OXFORD

Briefings in Bioinformatics, 2016, 1–15

doi: 10.1093/bib/bbw029
Paper

MuSERA: Multiple Sample Enriched Region Assessment

Vahid Jalili, Matteo Matteucci, Marco J. Morelli and Marco Masseroli

Corresponding author. Marco Masseroli, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy. Tel.: +39-02-2399-3553; Fax: +39-02-2399-3411; E-mail: marco.masseroli@polimi.it

Abstract

Enriched region (ER) identification is a fundamental step in several next-generation sequencing (NGS) experiment types. Yet, although NGS experimental protocols recommend producing replicate samples for each evaluated condition and their consistency is usually assessed, typically pipelines for ER identification do not consider available NGS replicates. This may alter genome-wide descriptions of ERs, hinder significance of subsequent analyses on detected ERs and eventually preclude biological discoveries that evidence in replicate could support. MuSERA is a broadly useful stand-alone tool for both interactive and batch analysis of combined evidence from ERs in multiple ChIP-seq or DNase-seq replicates. Besides rigorously combining sample replicates to increase statistical significance of detected ERs, it also provides quantitative evaluations and graphical features to assess the biological relevance of each determined ER set within its genomic context; they include genomic annotation of determined ERs, nearest ER distance distribution, global correlation assessment of ERs and an integrated genome browser. We review MuSERA rationale and implementation, and illustrate how sets of significant ERs are expanded by applying MuSERA on replicates for several types of NGS data, including ChIP-seq of transcription factors or histone marks and DNase-seq hypersensitive sites. We show that MuSERA can determine a new, enhanced set of ERs for each sample by locally combining evidence on replicates, and prove how the easy-to-use interactive graphical displays and quantitative evaluations that MuSERA provides effectively support thorough inspection of obtained results and evaluation of their biological content, facilitating their understanding and biological interpretations. MuSERA is freely available at http://www.bioinformatics.deib.polimi.it/MuSERA/.

Key words: next-generation sequencing; ChIP-seq and DNase-seq data analysis; combined evidence in replicates; integrated genome browser; genomic data visualization

Background

Next-generation sequencing (NGS) is a multi-purpose technology, which allows precise determination of DNA or RNA sequences within a sample of interest [1]. In particular, some strategies allow enriching for regions of cellular DNA characterized by some common property: chromatin immunoprecipitation followed by NGS (ChIP-seq) [2] reveals genome-wide DNA–protein interactions and chromatin modifications, e.g. histone marks, while

Vahid Jalili is a PhD candidate at Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy. His research on tertiary analysis of next-generation sequencing data is focused on systematic solutions for analytical and computational challenges.
Matteo Matteucci is associate professor of pattern recognition and machine intelligence at Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy. His research includes pattern recognition, machine learning, classification, robotics, computer vision and signal processing. He has co-authored more than 150 scientific international publications.
Marco Morelli is a researcher at the Center for Genomic Science of IIT@SEMM, Fondazione Istituto Italiano di Tecnologia, Milano, Italy. A physicist by background, his research interests span from bioinformatic analysis of next-generation sequencing data to dynamical models and machine learning algorithm for pattern recognition in big data.
Marco Masseroli is associate professor of bioinformatics and biomedical informatics at Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy. His research interests include distributed Internet technologies, biomolecular databases, biomedical terminologies and bio-ontologies to effectively retrieve, manage, analyze and semantically integrate genomic information with clinical and high-throughput genomic data. He is author of more than 170 scientific articles.
Submitted: 1 December 2015; **Received (in revised form):** 1 February 2016

DNase I sequencing (DNase-seq) [3] provides a global view of the open chromatin in a cellular sample through the identification of hypersensitive sites. The analysis of these NGS data returns, in both techniques, a list of enriched regions (ERs), often named 'peaks' and defined through their genomic coordinates; usually these peaks are also associated with a statistical significance score, i.e. a *P*-value. The availability of NGS data has opened the possibility of a comprehensive characterization of genomic and epigenomic landscapes; yet, extracting such biological information from raw data requires the use of complex computational pipelines, which include the identification of the ERs as a key step.

Although NGS experimental protocols recommend the production of at least two replicates for each sequenced sample, specific methods and tools currently used for ER calling (e.g. MACS [4] or ZINBA [5]) usually consider only a single sample at a time, and use global stringent thresholds to eliminate the noise in the data [6]; then, the ERs extracted from individual replicates are compared, and typically only the ERs identified in multiple replicates are retained (e.g. by simple intersection or using the irreproducibility discovery rate (IDR) method [7]). As we recently demonstrated [8], considering single samples and applying individual stringent thresholds lead to the discovery of the strongest ERs only, and it may discard true, although less intense, ERs, which in turn could be picked up by taking advantage of the increased sensitivity provided by replicates. Neglecting weak ERs may eventually distort the genome-wide picture of the genomic locations of the ERs of interest, hamper the significance of the following analyses on the identified ERs and ultimately prevent biological discoveries that could be supported by considering also the true, but less intense, ERs (i.e. genomic features) present in the NGS data. Alternative methods considering multiple samples exist, but were designed for other purposes (e.g. jMOSAiCS [9], which was designed to identify combinatorial patterns of enrichment across multiple ChIP-seq samples); they can be used also to discover ERs across replicates, but at the cost of a higher number of not validated peaks (false positives) [8].

Recently, we proposed a novel method that simultaneously considers multiple ChIP-seq replicates for transcription factors (TFs) and rigorously combines local evidence of ERs; the method provides new sample-specific peak lists taking into account the combined evidence of ERs, called with a threshold less stringent than usual [8]. In the tests performed on public ChIP-seq data sets for the Myc TF in K562 human cells, this method allowed to significantly extend the number of detected ERs, with respect to single-sample analysis with an equivalent significance threshold. The newly discovered ERs were validated by motif analysis and overlap with open chromatin regions. Furthermore, comparison with alternative methods (i.e. IDR and jMOSAiCS) showed that the method discovers more validated peaks than the former and less peaks than the latter, but with a better validation.

The authenticity of the ERs discovered by combining evidence depends on a variety of factors, including the quality of replicates and called ERs, as well as the choice of parameter values used to combine the evidence. An assessment of the resulting ERs should always be performed: for example, it could be achieved by visualizing the results in a genome browser, inspecting ER nearest-neighbour distributions, and/or comparing the ERs with known genomic annotations (functional analysis). These last two procedures involve the calculation of the distribution of distances between ERs of replicates, or between ERs and known genomic features (e.g. genes, promoters or other regulatory regions). Such distributions may show, for example, that ERs in different sets are relatively close to each other, but they are not overlapping, or that they are not at specific distances from known genes; if this is not expected (e.g. as in the case of replicates regarding ChIP-seq experiments of TFs), it may suggest a revision of the parameter values used for peak calling, or for combining ER evidence.

Addressing all the above aspects, here we review MuSERA, a novel, broadly useful, advanced graphical tool that efficiently implements, extends and generalizes the original method presented in [8], and, in addition, integrates commonly used analysis features that allow performing easily further assessments, genomic annotations and functional analyses on the identified ERs. Through its intuitive graphical interface, MuSERA provides several graphic displays that help the user in gaining a deeper insight and biological evaluation of the analysis results. We review the main MuSERA features, describing how they are implemented, and we apply MuSERA to several types of data, from ChIP-seq experiments of TFs or histone marks, both narrow and broad, to DNase-seq experiments. Finally, we review and discuss some examples of the analysis of these data using MuSERA, which show the relevance of the additional ERs identified with MuSERA, as well as the efficacy of the graphical displays of the computational results that MuSERA provides in supporting the biological interpretation of NGS experiments.

Notations

An ER is a unique independent entity, denoted by r_{ji}, belonging to the sample $R_j = \{r_{j1}, \ldots, r_{ji}, \ldots, r_{jI}\}$, with $U = \{R_1, \ldots, R_j, \ldots, R_J\}$ being a set of replicates; the index i, with $1 \le i \le I$, identifies the regions within a given replicate, and the index j, with $1 \le j \le J$, identifies the replicates. An ER is characterized by its genomic coordinates ($chromosome_{ji}$, $start_{ji}$, end_{ji}) and $P\text{-}value_{ji}$ (p_{ji}). The significance of r_{ji} is stratified by the 'stringent' (T^s) and 'weak' (T^w) thresholds, with $T^s < T^w$; accordingly, $R_j^s = \{r_{ji} \mid p_{ji} < T^s\}$, $R_j^w = \{r_{ji} \mid T^s \le p_{ji} < T^w\}$ and $R_j^b = \{r_{ji} \mid p_{ji} \ge T^w\}$ represent the sets of 'stringent', 'weak' and 'background' ERs, respectively, for the replicate sample j.

Let $R_{ji,*}$ be the set of all ERs intersecting with r_{ji} (including r_{ji}), where only the intersecting ER with the lowest/highest *P*-value of each sample is considered if multiple intersecting ERs exist in a sample, and let $K = |R_{ji,*}|$, where $1 \le K \le J$ by definition. According to the method in [8], the significance of an ER in $R_{ji,*}$ is assessed by computing a 'combined evidence' X^2 statistics (i.e. the sum, over the K ERs in $R_{ji,*}$, of $-2 \ln p_{ji}$), which, according to the Fisher's combined probability test [10], follows a χ^2 distribution with $2K$ degrees of freedom; the right-tail cumulative probability of this χ^2 distribution defines the ER combined evidence p_{ji}^{comb}, whose comparison with a 'stringency threshold' γ defines 'confirmed' ($R_j^c = \{r_{ji} \mid p_{ji}^{comb} \le \gamma\}$) and 'discarded' ($R_j^d = \{r_{ji} \mid p_{ji}^{comb} > \gamma\}$) sets of ERs for each replicate sample j. Subsequently, the method generates an 'output set' (R_j^o) for each replicate sample by applying a multiple testing correction procedure on the confirmed ERs of the sample. Additionally, for each replicate sample j, the method defines the following sets: (i) 'stringent confirmed' $R_j^{sc} = \{r_{ji} \mid p_{ji} < T^s \wedge p_{ji}^{comb} \le \gamma\} \subseteq R_j^c$, (ii) 'stringent discarded' $R_j^{sd} = \{r_{ji} \mid p_{ji} < T^s \wedge p_{ji}^{comb} > \gamma\} \subseteq R_j^d$, (iii) 'weak confirmed' $R_j^{wc} = \{r_{ji} \mid T^s \le p_{ji} < T^w \wedge p_{ji}^{comb} \le \gamma\} \subseteq R_j^c$, (iv) 'weak discarded' $R_j^{wd} = \{r_{ji} \mid T^s \le p_{ji} < T^w \wedge p_{ji}^{comb} > \gamma\} \subseteq R_j^d$, (v) 'multiple-testing confirmed' R_j^{mtc} (with $R_j^{mtc} = R_j^o$) and (vi) 'multiple-testing discarded' R_j^{mtd} (with $R_j^{mtc} + R_j^{mtd} = R_j^c$). In addition to the method from [8], MuSERA provides also a single 'unified output set' (R^{uo}) representing the confirmed ERs present in all the R_j^o sets of the

combined replicate samples. This R^{uo} set is built by merging all ERs in all the R_j^o sets (one for each replicate sample), so that, for each group of overlapping ERs in the R_j^o sets, only a single ER is present in R^{uo}, having as left-end and as right-end the left-most left-end and the right-most right-end of the overlapping ERs, respectively. A significance score is assigned to each ER in R^{uo}, calculated by rigorously combining the significance of the overlapping ERs in the R_j^o sets using the Fisher's method [10].

