scientific reports

Check for updates

OPEN

Productive visualization of high-throughput sequencing data using the SeqCode open portable platform

Enrique Blanco[1]✉, Mar González-Ramírez[1] & Luciano Di Croce[1,2,3]✉

Large-scale sequencing techniques to chart genomes are entirely consolidated. Stable computational methods to perform primary tasks such as quality control, read mapping, peak calling, and counting are likewise available. However, there is a lack of uniform standards for graphical data mining, which is also of central importance. To fill this gap, we developed SeqCode, an open suite of applications that analyzes sequencing data in an elegant but efficient manner. Our software is a portable resource written in ANSI C that can be expected to work for almost all genomes in any computational configuration. Furthermore, we offer a user-friendly front-end web server that integrates SeqCode functions with other graphical analysis tools. Our analysis and visualization toolkit represents a significant improvement in terms of performance and usability as compare to other existing programs. Thus, SeqCode has the potential to become a key multipurpose instrument for high-throughput professional analysis; further, it provides an extremely useful open educational platform for the world-wide scientific community. SeqCode website is hosted at http://ldicrocelab.crg.eu, and the source code is freely distributed at https://github.com/eblancoga/seqcode.

Current high-throughput sequencing techniques (e.g. ChIP-seq, ATAC-seq, and RNA-seq) can use a single run to identify the repertoire of functional characteristics of the genome. Therefore, an accurate interpretation of the results is fundamental to understand how the transcriptomic and epigenomic landscape of cells evolve during developmental and/or disease stages[1–3]. Powerful bioinformatic tools are available to manage this volume of data at a primary stage: (i) quality control profilers evaluate distinct scoring metrics on raw information[4–6]; (ii) mapping algorithms identify the location of each read on the genome[7–9]; (iii) peak callers find clusters of reads significantly enriched in certain genomic regions in the sample map file[10–12]; (iv) genome browsers are useful to visualize genome-wide binding profiles and peaks[13–16]; and (v) other auxiliary applications convert intermediate files into the appropriate data formats[17–20].

However, further work is necessary to finally extract useful knowledge from processed data (such as meta-plots, heatmaps, feature charts, boxplots of signal intensity, etc.). Surprisingly, only moderate interest has been shown until now for standardizing the design of this kind of graphical information. As a result, researchers in many cases develop their own scripts to address these issues, which is likely to compromise reproducibility and comparisons. In parallel, the reduction in the cost of the next-generation sequencing (NGS) techniques has dramatically increased the number of high-throughput sequencing experiments that are produced every year, making it urgent to find methods that efficiently process this information[21].

Here, we first illustrate the main characteristics of SeqCode, introduce the collection of principal SeqCode features to perform high-quality graphical analysis of sequencing data, and propose a standardized nomenclature of representations. SeqCode is entirely focused on the graphical analysis of 1D genomic data (e.g. ChIP-seq, RNA-seq). For information regarding visualization of 3D data (such as Hi-C interactions), we refer the reader to existing comprehensive reviews[22–25]. Next, we discuss our web portal, which offers a user-friendly interface for SeqCode functions to generate images from NGS data sets. From technical and biological point of views, we then assess the soundness of SeqCode results in several realistic scenarios. Finally, we comprehensively review the existing literature on similar tools to evaluate our software in comparison to current approaches.

[1]Centre for Genomic Regulation (CRG), Barcelona Institute for Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain. [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3]ICREA, Passeig Lluis Companys 23, 08010 Barcelona, Spain. ✉email: enrique.blanco@crg.eu; luciano.dicroce@crg.eu

