes. Using data from RNA-Seq and microarray experiments on liver and kidney samples, we have shown that Pyicos can recover differentially expressed genes with high accuracy and that is in fact comparable to methods specifically designed for differential expression analysis from RNA-Seq. Moreover, Pyicos performs well using different inputs: replicated or non-replicated, read-counts or RPKM. DEGseq, which can also work with RPKM, shows similar accuracy as Pyicos and both correlate well in terms of the predicted Z-scores. The other two methods, DESeq and edgeR, were not designed to work with RPKM, hence could not be included in the comparison. The fact that Pyicos performs well with simulated replicated data presents the advantage of making possible to analyse many of the published datasets that have been produced without replica.

It is surprising that the highest accuracy is achieved on count data, since it is known that using counts leads to a length bias in DE gene detection. This is probably because the genes selected from the microarray for benchmarking are not much affected by the length-bias, since the accuracy hardly changes when using RPKM. This also suggests that RPKM alone does not provide the optimal normalization method. However, using a new density definition, TPK, which combines RPKM with the TMM normalization, the length differences between DE and non-DE genes are reduced to a level close to that of the microarray. Nonetheless, the TMM normalization is based on the assumption that the majority of regions tested do not change significantly, which might not hold true for some pairs of samples or for certain sets of regions. Accordingly, Pyicos implements this as an option to the user.

The flexibility of Pyicos is further demonstrated by applying the enrichment analysis (EA) protocol to broad ChIP-Seq data. In particular, using ENCODE ChIP-Seq data for H3K36me3 and RNAPII from the cell-lines K562 and NHEK, we obtain a high correlation for the Z-scores of these signals when compared to the enrichment Z-scores for RNA-Seq between the same cell-lines. Thus the EA of Pyicos using RPKM provides a tool to analyze broad ChIP-Seq data of various sorts, which would otherwise require a combination of approaches to be analysed (Young et al., 2011). A further advantage is that Pyicos can directly accept BED files with mapped reads and regions of interest, unlike DESeq and edgeR. Pyicos can also calculate enriched regions de novo genome-wide, using the reads from two experiments that are overlapping or sufficiently close in position. This is particularly useful for the analysis of signals for which the user does not know where to expect the enrichment relative to genome annotations, like enhancer elements. We have further shown that Pyicos can also be used to process CLIP-Seq data without a control. We expect that many of the basic Pyicos operations could be applicable to other datatypes and could possibly be combined to generate new analysis protocols.

Some of the operations described can become impracticable for some analysis tools due to the amount of reads produced by a single HTS experiment nowadays. The bottleneck of HTS data analysis mostly lies in the memory usage of the software and in the storage and retrieval of data. Indeed, HTS data generation seems to be outpacing the improvements in CPU and disk storage (Kahn, 2011). Moreover, HTS data has become ubiquitous in genomic research; hence, we should aim to provide software that can be adapted to the available computing resources and, in particular, that can run on the average desktop computer. For these reasons, we developed Pyicos minimizing RAM usage and maintaining reasonable CPU time usage. Pyicos callpeaks protocol for ChIP-Seq data outperforms the other tested methods concerning memory usage while it stays competitive in running time. FindPeaks and USeq algorithms load entire files in memory, allowing for better time performance. However, this is only practical on a computer with enough RAM. For example, an experiment with 100 million reads using as input a BED file, would require more than 4 Gigabytes of available RAM. For EA, the main bottleneck lies in the calculation of read-counts or RPKM on the input regions. Although Pyicos is not as fast as BEDTools combined with DESeq or edgeR, its memory performance is far superior, allowing the handling of very large datasets. As made patent in the last few years, read files are increasing enormously in size with the development of the technology, making memory usage a critical feature in HTS analysis software.

We conclude that the added value of having a modular tool is not at the cost of accuracy or performance; hence, Pyicos provides a useful framework for the analysis and integration of heterogeneous HTS data. Finally, as Pyicos is open-source, we encourage the addition of new operations in order to combine them with already existing ones and possibly to create new analysis protocols.

5 ACKNOWLEDGEMENTS

The authors would like to thank Christian Perez-Llamas, David González, Julien Lagarde and Robert Castelo for useful discussions and Alba Jene for discussions and comments on the manuscript. We also acknowledge Mark Kon, Manolis Kellis and Ana Rojas for providing us the opportunity to present Pyicos in their institutions.

Funding: SA is supported by an FI grant from the Generalitat de Catalunya and JGV is supported by a FPI grant from the Spanish Ministry of Science (MICINN). This work was further supported by the grant BIO2008-01091 (Plan Nacional I+D from MICINN) and by the project EURASNET-LSHG-CT-2005-518238 from the European Commission.

6 REFERENCES