MuSERA features

MuSERA combines replicates to increase the statistical significance of ERs. It assigns ERs to different sets and provides an integrated genome browser for their visualization. Furthermore, for the evaluation of the replicate-combined results, it offers several additional features, including 'genomic annotation and functional analysis of enriched regions', 'nearest enriched region distance distribution' and 'global correlation assessment of enriched regions', for in-depth investigation of each of the ER sets. MuSERA bins distances based on a user-modifiable window size, shows results on tables and plots (supporting user-friendly zoom and pan) and allows operations to be applied on user-selected chromosomes. These and other MuSERA features, including 'interactive and batch execution' as well as 'input/output standard data formats', are reviewed in the following sections, where we show the relevance and utility of MuSERA for biological investigation. MuSERA is a .NET application implemented in C# that runs primarily on Microsoft Windows® and may be run also on other operating systems using an Oracle Virtual Box virtual machine freely provided for non-commercial use.

Combining replicates

To combine ER evidence present in sample replicates, MuSERA extends the method described in [8], and implements it in an efficient multi-threaded environment. Each ER is categorized as 'stringent', 'weak' or 'background' with respect to the significance of the ER according to user-defined stringent (T^s) and weak (T^w) thresholds, with only 'stringent' and 'weak' ERs being considered for replicate evidence combination. The algorithm combines the P-values of intersecting ERs using the Fisher's method [10], if and only if the number of such intersecting ERs is above or equal to a user-defined lower bound (C); accordingly, it assigns the property of 'confirmed' or 'discarded' to each of the intersecting ERs if the combined evidence p_{ji}^{comb} is below or is not below, respectively, the user-defined combined stringency threshold γ (see the 'Notations' section). Besides, overlapping ERs located in a number of samples below the required value of the parameter C are 'discarded'. Each replicate sample contributes to the evidence combination with single evidence only; hence, if a sample has multiple ERs overlapping with a single ER of another sample, only the most/least stringent (according to user definition) overlapping ER of the former replicate is considered for the evaluation of the ER of the latter replicate.

Genomic annotation and functional analysis of ERs

An ER can overlap known genomic loci, like promoters or other regulatory elements of genes. Besides, a gene might be regulated by a TF bound to a DNA regulatory element far from its promoter (e.g. regulatory elements called 'enhancers' [11] can be located far from transcription start, like for the Sonic hedgehog (Shh) gene in mouse [12]), even interspersed with other non-regulated genes [13, 14]. MuSERA can efficiently assign an ER to the closest up-/down-stream genomic feature [e.g. gene transcription start site, promoter region, Coding DNA Sequence (CDS) or enhancer], thanks to its optimized implementation using an adaptive binning of data (see 'Implementation' section and Figures 1B and 2). Furthermore, MuSERA estimates the 'ER-to-feature overlap score', by determining the number of ERs intersecting with genomic annotations (e.g. known genes, 3′/5′ untranslated regions, CDSs, intergenic regions (IGR), introns, promoter regions), or with any experimentally verified binding sites uploaded in MuSERA by the user as annotations in General Transfer Format (GTF). Additionally, it estimates the 'ER-to-feature distance distribution' between the ERs and the closest up-/down-stream features per functional group. All these options allow better biological evaluation of the distribution of the ERs in the genomic context.

Nearest ER distance distribution

MuSERA can compute the ER nearest neighbour distance distribution (NND). In each analysis session consisting of at least two samples, the ERs of each sample are grouped into different sets before ('stringent' or 'weak' set) and after ('stringent confirmed', 'weak confirmed', 'stringent discarded', and 'output' set) the multiple-sample analysis. To estimate the NND, after the user chooses the desired sample(s) and set(s) to be considered, for each ER, MuSERA determines the distance to the nearest ER; an option is available to treat all selected samples and sets either as a single entity or as distinct entities. In the case of single entity, the closest neighbour of an ER could be an ER belonging to any set of any sample of the analysis session. In the case of distinct entities, the closest neighbour of an ER is determined within the same set and sample of the ER.

Global correlation assessment of ERs

The similarity between replicates is frequently assessed either before peak calling, using genome-wide read densities, or after peak calling, using the identified ERs. Pearson's product-moment correlation coefficient (PCC) [15] is a threshold-independent and scale-invariant method [16] commonly used to compute a global correlation assessment between replicates. PCC is also used after the peak calling when binned signal intensities are provided, either in a separate 'wiggle' file per sample or as numerical vectors per identified ER (e.g. data set chipseq_mES of [17]). Similarly, the Jaccard Similarity Coefficient (JSC) is a statistical method for correlation/diversity assessment of samples, consisting on the ratio between the cardinalities of the intersection and the union of two sets; it can be used both before peak calling (e.g. [18] increases genes detection power of RNA-seq data using JSC for global similarity filtering) or as a post peak calling correlation assessment procedure (e.g. [19, 20]).

MuSERA determines both region-level and base-pair-level correlations between all pairs of sets using JSC (see Figures 2 and 3). They are respectively computed as the ratio between the number of overlapping regions (region-level correlation), or genomic bases (base-pair-level correlation), and the total number of regions, or genomic bases, in the considered sets. Base-pair-level correlation is more stringent and is to be preferred when the position of the ERs is known with more certainty, or when

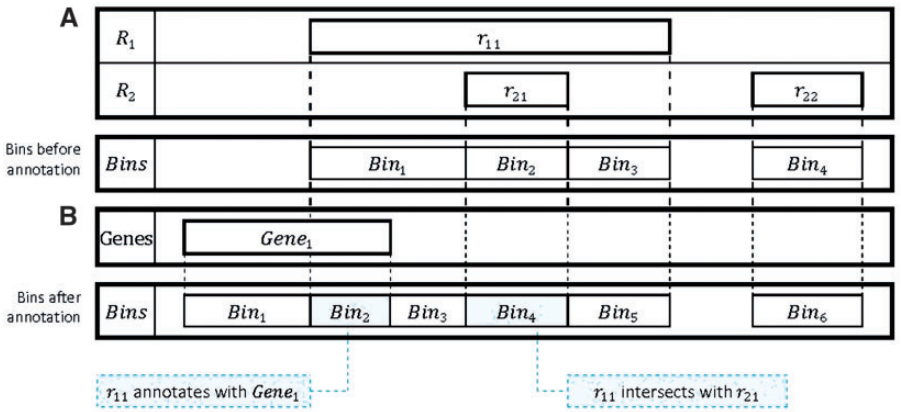


Figure 1. Binned data. (**A**) A set of bins is created with respect to the ERs of the replicates. (**B**) The bins are then modified with respect to known binding sites/genomic annotations. Each bin contains all available information for the segment of genome it represents; for instance, in (**B**), Bin_2 corresponds to r_{11} intersecting with $Gene_1$ at the genome position determined by the Bin_2 coordinates. Bins are orderly stored by their genomic position, which enables a binary search for a specific bin. An ER is possibly represented by more than one bin, i.e. by all bins that start/end within the ER coordinates (e.g. in (**B**), the ER r_{11} is represented by bins Bin_2, Bin_3, Bin_4 and Bin_5); therefore, comprehensive information about an ER is provided by the union of all bins spanning it.

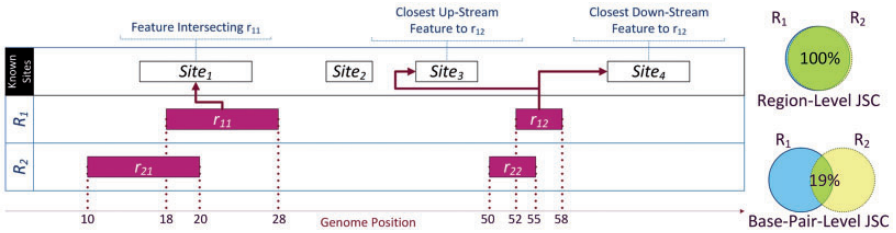


Figure 2. Genomic annotation and correlation assessment. For each ER, MuSERA computes the distance between the ER and the closest known genomic feature (site). If an ER overlaps a feature (e.g. r_{11} and $site_1$), their distance is 0; otherwise two distances are computed between the ER and the closest up-stream and down-stream features, respectively. MuSERA determines the correlation between samples in terms of the Jaccard Similarity Coefficient (JSC), both at region level and base-pair level. The right-hand side of the figure highlights the possibility of considerable difference between the two levels.

the experimental protocols have a low level of noise. Region-level correlation is instead more permissive, as it scores the overlap of entire regions rather than quantifying the magnitude of this overlap; this correlation measure is then to be preferred in the presence of heterogeneous or noisy data sets.

Interactive and batch execution

MuSERA implements two execution procedures: 'Interactive' and 'Batch' processing. The 'Interactive mode' is provided through a graphical user interface (GUI) with a wide range of review graphical features; it is intended for processing a limited number of samples, where results need to be reviewed through multiple user iterations for parameter tuning (see Figure 4 for cross-functional flowchart). The 'Batch mode' is suitable for processing a high number of samples with a given set of parameters; it reads 'jobs' defined in a simple way through an Extensible Markup Language (XML) file and it has a limited set of review features. The XML file is compliant with the World Wide Web Consortium Document Object Model (DOM) level 1 core and DOM level 2 core recommendations, and its schema has been defined so to ease the work of the end user in the definition of 'jobs'.

Input/output standard data formats

MuSERA processes ERs and allows further investigation of results using genome annotations as references. ERs can be read from tab-delimited files consisting of the ER genomic interval attributes (chromosome, start, end and *P*-value) as essential fields; common standard tab-delimited formats such as Browser Extensible Data (BED), ENCODE narrowPeak and ENCODE broadPeak are of such kind. Genome annotations like Reference Sequence (RefSeq) or GENCODE genes can be parsed and loaded from files in standard formats such as the GTF.