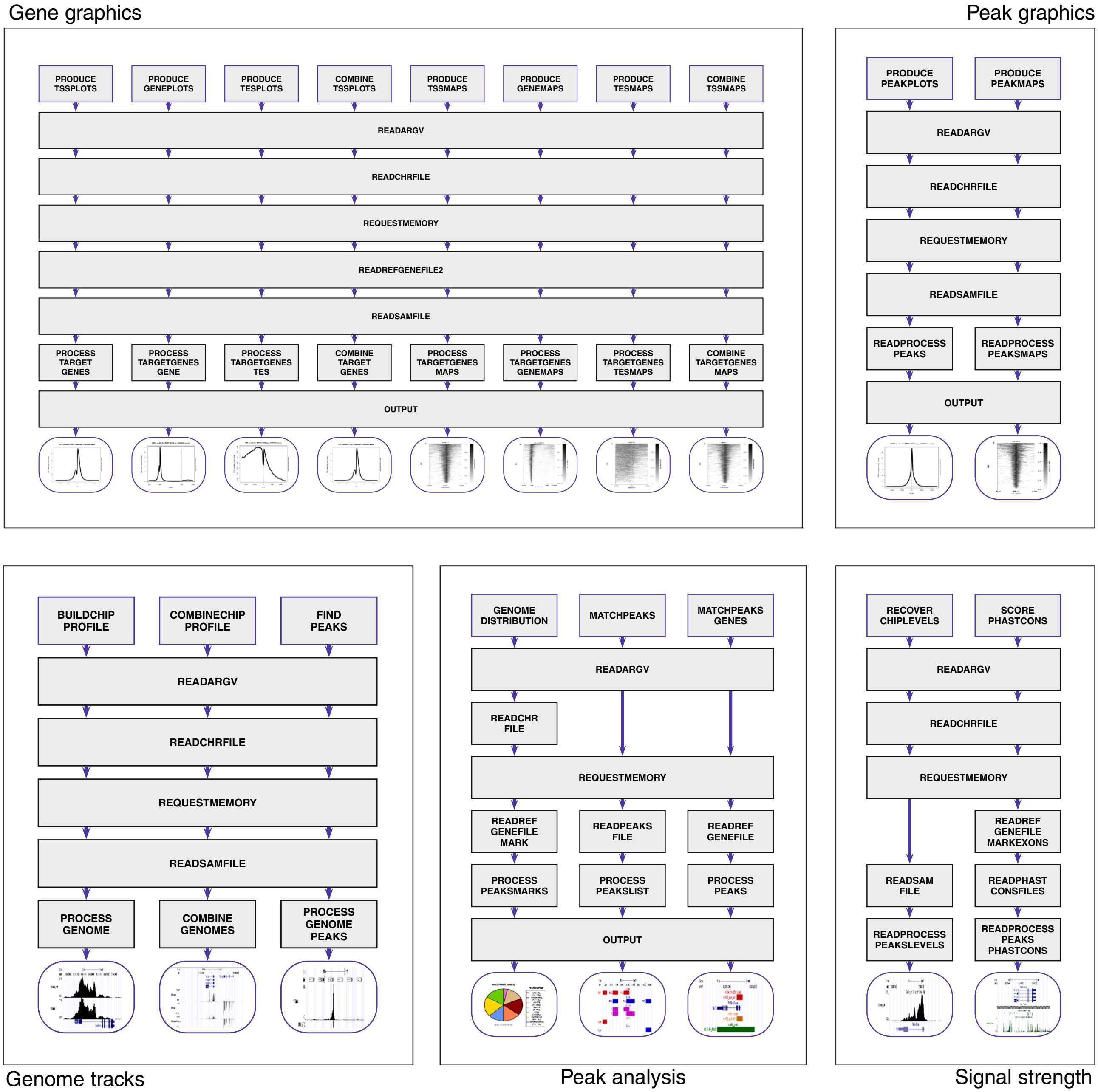


Figure 1. Modular architecture of SeqCode. Diagram of subroutines implemented in SeqCode to perform multiple tasks. Functions were classified depending on the object of analysis: genes, genomic regions or peaks, genome tracks, and signal levels. Functions are represented as boxes; arrows indicate the dataflow of each pipeline.

Results

Overview of the SeqCode software.

SeqCode is a command-line toolkit that produces top-quality images from high-throughput sequencing data. It has been implemented in ANSI C following a modular architecture of blocks (Fig. 1). All commands share a common standard UNIX interface, which makes our software a suitable candidate for integration into most bioinformatic pipelines. Each command is configurable, allowing the user to customize multiple attributes for the final plots (Table 1 for a comprehensive list of functions). SeqCode is fully functional over almost every genome that has been made available to scientific community. Information on a particular genome assembly is loaded from two external files that must be supplied by the user: (i) the chromosome size file (`ChromInfo.txt`), and (ii) the gene transcript annotations (`refGene.txt`), as provided by the RefSeq consortium[26]. Both archives can be downloaded from the UCSC genome browser for every genome release[14]. Moreover, users must specify the filename containing the reads mapped over the same genome assembly in SAM/BAM format[18]. Each analysis is usually performed across a list of target genes or a subset of genomic intervals—structured as BED records—that are also provided by the user within a plain