MuSERA exports each of the resulting ER sets in a separate BED file. Additionally, an XML file is created for each R_j^o, R_j^c and R_j^d set, containing extensive explanatory information for each included ER, such as (i) ER signature (i.e. chromosome, start, end, name, *P*-value), (ii) initial categorization (i.e. stringent or weak), (iii) computed combined *P*-value (X^2) and corresponding right-tail probability (p^{comb}) and (iv) signatures of the ERs it is combined with, including the sample name they belong to. Chromosome-wide basic statistics of each input sample (e.g. widest/narrowest peak, lowest/highest *P*-value and average/median/standard-deviation of *P*-values) are provided in a separate text file. When running in 'Batch Mode', MuSERA also exports a text file for each analysis session, providing comprehensive information about the parameters and the overall analysis results for any future reference.

Implementation

Overview

MuSERA is a .NET application written in Model-View-ViewModel (MVVM) pattern [21], with a GUI developed in Windows Presentation Foundation (WPF) graphical system and

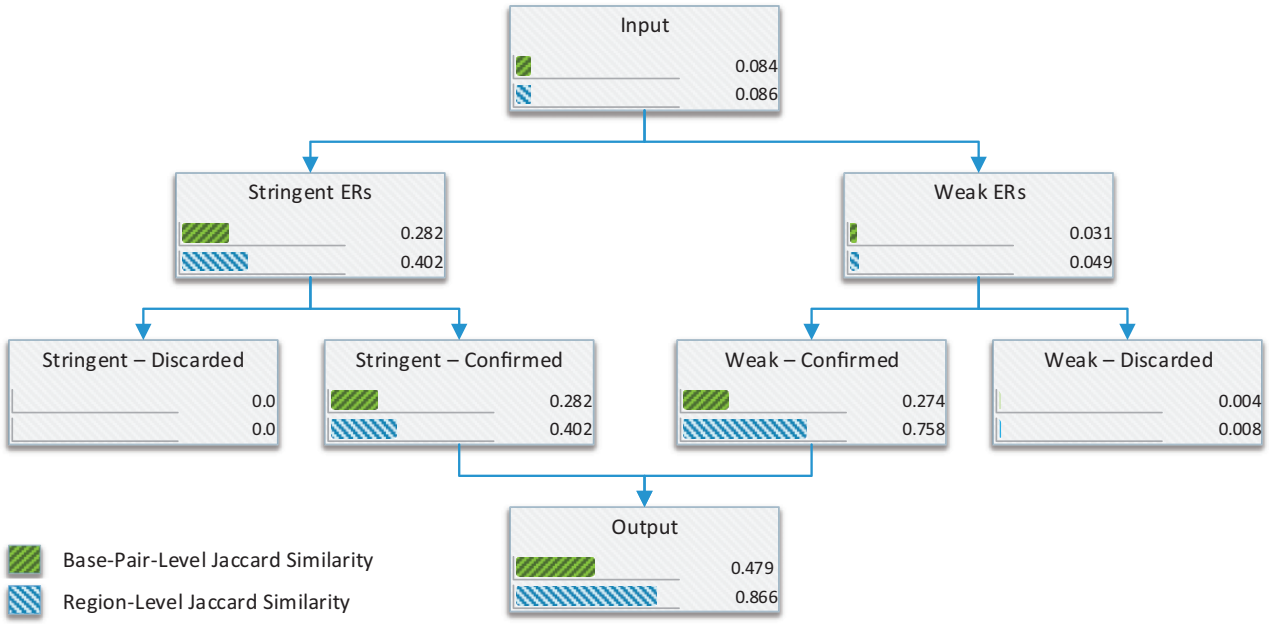


Figure 3. Correlation assessment hierarchy. During processing of two replicate samples, MuSERA estimates the Jaccard Similarity Coefficient between the two samples and for each of their computed ER sets. Values are shown for the ENCODE samples wgEncodeSydhTfbsK562CmycIfna30StdAlnRep1 and wgEncodeSydhTfbsK562CmycIfna30StdAlnRep2 (processed with analysis parameters: BioRep, $T^s = 10^{-8}$, $T^w = 10^{-4}$, $\gamma = 10^{-8}$, $C = 1$), which overall show a rather low correlation (Input). In these samples the peaks are called using MACS2.0 [4] with 0.001 *P*-value threshold; hence, such low correlation is expected because of low signal-to-noise ratio. Initial classification of the ERs in each replicate (i.e. Stringent ERs versus Weak ERs) confirms that in the replicates stronger evidence correlates better than weaker one. Combining the samples, each of the two initial categories is divided into the Confirmed and Discarded sub-categories; ERs in the Confirmed sub-categories result to be considerably more correlated compared with their corresponding ERs in the Discarded sub-categories.

business logic written in C# programming language. Being the GUI implemented in WPF, leveraging on DirectX and on the Graphical Processing Unit in a Multithreaded Apartment (MTA) model, MuSERA delivers a high-end smooth, interactive and user-friendly graphics. The MVVM pattern and MTA model enable separation of business logic and GUI process load, which avoids any possible lag on either side. As far as code metrics are concerned (calculated by Microsoft Visual Studio), MuSERA consists of roughly 6500 lines of code with a maintainability index of 82, cyclomatic complexity of 2.000 and 9 maximum depth of inheritance [22].

MuSERA source code is freely available under open-source GPLv3 license at http://musera.codeplex.com/; its implementation for MS-Windows systems and an Oracle Virtual Box virtual machine for its evaluation on other systems (e.g. Linux, Mac) are freely available for downloading for non-commercial use from http://www.bioinformatics.deib.polimi.it/MuSERA/, where the MuSERA user manual (Supplementary file 1) is also available.

Interactive and batch execution

The 'Interactive mode' is implemented using the Multi-Threaded Apartment model, while the 'Batch mode' uses the Single-Threaded Apartment model. The two modes use common thread-safe components that enable concurrent execution of modes with no intervention, and the possibility to set the process priority of the 'Batch mode'.

The 'Batch mode' executes a series of 'jobs' collected in an 'at-job' that is defined in an XML file compliant with the DOM specifications. An 'at-job' consists of three parts: (i) properties (e.g. 'Height', 'Width', 'Font size') for all generated plots, (ii) path of the file where the batch log writes and (iii) a collection of 'jobs'. A 'job' is entitled as 'Session' and has three sets of parameters:

- Load and Save parameters, which define the full path of input files and the folder where to save the results; additionally, they enable/disable saving different ER sets to separate files.
- Analysis parameters, which set analysis properties such as T^s and T^w.
- BED parser parameters, which set properties such as *P*-value column number in input BED files to correctly load them.

A sample portion of an 'at-job' XML file is shown in Supplementary file 2. The 'at-job' is executed by a 'managed code' with least possible footprints, all being memory resources freed-up at 'job' execution termination. Hence, the 'Batch mode' memory requirement is limited to the amount needed for a single 'job' execution.

Determination of intersecting ERs

Cross-replicate, co-localized ERs shall be combined for overall significance determination of evidence; for each ER i of each sample j (i.e. r_{ji}), MuSERA combines the ERs in $R_{ji,*}$, i.e. the set of ERs in the replicates that intersect with r_{ji} (including r_{ji}), using the Fisher's method [10]. The set $R_{ji,*}$ can be determined using various efficient methods, such as algorithms based on ordered lists, i.e. by scanning all lists in parallel and linearly grouping ERs. However, the performance of such algorithms degrades when the intersection size is considerably smaller than the input size, or when input sizes vary significantly between the ER sets [23].

Algorithms based on variants of self-balancing binary search trees, such as interval trees [24] (i.e. an augmentation of red-black trees [25]) or segment trees [26], are asymptotically

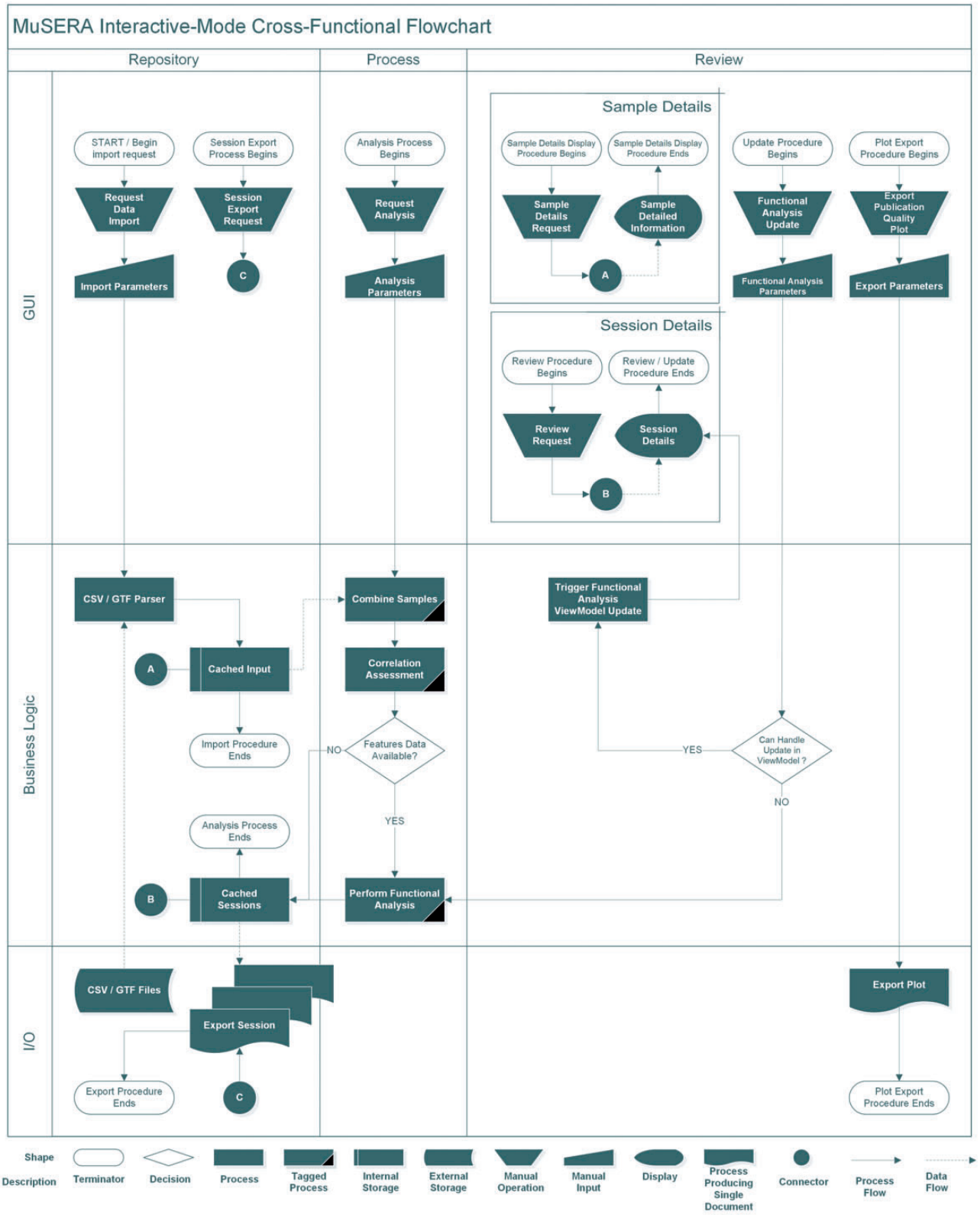


Figure 4. Cross-functional flowchart for the MuSERA Interactive mode. The flowchart shows a simplified flow of the major Interactive mode uses. In the Process part of the Business Logic section of the flowchart, the processes (rectangles) tagged with a black triangle in their bottom-right corner are time-consuming concurrent processes that allow executing other paths while MuSERA is busy computing them.

optimal data structures that store intervals and efficiently support queries for intervals overlapping a given interval/point. An ER is an interval on the genomic domain with respect to its 'chromosome', 'start' and 'end' attributes; this makes interval trees an appropriate data structure for the determination of $R_{ji,*}$ sets. Additionally, interval trees do not require the input to be sorted, which saves the time for sorting a possibly unsorted input.

MuSERA creates distinct interval trees, one for each chromosome of each replicate. The query time of an interval tree is order of $O(k\ Log_2\ n)$ for reporting k intervals when the tree holds n items. Therefore, $R_{ji,*}$ determination has $O(J\ k\ Log_2\ n)$ complexity, as it requires processing J distinct interval trees, each representing the same chromosome for one of the J replicates. Additionally, MuSERA processes chromosomes independently, and hence it is parallelized by distributing individual chromosome processes on available threads.

Genomic annotation of ERs

Once replicates are combined, MuSERA automatically annotates ERs with user-provided genomic features (e.g. genes, promoters, CDSs, binding sites of other TFs), independently for each of the ER sets (e.g. R_j^s, R_j^w, R_j^c, R_j^d). MuSERA is an interactive tool, where the user can tune a few parameters to achieve better results; hence, response time to update each annotation parameter should be reasonably fast. MuSERA can linearly group ERs and known binding sites/genomic annotations that overlap; however, this would require re-running the algorithm in case of any user-defined parameter is changed. To avoid this, the genomic annotation algorithm of MuSERA pre-processes data by defining genome-wide dynamic bins with coordinates determined by the ERs of the considered set (Figure 1A) and the known binding sites/genomic annotations (Figure 1B), the bins being stored and sorted according to their 'start' coordinate.

A bin spans a segment on the genome determined by two consecutive start/end coordinates of ERs or genomic annotations (i.e. start-start, start-end, end-start or end-end; see Bin_1, Bin_2, Bin_3 and Bin_5, respectively, in Figure 1B), and it includes all available information for that segment of DNA; hence, it enables constant access for the biological interpretation of the segment. This aspect avoids re-running the annotation process in case of changing any user-defined annotation parameter, such as the filter option (e.g. considering only TF binding sites or CDSs as known binding sites/genomic annotations). Additionally, given an ER, the corresponding DNA segments (i.e. bins) are determined in logarithmic time, because this requires a binary search on sorted elements (bins), and the element annotations are determined in constant time; therefore, an ER annotation is optimally computed in $O(Log_2\ n)$, where n is the number of defined bins.

Integrated genome browser

MuSERA implements also a flexible and highly interactive set of plotting features based on the Dynamic Data Display [27] package, allowing real-time interactive zoom and pan on genome-scale samples. Having combined samples, MuSERA automatically creates bins independently for each of the determined sets (e.g. R_j^s, R_j^w, R_j^c, R_j^d), as already shown in Figure 1, and displays in tabular format all the ERs of the sets with their corresponding information (e.g. 'chromosome', 'start', 'end', 'P-value', X^2). By double-clicking on any of the listed ERs, MuSERA plots it together with all the ERs (in different colours according to the set they belong, i.e. 'stringent confirmed', 'stringent discarded', 'weak confirmed' or 'weak discarded') and annotations, if any, within a window of user-defined size (e.g. see Figure 5); then, this can be easily scrolled, panned and zoomed to interactively explore the location on the DNA also of all the other ERs and annotations.

Use case results and practical guidance

In this section, we first illustrate how sets of significant ERs are expanded by applying MuSERA on replicates, for several types of NGS data, such as ERs from ChIP-seq of TFs and broad and narrow histone marks, and DNase-seq hypersensitive sites. We show that MuSERA is able to correctly determine a new set of ERs by locally combining their evidence on replicates, and we prove how the integrated graphical features that MuSERA provides well support thorough inspection of the obtained results and evaluation of their biological content.

Used data sets

We applied MuSERA on publicly available NGS data sets from the ENCODE repository, which always provides at least two biological replicates for each experiment [28]; we considered data sets regarding K562 (acute myelogenous leukaemia) human cells. To test MuSERA against a variety of different types of data and peak shapes, we decided to consider nine different data

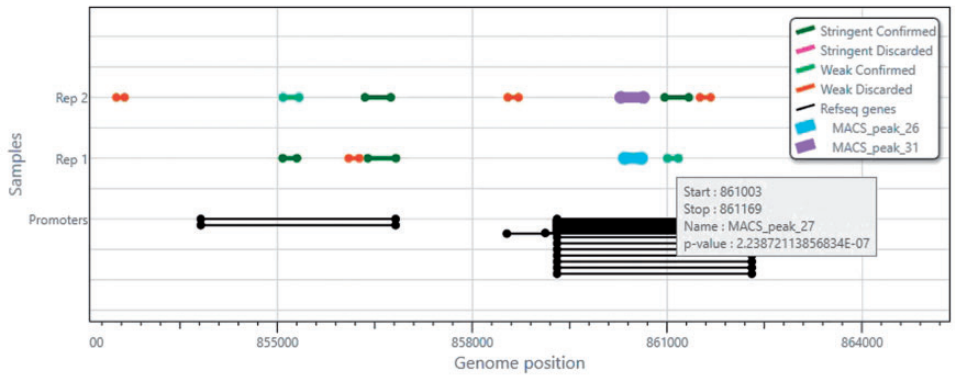


Figure 5. An example view of the integrated genome browser. For a selected ER (e.g. the ER represented by the light blue thick interval on Rep1 line, named MACS_peak_26), the ER(s) it is combined with (e.g. the ER represented by the purple thick interval on Rep2 line, named MACS_peak_31) and all surrounding ERs (coloured according to the set they belong) and available annotations are plotted; hovering the cursor on an ER, a tooltip is opened to show the corresponding information (e.g. start, stop, name, P-value).

sets: two ChIP-seq data sets of the TF CTCF (CTCF1, with three replicates, and CTCF2, with two replicates), one ChIP-seq data set of the TF JunD (JunD, two replicates), one ChIP-seq data set of the RNA Polymerase II molecule, responsible for gene transcription (Pol2, two replicates), one ChIP-seq data set of the histone mark H3K4me3, marking active promoters (H3K4me3, two replicates), which usually generates narrow, TF-like peaks, one DNase-seq data set, corresponding to open chromatin regions (DNaseI, two replicates) and three data sets of histone marks, which are deposited over large genomic regions (H3K9me3 and H3K27me3, two replicates each, marking the body of repressed genes; H3K36me3, two replicates, marking actively transcribed gene bodies). The details of the samples are given in Table 1.

Results of combining ER evidence on replicates and their validation

MuSERA has been run with parameters $T^s = 10^{-8}$, $T^w = 10^{-4}$, $C = 1$ in the 'biological replicate' mode, meaning that the validation of an ER overlapping with ERs in the other replicate samples is sufficient to validate all overlapping ERs. In particular, with the choice $C = 1$ we decided to automatically confirm each stringent peak, regardless of its overlap with peaks in the other replicates, as we found this was the best strategy for the TF Myc [8].

As we can see from Figure 6 (panels A–I), showing the sizes of the different ER sets determined, by combining ER evidence on replicates MuSERA allows the 'rescue' of a large number of peaks ($R^{w,c}$ set, dark green) below the chosen significance threshold T^s in a single sample. The number of 'rescued' (i.e. weak, confirmed) peaks in the output set R^o ranges from 12% to more than the double of the size of the original single-sample stringent set R^s (panel J): the presence of a biological replicate allows a consistent expansion of the set of 'good' peaks by locally lowering the sensitivity threshold. The highest efficiency is found for the Pol2, DNaseI, H3K27me3 and H3K36me3 samples, where the sample output set R^o has more than double (up to more than triple for H3K27me3) of the peaks in the stringent set R^s.

In all determined ER sets of the CTCF samples, we validated our results by looking for the presence of the CTCF motif (coded as a Position Weight Matrix in the JASPAR CORE Vertebrata database entry MA0139.1 [29]) recognized on the genome in the sequences spanned by the peaks in the different sets. We found the motif enriched in all the CTCF R^s and R^o sets, as expected, but also in all the $R^{w,c}$ and in two of five $R^{w,d}$ sets, even if the P-values of the enrichments are higher (i.e. less significant) in the $R^{w,d}$ set case. This result fully validates the 'rescue' process proposed by MuSERA, and also suggests that our peak call has been rather stringent. The details of the validation results are shown in Table 2.

Use case result evaluation with MuSERA graphical features

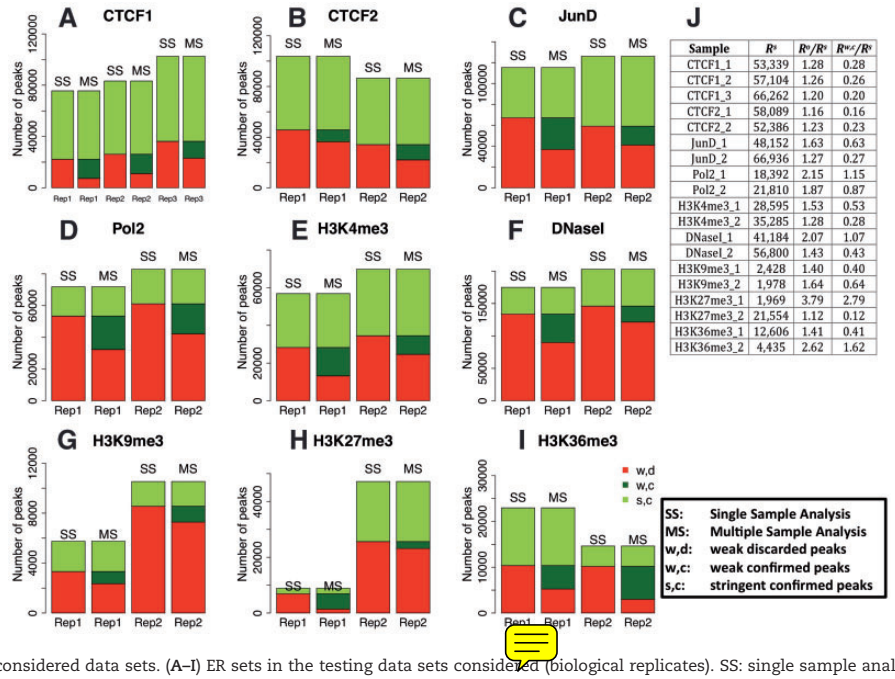
Through its graphical interface, MuSERA allows a quick inspection of the analysis results and a thorough evaluation of their biological content. Figure 7 shows the MuSERA 'Overview' panel providing a general overview of the ER sets determined for the CTCF1_1 sample, and including a global view of the parameter values used. All ERs of each set are listed in a table view, together with all their quantitative values and computed statistics; with just a double click, the ERs can be easily displayed in the genomic context along the DNA, thanks to the MuSERA integrated genome browser (Figure 5).