	Name	Description	Input	Output
Custom tracks for genome browsers	buildChIPprofile	Generates a custom track from a sequencing experiment to be visualized in current genome browsers	One SAM/BAM file	The custom track in BedGraph format
	combineChIPprofiles	Generates a custom track by subtracting the second sequencing experiment from the first one to be visualized in current genome browsers	Two SAM/BAM files	
Average occupancy plots	combineTSSplots	Draws the average distribution by subtracting the second sequencing experiment from the first one around the TSS of selected genes	Two SAM/BAM files, a list of genes	The average plot in PDF, the signal values and the R script
	produceGENEplots	Draws the average distribution of a sequencing experiment along the body of the meta-gene of selected genes	One SAM/BAM file, a list of genes	
	producePEAKplots	Draws the average distribution of a sequencing experiment around the center of selected peaks	One SAM/BAM file, a list of BED peaks	
	produceTESplots	Draws the average distribution of a sequencing experiment around the TES of selected genes	One SAM/BAM file, a list of genes	
	produceTSSplots	Draws the average distribution of a sequencing experiment around the TSS of selected genes	One SAM/BAM file, a list of genes	
Density heatmaps	combineTSSmaps	Draws the heatmap by subtracting the second experiment from the first one around the TSS of selected genes	Two SAM/BAM files, a list of genes	The heat map in PDF, the rank of signal values and the R script
	produceGENEmaps	Draws the heatmap of reads of an experiment along the body of the meta-gene of selected genes	One SAM/BAM file, a list of genes	
	producePEAKmaps	Draws the heatmap of reads of an experiment around the centre of selected peaks	One SAM/BAM file, a list of BED peaks	
	produceTESmaps	Draws the heatmap of a sequencing experiment around the TES of selected genes	One SAM/BAM file, a list of genes	
	produceTSSmaps	Draws the heatmap of a sequencing experiment around the TSS of selected genes	One SAM/BAM file, a list of genes	
Signal levels	recoverChIPlevels	Calculates the average, maximum, and total number of normalized reads of a sequencing experiment inside a set of regions of the genome	One SAM/BAM file, a list of regions in BED format	Average, maximum and total number of reads inside each region
Peak analysis	genomeDistribution	Distributes a list of regions of the genome into distinct gene features	A list of regions in BED format	The genome distribution in PDF, the annotation of the regions
	matchpeaks	Calculates the overlap between the components of two lists of peaks	Two lists of BED peaks	The list of overlapping peaks and the subsets of peaks that do not overlap on the other set
	matchpeaksgenes	Identifies genes genome-wide that contain one (or more) peaks from a list of selected peaks defined by the user accordingly to a set of rules and distances	A list of BED peaks	The list of genes in the genome that are target of the peaks
Evolutionary conservation	scorePhastCons	Calculates the average, maximum value, and total PhastCons score inside a set of genomic regions	A list of regions in BED format	Average, maximum and total score inside each region

Table 1. List of SeqCode functions.

text file. SeqCode output typically consists of graphical summaries of a sequencing experiment for the fraction of genes or genomic regions selected by the user. SeqCode tools routinely normalize data by sequencing depth (i.e. total read count in the experiment), although users can opt for performing the normalization by spike-in correction (i.e. total read count in the exogenous material). R is internally used to generate the resulting images to PDF. Intermediate plain text files and R scripts of each run are also delivered into the output folder labelled with the name of the job.

SeqCode can be executed in UNIX platforms under heterogeneous processor and memory configurations. We have tested SeqCode on iMac (Mac OS-X) and Personal Computer (Ubuntu Linux) platforms. The full source code is distributed upon open GNU license in GitHub (https://github.com/eblancoga/seqcode). We provide a Makefile to generate the binaries from the standalone version, which can be immediately used in the working path of the user machine. Users can automatically check that SeqCode services are functioning correctly with fragments of published sequencing samples using a collection of Perl scripts that is integrated into the code distribution. Moreover, we provide binary files for UNIX platforms, Oracle VM virtualbox™ appliances and Docker containers built on Linux Ubuntu, for different memory and storage configurations with the latest version of SeqCode internally pre-installed, through our website (http://ldicrocelab.crg.eu/). Our interactive web platform also offers front-end graphical interfaces implemented in PHP for SeqCode core functions, and other auxiliary tools designed in R, to generate graphical representations of user datasets (e.g. boxplots, heatmaps, PCA plots, scatter plots, UpSet charts[27], alluvial charts, and Venn diagrams).

Components of SeqCode.

SeqCode can generate quantitative and qualitative graphical representations, thereby allowing novel biological knowledge to be extracted from sequencing data. We briefly introduce each component of our suite of programs (Table 1; see our website for further information).