Furthermore, several other quantitative features that MuSERA automatically computes can be straightforwardly displayed; some of them are shown in Figure 8: the stratification of the ER sets over the different chromosomes (panel A), the distribution of the combined significance (X^2) of the ERs in each set (panel B, Output Set of the CTCF1_1 sample) and the distribution of the distance of the ERs in each set from the closest genomic feature chosen (panel C, Output Set of the CTCF1_1 sample; the

Table 1. ENCODE alignment files used and their quantitative features

Sample name	Short name	Aligned reads	R^s	R^w
wgEncodeOpenChromChipK562CtcfAlnRep1	CTCF1_1	6 051 439	53 339	22 290
wgEncodeOpenChromChipK562CtcfAlnRep2	CTCF1_2	6 211 475	57 104	26 177
wgEncodeOpenChromChipK562CtcfAlnRep3	CTCF1_3	11 988 569	66 262	36 278
wgEncodeSydhTfbsK562CtcfbIggrabAlnRep1	CTCF2_1	26 957 114	58 089	45 727
wgEncodeSydhTfbsK562CtcfbIggrabAlnRep2	CTCF2_2	26 437 775	52 386	34 130
wgEncodeSydhTfbsK562JundbIggrabAlnRep1	JunD_1	16 175 565	48 152	67 154
wgEncodeSydhTfbsK562JundbIggrabAlnRep2	JunD_2	28 086 672	66 936	59 105
wgEncodeSydhTfbsK562Pol2IggmusAlnRep1	Pol2_1	17 762 352	18 392	53 489
wgEncodeSydhTfbsK562Pol2IggmusAlnRep2	Pol2_2	19 293 573	21 810	61 160
wgEncodeBroadHistoneK562H3k4me3StdAlnRep1	H3K4me3_1	9 512 593	28 595	28 271
wgEncodeBroadHistoneK562H3k4me3StdAlnRep2	H3K4me3_2	15 640 462	35 285	34 526
wgEncodeOpenChromDnaseK562AlnRep1	DNaseI_1	9 993 542	41 184	133 829
wgEncodeOpenChromDnaseK562AlnRep2	DNaseI_2	29 472 357	56 800	146 113
wgEncodeBroadHistoneK562H3k9me3StdAlnRep1	H3K9me3_1	15 816 227	2428	3324
wgEncodeBroadHistoneK562H3k9me3StdAlnRep2	H3K9me3_2	33 939 687	1978	8555
wgEncodeBroadHistoneK562H3k27me3StdAlnRep1	H3K27me3_1	12 210 065	1969	6916
wgEncodeBroadHistoneK562H3k27me3StdAlnRep2	H3K27me3_2	12 119 288	21 554	25 603
wgEncodeBroadHistoneK562H3k36me3StdAlnRep1	H3K36me3_1	14 803 144	12 606	10 365
wgEncodeBroadHistoneK562H3k36me3StdAlnRep2	H3K36me3_2	10 393 298	4435	10 189

Peaks were called with the software package MACS2.0 [4] using the parameters '-auto-bimodal -p 0.01 -g hs' (thus setting a P-value threshold of 10^{-2}). R^s: stringent ER set (ERs with P-value $< T^s$). R^w: weak ER set (ERs with $T^s \leq P$-value $< T^w$). $T^s = 10^{-8}$, $T^w = 10^{-4}$. Peaks for the histone marks H3K4me3, H3K9me3, H3K27me3 and H3K36me3 were called with the 'broad' option.



J

Sample	R^s	R^o/R^s	$R^{w,c}/R^s$
CTCF1_1	53,339	1.28	0.28
CTCF1_2	57,104	1.26	0.26
CTCF1_3	66,262	1.20	0.20
CTCF2_1	58,089	1.16	0.16
CTCF2_2	52,386	1.23	0.23
JunD_1	48,152	1.63	0.63
JunD_2	66,936	1.27	0.27
Pol2_1	18,392	2.15	1.15
Pol2_2	21,810	1.87	0.87
H3K4me3_1	28,595	1.53	0.53
H3K4me3_2	35,285	1.28	0.28
DNaseI_1	41,184	2.07	1.07
DNaseI_2	56,800	1.43	0.43
H3K9me3_1	2,428	1.40	0.40
H3K9me3_2	1,978	1.64	0.64
H3K27me3_1	1,969	3.79	2.79
H3K27me3_2	21,554	1.12	0.12
H3K36me3_1	12,606	1.41	0.41
H3K36me3_2	4,435	2.62	1.62

Figure 6. ER sets for the considered data sets. (**A–I**) ER sets in the testing data sets considered (biological replicates). SS: single sample analysis; MS: multiple sample analysis. In each panel, the SS stacked bars represent R^s (light green/gray) and R^w (green/gray) sets in the replicates, while the MS bars show the same peaks, confirmed or discarded according to the MuSERA output: $R^{s,c}$ (light green/gray), $R^{w,c}$ (dark green/gray) and $R^{w,d}$ (red/black) sets. Note that setting the parameter $C = 1$, the $R^{s,d}$ set is always empty. (**J**) General statistics on the cardinality of the ER sets. See Table 2 for the validation results of the CTCF peaks.

Table 2. Validation results of the peaks for the CTCF samples

Sample	R^s	R^o	$R^{w,c}$	$R^{w,d}$
CTCF1_1	7.4 e-3575	3.4 e-3786	5.2 e-338	–
CTCF1_2	1.6 e-3586	2.2 e-3846	4.1 e-147	–
CTCF1_3	8.6 e-4115	3.8 e-4321	6.9 e-199	2.4 e-128
CTCF2_1	2.6 e-3610	5.2 e-3676	2.7 e-116	9.4 e-67
CTCF2_2	7.5 e-3414	1.0 e-3517	2.8 e-267	–

P-values for the enrichment of the CTCF binding motif (JASPAR CORE Vertebrata database entry MA0139.1), as estimated with the DREME package [30] in the 150 bp around the peak midpoint. R^s: stringent ER set. R^o: output ER set. $R^{w,c}$: weak, confirmed ER set. $R^{w,d}$: weak, discarded ER set. '-': no enriched motif.

chosen genomic feature is the set of promoters in the human genome hg19).

The 'genomic annotation and functional analysis of enriched regions' and the 'nearest enriched region distance distribution' that MuSERA supports can provide better understanding and improved biological interpretation of the obtained results. For example, looking at the peak-to-peak distance across the different samples (Figure 9, panel A: narrow ERs; panel B: broad ERs), which MuSERA automatically quantifies to build the 'nearest enriched region distance distributions', we can see that this quantity is higher in weak confirmed peaks than in stringent confirmed peaks for the considered samples of the CTCF TF and H3K4me3 and H3K9me3 histone marks; whereas, this distance is roughly the same for the DNase I Hypersensitive Site (DHS), RNA Polymerase II and for some of the broad histone mark samples considered. This probably depends on the fact that the stringent confirmed CTCF peaks correspond to high-affinity binding sites of the TF, while the CTCF weak peaks could be generated by transient interactions with the DNA, which are not stabilized by a specific target, and therefore are scattered across the genome. A similar argument holds for the H3K4me3 and H3K9me3 histone marks, although in this case the strength of the signal is just an indication of the fraction of cells bearing the modification, and it is more difficult to identify the mechanism responsible for this difference; a good guess is that it could be related to the local balance of the enzymes transferring and removing the methyl groups to the histone proteins. On the other hand, DHS ERs are found throughout the genome and do not have preferred genomic locations where the signal is stronger; therefore, in this case the peak-to-peak distance distribution is similar for strong and weak peaks. The mixed behaviour of sample H3K27me3 may depend on the large differences in the number of peaks across the two replicates: replicate 1 almost quadruplicates its number of ERs in the R^o set, thanks to the high number of ERs in replicate 2, and probably the few stringent confirmed peaks were more scattered around the genome than the many weak confirmed peaks. A similar case, although with lesser intensity, may hold true for replicate 2 in sample H3K36me3.

The case of RNA Polymerase II is rather surprising: RNA Polymerase II is the molecule transcribing the genome, and it is mostly localized on genes and promoters, although recent studies indicate that most of the genome has the potentiality of being transcribed [31]. To gain a better insight on this aspect, we took advantage of the 'genomic annotation and functional analysis' available in MuSERA to inspect the genomic location of the RNA Polymerase II peaks. Using the 'ER-to-feature overlap score' that MuSERA automatically calculates when promoters, intragenic regions or IGR are selected as genomic features, respectively, we found that the fraction of RNA Polymerase II peaks located on these regions is unchanged in the stringent and weak ER sets and across the replicates (Figure 10). Thus, the features that MuSERA computes and graphically shows enabled us to conclude that RNA Polymerase II binds with a wide range of intensities to both genes and intergenic elements.

Finally, the 'global correlation assessment' provided by MuSERA, through the evaluation of the JSC for each type of ER set determined, confirms that the obtained output sets are

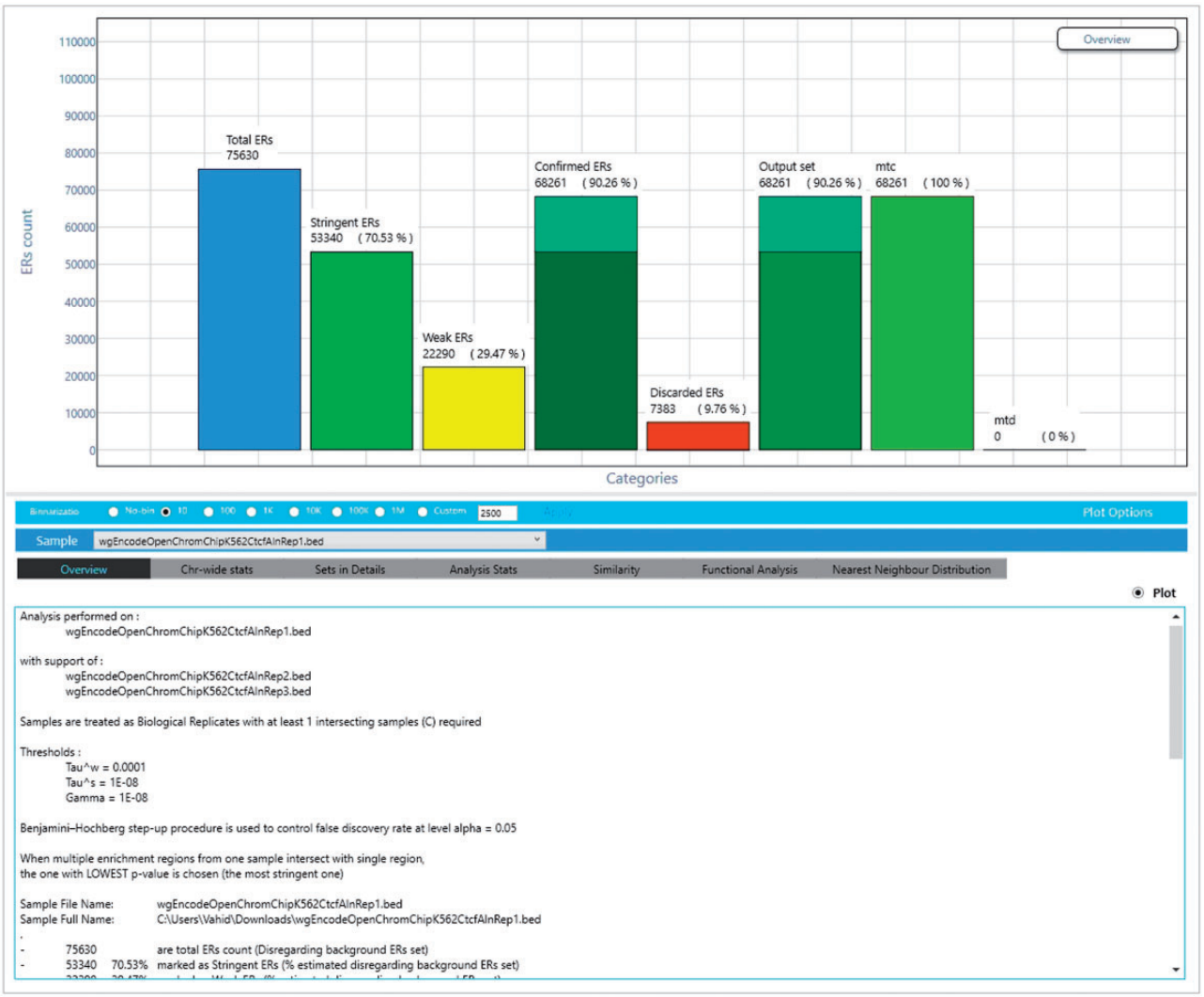


Figure 7. Overview panel of the MuSERA graphical interface. After an analysis is performed, MuSERA shows the statistics of the analysed sets in the 'Overview' panel. Data shown regard the CTCF1_1 sample.

much more congruent than the stringent sets, which would have been used as outputs in the absence of MuSERA (Figure 11, left panel); besides, even for weak ERs, which may include a higher fraction of spurious binding sites, the JSC value increases considerably for the weak confirmed sets, confirming the validity of the 'rescue' process that MuSERA performs (Figure 11, right panel). The evidence in the replicates of a data set is therefore combined in sets of ERs that are more coherent between themselves than the outputs of single-sample analyses.

Performance

We benchmarked the MuSERA performance for a variety of operations, including loading data, combining replicates and preprocessing of analysis results for further assessments through e.g. genomic annotation, similarity search or integrated genome browser. Tests were performed on a standard laptop computer running Microsoft Windows® 10, with Intel® Core™ i3 (2.10 GHz) CPU and 6 GB of RAM. The benchmark was performed on multiple ENCODE ChIP-seq and DNase-seq data sets regarding K562 human cells, including two to three replicates each, where the overall number of ERs in the replicate samples of each data set spanned few thousands to millions of ERs. Additionally, a data set of human genome hg19 promoters (counting 82 960 promoter regions) was imported for genomic annotation performance benchmarking.

In general, MuSERA performance is in the scale of seconds, spanning few tens to hundreds of seconds depending on operation and number of ERs on replicates, from few tens of thousands to millions of ERs (Figure 12). The process of parsing and loading ERs from input sample files runs in a handful of seconds for most samples. The algorithm of combing replicates is highly optimized and runs in few tens of seconds for two to three replicates with a few hundreds of thousands of ERs each. The correlation between replicates is assessed once replicates are combined; the algorithm runs instantaneously (hence it is not explicitly included in Figure 12). Some operations (e.g. nearest neighbour search, genomic annotation and genome browsing) depend on a data structure that is automatically populated once an analysis session is selected. Such process is executed in background to minimize its effect on other MuSERA independent operations and maximize user experience (i.e. the user can benefit the other independent features of MuSERA while the

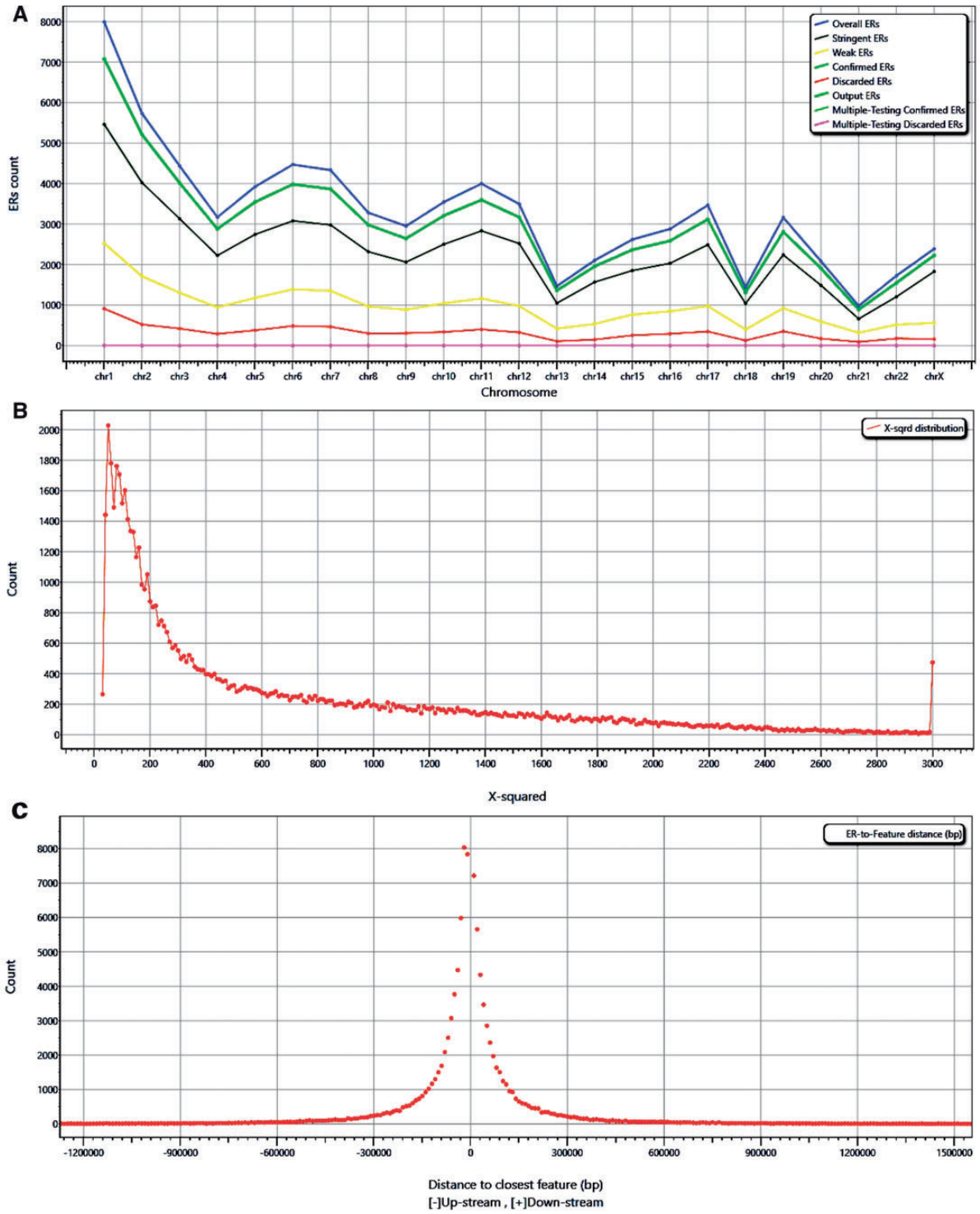


Figure 8. Some graphical analyses performed by MuSERA. The different panels of MuSERA plot the features computed in the analysis. For example, we show here (**A**) elements in the sets, stratified by chromosomes; (**B**) distribution of the combined significance (X^2) of the output set (R^o) ERs for the CTCF1_1 sample; (**C**) distance of ERs in the output set (R^o) of the CTCF1_1 sample from human promoters: clearly, the CTCF TF prefers to bind the DNA close to the regulatory regions of a gene.

required data structure is being populated in background); the process completes in few tens of seconds, depending on the number of considered ERs and genomic features. Once the data structure is populated, the operations, such as genome browsing, are instantaneous.

Discussion and conclusions

We reviewed MuSERA, an effective, efficient and easy-to-use graphical tool of broad utility to combine evidence across ChIP-seq or DNase-seq replicates, and to evaluate them and their

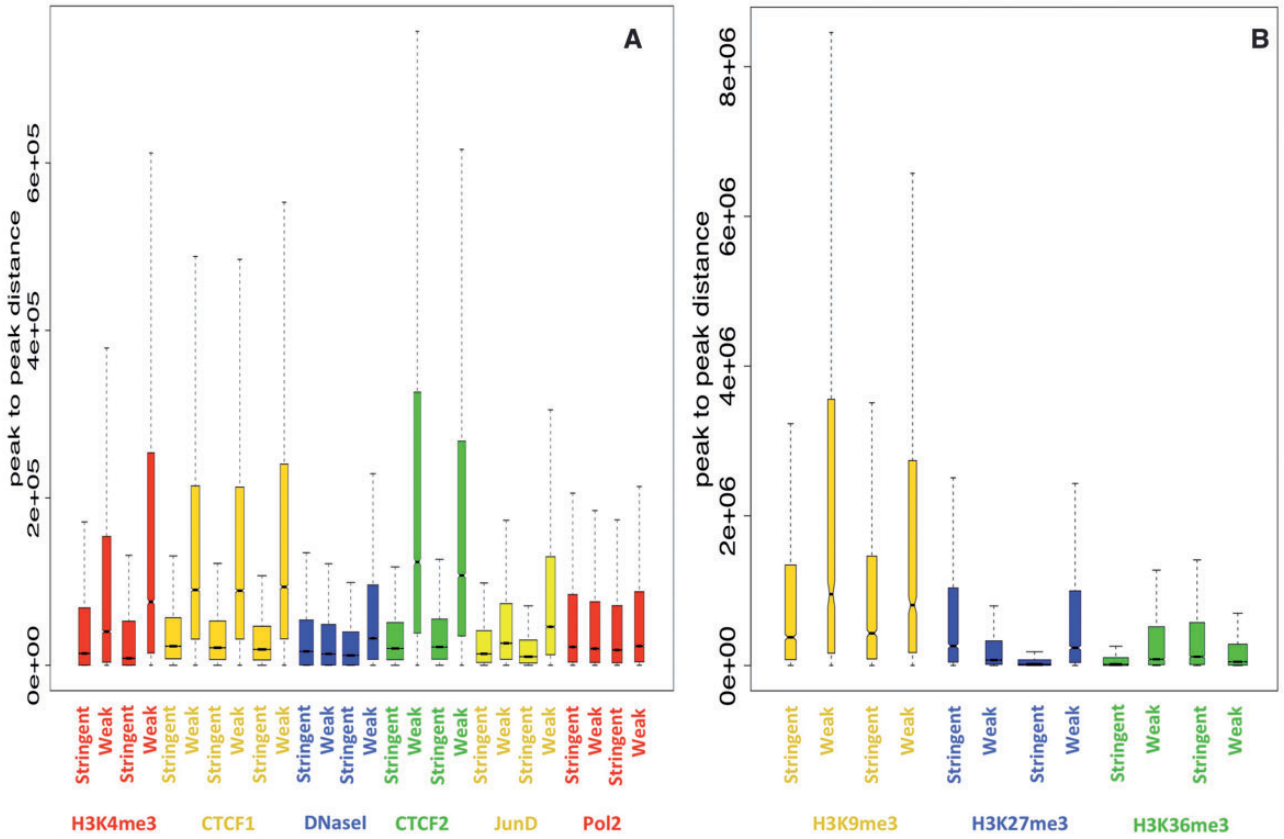


Figure 9. Peak-to-peak distance. The boxplots represent the peak-to-peak distance for the stringent confirmed and weak confirmed ER sets for all the samples considered. Samples displaying narrow, Gaussian-like peaks (**A**) are shown separated from samples having ERs with a broader shape (**B**). While this distance is on average greater in the weak confirmed ER sets for the considered TFs (CTCF and JunD) and the H3K4me3 and H3K9me3 histone marks, it stays roughly constant in the two sets for DHS and RNA Polymerase II, and in the remaining histone marks.

biological relevance in the genomic context. MuSERA allows the annotation of samples with user-defined genomic features and the visualization of results in an integrated genome browser. Furthermore, it provides a rich set of quantitative evaluations and interactive graphical displays, which greatly help the understanding and biological interpretations of results.

Common tools used to analyse NGS data are usually designed for scientists with training in bioinformatics or other quantitative disciplines, as they usually involve command-line interfaces and heavily rely on extensive coding abilities. This naturally poses a barrier against biologists who generated the data, and would like to directly perform simple analyses on them. Only few tools make use of a GUI to reach out to larger audiences, the Galaxy project being the most prominent example [32, 33]. However, this large, all-purpose tool can become rather complex to use despite the presence of a GUI, and usually requires powerful computing facilities to run analysis applications on NGS data files, which are typically large. MuSERA, on the other hand, is a dedicated tool efficiently performing integrated analysis of replicated NGS data sets involving ERs, which can be directly used on any personal computer and mastered in a short time. Some tools, like Nebula [34], provide a more focused GUI centred on the processing of ChIP-seq data, and yet, they do not consider the presence of replicates. This same and relevant limitation generally applies to the available tools commonly used for NGS data evaluation, including GenometricCorr [35], which is focused on the detection of genome-wide correlations between pairs of samples; it includes four different methods to compute these correlations ('relative distance', 'absolute distance', 'projection' and 'Jaccard'), together with appropriate null models and statistical tests to evaluate the significance of the correlations. We note that the 'ER-to-feature overlap score' implemented in MuSERA can be thought of a specific case of the 'absolute distance' method implemented in GenometricCorr, where all distances >0 (i.e. considering only non-overlapping regions in the pair of samples considered) are discarded. Some other tools, like PAPST [36], focus on co-localization of different types of ERs, but do not consider the significance of the ERs in their analysis. Instead, MuSERA uniquely combines a rigorous approach for jointly evaluating ERs in replicates [8] with an intuitive GUI and an array of useful downstream analyses, both computational and graphical. Moreover, it leverages on high-end data structures to minimize the runtime of common analysis procedures, and executes time-consuming operations in the background, resulting in high user-friendly interaction with minimal lag. Additionally, while batch processing on common tools requires scripting and/or coding knowledge, MuSERA facilitates batch execution specification by providing a simple XML structure to define batch jobs.

We applied MuSERA to ChIP-seq data sets of TFs and histone marks, and to a DNase-seq data set, and we found that the efficiency of the 'rescue' of weak ERs varies between 12% and 279%, thus potentially making a big impact on the final list of sample-specific confirmed ERs. Variation of MuSERA efficacy depends on many factors, including the quality of replicates and the

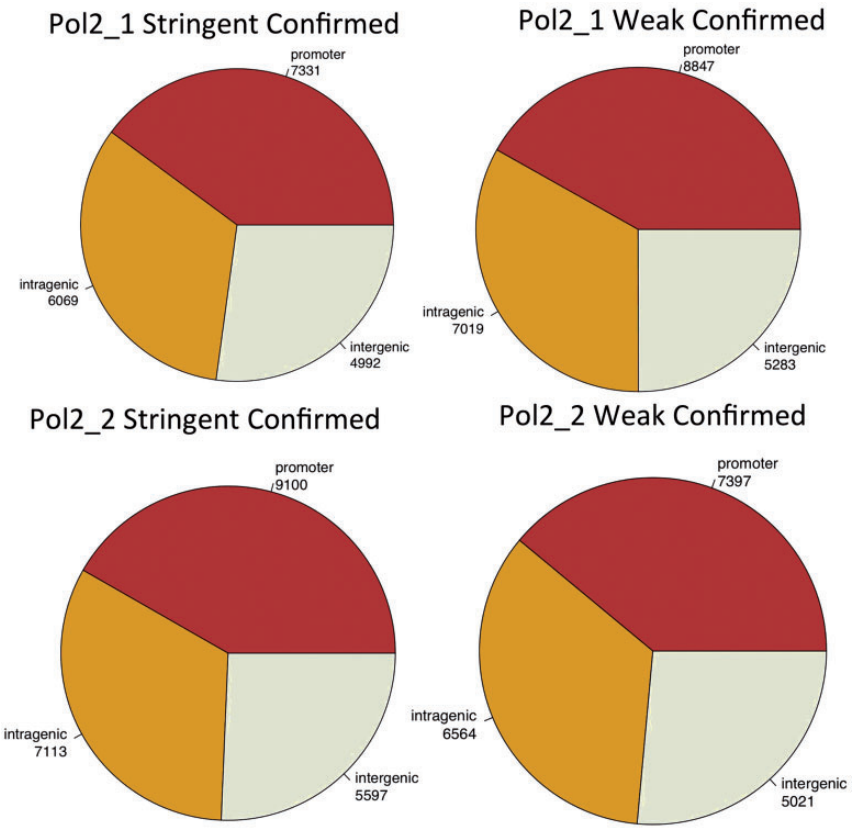


Figure 10. Distribution of RNA Polymerase II ERs in the genome. RNA Polymerase II (Pol2) ERs fall mostly around genes (promoter, intragenic), but a considerable fraction is located in intergenic regions. This behaviour is highly conserved across the two replicates considered and across stringent confirmed and weak confirmed ERs.

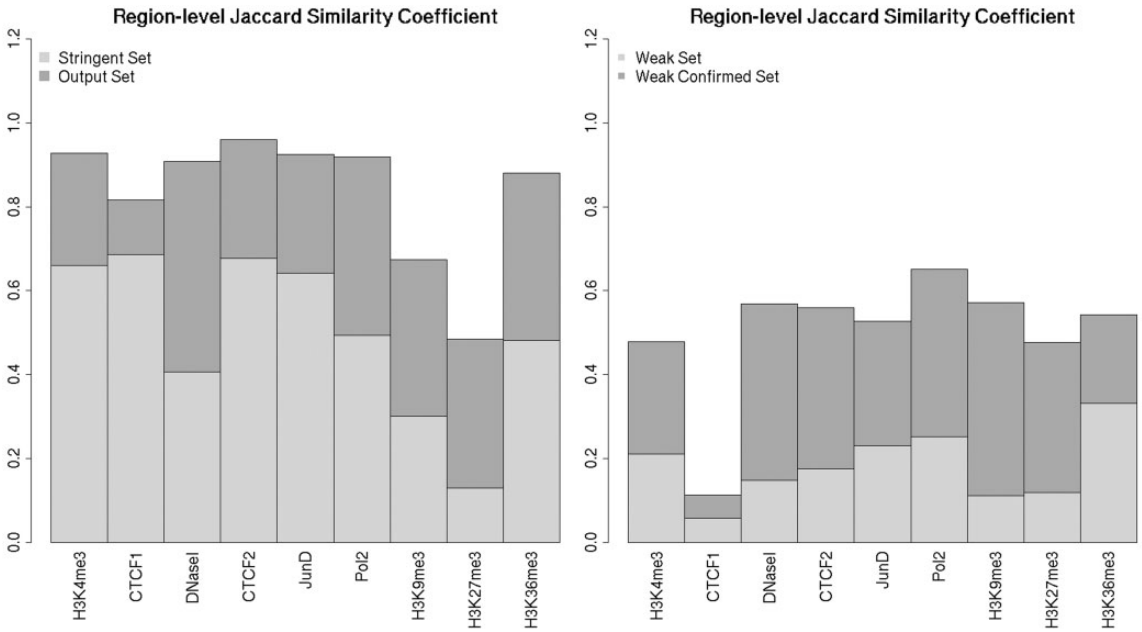


Figure 11. Region-level Jaccard Similarity Coefficient (JSC). The JSC measures the similarity between two or more sets, and it is automatically computed by MuSERA for any type of ER set determined. The figure shows that, for each of the data sets considered, the similarity between the output sets is much higher than the similarity between the stringent sets (left panel). For the sets of weak ERs (right panel), which usually contain a higher fraction of binding sites, the JSC value is rather low, but it considerably increases for the weak confirmed sets, supporting the validity of the evidence-combining process that MuSERA performs. Finally, we note that the CTCF1 data set, which has three replicates, has a lower JSC value owing to the evaluation of an additional sample in the overall ER overlaps.

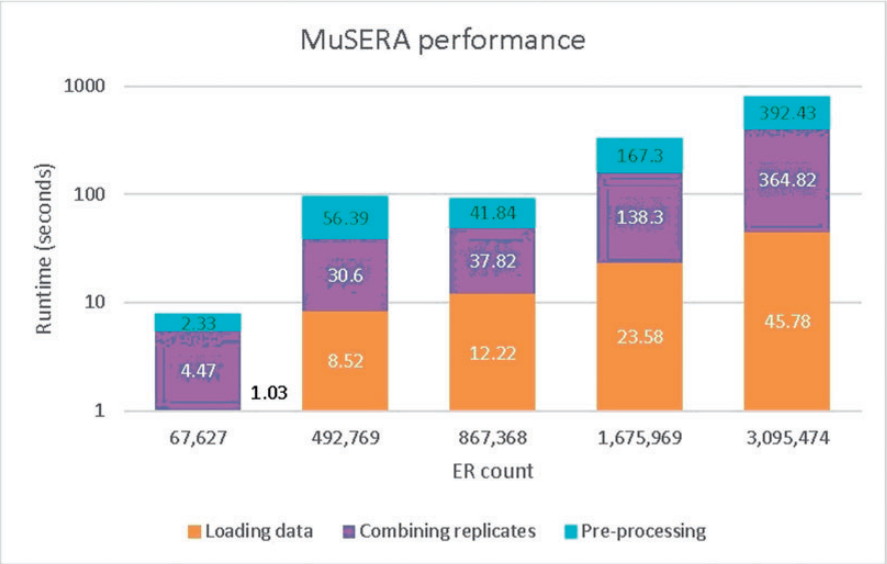


Figure 12. Benchmarking of MuSERA main operations. Operation runtime, on a logarithmic scale, for increasing number of ERs in combined replicates of ENCODE ChIP-seq or DNase-seq data sets (two to three replicates for each data set). The data sets considered were downloaded from ENCODE; for the ER counts in the figure they were, from left to right: wgEncodeSydhHistoneK562bH3k4me3bUcdAlnRep1/2, wgEncodeUwDnaseK562Znf4g7d3AlnRep1/2, wgEncodeUwDnaseK562Znf2c10c5 AlnRep1/2, wgEncodeOpenChromDnaseK562NabutAlnRep1/2 and wgEncodeOpenChromDnaseK562G1phaseAlnRep1/2/3.

biological characteristics of the ERs. For example, for the TF CTCF we observed the lowest rate of rescue of weak peaks among all the replicates: this TF makes contact with the DNA through 11 distinct zinc-finger domains [37], and therefore binds in an extremely strong way. In this case, most of these interactions correspond to a clear signal in the ChIP-seq experiment, and the corresponding ERs are inevitably classified as stringent. Therefore, for this particular TF, the replicates are more coherent than usual, and most of the weak interactions are classified as noise. On the opposite, DNaseI hypersensitive sites have been shown to display a continuum of intensities, which does not saturate even at high sequencing depths [38]; therefore, for these experiments, the border between weak and stringent ERs is somewhat arbitrary and many of the ERs classified as weak in a single sample correspond to true open chromatin regions, consistently observed across replicates. In this case, MuSERA is particularly successful in expanding the set of the confirmed ERs.

The integrated genome browser and the several graphical features that MuSERA offers for genomic annotation and functional analysis of ERs, nearest ER distance distribution and global correlation assessment of ERs proved useful for the evaluation and biological interpretation of the obtained ERs within the genomic context, and could be the starting point of deeper functional analyses based on more refined measures, as currently implemented in other tools [9, 35, 36]. Moreover, the method that MuSERA implements to obtain the ERs was proved, with respect to other approaches, to optimally address the specific task of combining evidence over replicates [8]. Its output, designed to allow quick pipelining to downstream analyses, provides both sample-specific BED files of the different ER sets determined, and a single BED file unifying the significant confirmed ERs present in the combined replicate samples; all these files can be directly analysed with common tools like BEDTools [39], BEDOPS [40] or Bioconductor [41]. In addition, the overview XML files generated give all the details about the performed combination of multiple evidence across replicates, and allow tracking down the individual overlapping events among ERs. All this makes MuSERA a tool likely to be of broad utility that represents a significant advance over previously published software.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Replicates in next-generation sequencing experiments are recommended, but their full potential, especially in experiments involving the identification of enriched regions (ER), is often neglected.
- MuSERA is a tool that allows combining local evidence in replicates to improve ER calling, and provides quantitative evaluations and graphical features to assess the biological relevance of each determined ER set within its genomic context; they include genomic annotation of determined ERs, nearest ER distance distribution and global correlation assessment of ERs.
- MuSERA comes with an intuitive graphical user interface, making it immediate to use, which provides an integrated genome browser and an array of graphical displays that greatly support understanding and biological interpretations of the results.
- By applying MuSERA to different data types, including ChIP-seq of transcription factors or histone marks and DNase-seq hypersensitive sites, we always found enhanced sets of ERs and proved its effective support in the inspection of obtained results and evaluation of their biological content.
- MuSERA represents a significant advance over previously published software, as we discuss and comparatively demonstrate.

Funding

This work was supported by the Fondazione Istituto Italiano di Tecnologia and by AIRC [IG_13182] and the Italian Ministry of the University and Research (MIUR) ['Data-Driven Genomic Computing (GenData 2020)' PRIN project (2013-2015)].

References

1. van Dijk E, Auger H, Jaszczyszyn Y, *et al*. Ten years of next-generation sequencing technology. *Trends Genet* 2014;**30**(9):418–26.
2. Park P. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Gen* 2009;**10**:669–80.
3. Cockerill P. Structure and function of active chromatin and Dnase I hypersensitive sites. *FEBS J* 2011;**278**:2182–210.
4. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**(9):R137.
5. Rashid NU, Giresi PG, Ibrahim JG, *et al*. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* 2011;**12**(7):R67.
6. Chen Y, Negre N, Li Q, *et al*. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat Methods* 2012;**9**(6):609–14.
7. Li Q, Brown JB, Huang H, *et al*. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 2011;**5**(3): 1752–79.
8. Jalili V, Matteucci M, Masseroli M, *et al*. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* 2015;**31**(17):2761–9.
9. Zeng X, Sanalkumar R, Bresnick EH, *et al*. jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol* 2013;**14**(4):R38.
10. Fisher RA. *Statistical Methods for Research Workers*. Guildford, UK: Genesis Publications, Ltd, 1925.
11. Bulger M, Groudine M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol* 2010;**339**(2):250–7.
12. Lettice LA, Heaney SJ, Purdie LA, *et al*. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* 2003;**12**(14):1725–35.
13. Glassford WJ, Rebeiz M. Assessing constraints on the path of regulatory sequence evolution. *Philos Trans R Soc Lond B Biol Sci* 2013;**368**(1632):20130026.
14. MacQuarrie KL, Fong AP, Morse RH, *et al*. Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet* 2011;**27**(4):141–8.
15. Pearson K. Note on regression and inheritance in the case of two parents. *Proc R Soc Lond* 1895;**58**(347-352):240–2.
16. Bardet AF, He Q, Zeitlinger J, *et al*. A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc* 2012;**7**(1):45–61.
17. Zhao X, Valen E, Parker BJ, *et al*. Systematic clustering of transcription start site landscapes. *PLoS One* 2011;**6**(8):e23409.
18. Rau A, Gallopin M, Celeux G, *et al*. Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* 2013;**29**(17):2146–52.
19. Giannopoulou EG, Elemento O. An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics* 2011;**12**(1):277.
20. Ashoor H, Hérault A, Kamoun A, *et al*. HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics* 2013;**29**(23):2979–86.
21. Smith J. WPF apps with the Model-View-ViewModel design pattern. *MSDN Magazine* 2009;**24**(2):dd419663.
22. Visual Studio Code metrics values. 2015. https://msdn.microsoft.com/en-us/library/bb385914.aspx (30 January 2016, date last accessed).
23. Hwang FK, Lin S. A simple algorithm for merging two disjoint linearly ordered sets. *SIAM J Comput* 1972;31–9.
24. Cormen TH, Leiserson CE, Rivest RL, *et al*. Section 14.3: interval trees. In: *Introduction to algorithms*. 3rd edn. Cambridge, MA: MIT Press and McGraw-Hill, 2009, p. 348354.
25. Bayer R. Symmetric binary B-trees: data structure and maintenance algorithms. *Acta Inform* 1972;**1**(4):290–306.
26. Bentley JL. *Solutions to Klee's rectangle problems*. Pittsburgh, PA: Carnegie-Mellon University, 1977.
27. CodePlex. Dynamic Data Display. 2011. http://dynamicdatadisplay.codeplex.com/ (30 January 2016, date last accessed).
28. Landt SG, Marinov GK, Kundaje A, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;**22**(9):1813–31.
29. Mathelier A, Zhao X, Zhang AW, *et al*. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2014;**42**: D142–7.
30. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 2011;**27**(12):1653–9.
31. Djebali S, Davis CA, Merkel A, *et al*. Landscape of transcription in human cells. *Nature* 2012;**489**(7414):101–8.
32. Blankenberg D, Taylor J, Schenk I, *et al*. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Biol* 2007;**17**(6):960–4.
33. Boekel J, Chilton JM, Cooke IR, *et al*. Multi-omic data analysis using Galaxy. *Nat Biotechnol* 2015;**33**:137–9.
34. Boeva V, Lermine A, Barette C, *et al*. Nebula – a web server for advanced ChIP-seq data analysis. *Bioinformatics* 2012;**28**(19):2517–19.
35. Favorov A, Mularoni L, Cope LM, *et al*. Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol* 2012;**8**(5):e1002529.
36. Bible PW, Kanno Y, Wei L, *et al*. PAPST, a user friendly and powerful Java platform for ChIP-seq peak co-localization analysis and beyond. *PLoS One* 2015;**10**(5):e0127285.
37. Phillips JE, Corces VG. CTCF: master weaver of the genome. *Cell* 2009;**137**(7):1194–211.
38. Neph S, Viestra J, Stergachis AB, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 2012;**489**(7414):83–90.
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**(6): 841–2.
40. Neph S, Kuehn MS, Reynolds AP, *et al*. BEDOPS: high-performance genomic feature operations. *Bioinformatics* 2012;**28**(14):1919–20.
41. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.