Custom profiles for quick visualization of fresh data in genome browsers.

Mapped reads can be piled-up along chromosomes to visually inspect the sequencing profiles in particular regions. The `buildChIPprofile` function produces genome-wide distributions in BedGraph format from a SAM/BAM file that can be uploaded in genome browsers (Fig. 2 and Supp. Figure 1). Multiple aspects of the profiles are customizable (e.g. name, color, and graphical resolution). SeqCode can also efficiently process paired-end and strand-specific libraries, and spike-in normalization factors can be introduced to weigh the distinct profiles. The distribution of two sequencing experiments can be graphically compared in a single track by subtraction. The `combineChIPprofiles` function generates genome-wide distributions in BedGraph format from two SAM/BAM files to be visualized in the same manner (see the RNA-seq strand-specific profiles in Fig. 2).

Averaged occupancy plots to spot the genomic distribution of a transcription factor/histone modification.

Occupancy plots (also known as aggregated plots or meta-plots) show the distribution of functional elements (which are typically reported from a ChIP-seq experiment) around a particular area of selected genes or within genomic intervals, with the aim of uncovering a characteristic pattern. The routines `produceTSSplots`, `produceGENEplots`, `produceTESplots`, and `producePEAKplots` generate average profiles of large-scale data from BAM files. It is therefore possible to select the viewpoint that is used as a reference to compute the average profile (Fig. 3a), which can be the transcription start site (TSS) of a gene, a normalized meta-gene body (GENE), transcription ending sites (TES), and the center of genomic intervals (PEAK). Moreover, users can configure multiple graphical features of the resulting image (background/foreground colors, flanking area width, etc.). Chromatin accessibility experiments (ATAC-seq) can be processed as well. The occupancy plot of a sequencing experiment can be corrected by subtraction of a second sample with the `combineTSSplots` function.

Density heatmaps to display the strength of the signal on each target.

Density heatmaps are barcodes that denote the signal strength pattern of a high-throughput experiment from a list of selected genes or a set of genomic intervals. The functions `produceTSSmaps`, `produceGENEmaps`, `produceTESmaps`, and `producePEAKmaps` generate density heatmaps from a set of target genes or peaks in BED format for a given SAM/BAM file (Fig. 3b). Multiple viewpoints are thus possible: TSS/TES of genes, meta-gene bodies or genomic regions. Customizing the graphical appearance of the final images (e.g. background/foreground colors and resolution) is straightforward in SeqCode (Supp. Figure 2 and our website). A heatmap corresponding to one sequencing experiment can be corrected by a second sample using the `combineTSSmaps` function.

Read counts for signal quantification and further dataset comparisons.

SeqCode executes the `recoverChIPlevels` method to determine the maximum, average, and total number of reads of a large-scale experiment within a set of genomic intervals in BED format given in a SAM/BAM file (Fig. 3d). Of note, not only ChIP-seq peaks but also gene annotations converted into this format can be used to evaluate the strength of a sequencing experiment within. Users can choose to normalize the output values with the total number of reads of the experiment or the number of spike-in reads, if available. Resulting distributions can be computed for multiple datasets for further statistical analysis of significance in boxplots or scatter plots.

Classification of peaks into different genomic features.

The `genomeDistribution` function classifies a list of genomic intervals in BED format (e.g. ChIP-seq peaks) into different features of the genome. Users can produce pie charts and annotation files under distinct degrees of detail (Fig. 3c). Using RefSeq gene transcript annotations, SeqCode classifies each region as promoter (proximal or distal), intergenic, or intragenic (5′ UTR, 3′ UTR, coding sequence [CDS], and introns) (Supp. Figure 3). Peaks overlapping more than one genomic feature are counted as many times as the number of genomic features they contain. On the other hand, superimposed pie charts (spie charts) can highlight the significance of the value distribution elements by using a second distribution as a reference[28]. Thus, when compared to the full-genome distribution of the same features, the results obtained with the `genomeDistribution` function for a particular sample can be further recycled to generate one spie chart, which depicts the relevance of the frequency of each element in the high-throughput experiment and along the genome (see examples in our website).

Identification of target genes and peak comparison.

ChIP-seq and ATAC-seq peaks are typically associated to the TSS of the nearest gene to further analyze genesets using ontology term enrichment analysis. The `matchpeaksgenes` routine allows the user to define the conditions of matching a given peak to a certain gene, by defining which regions around the TSS or the gene body (according to RefSeq) can be used in the

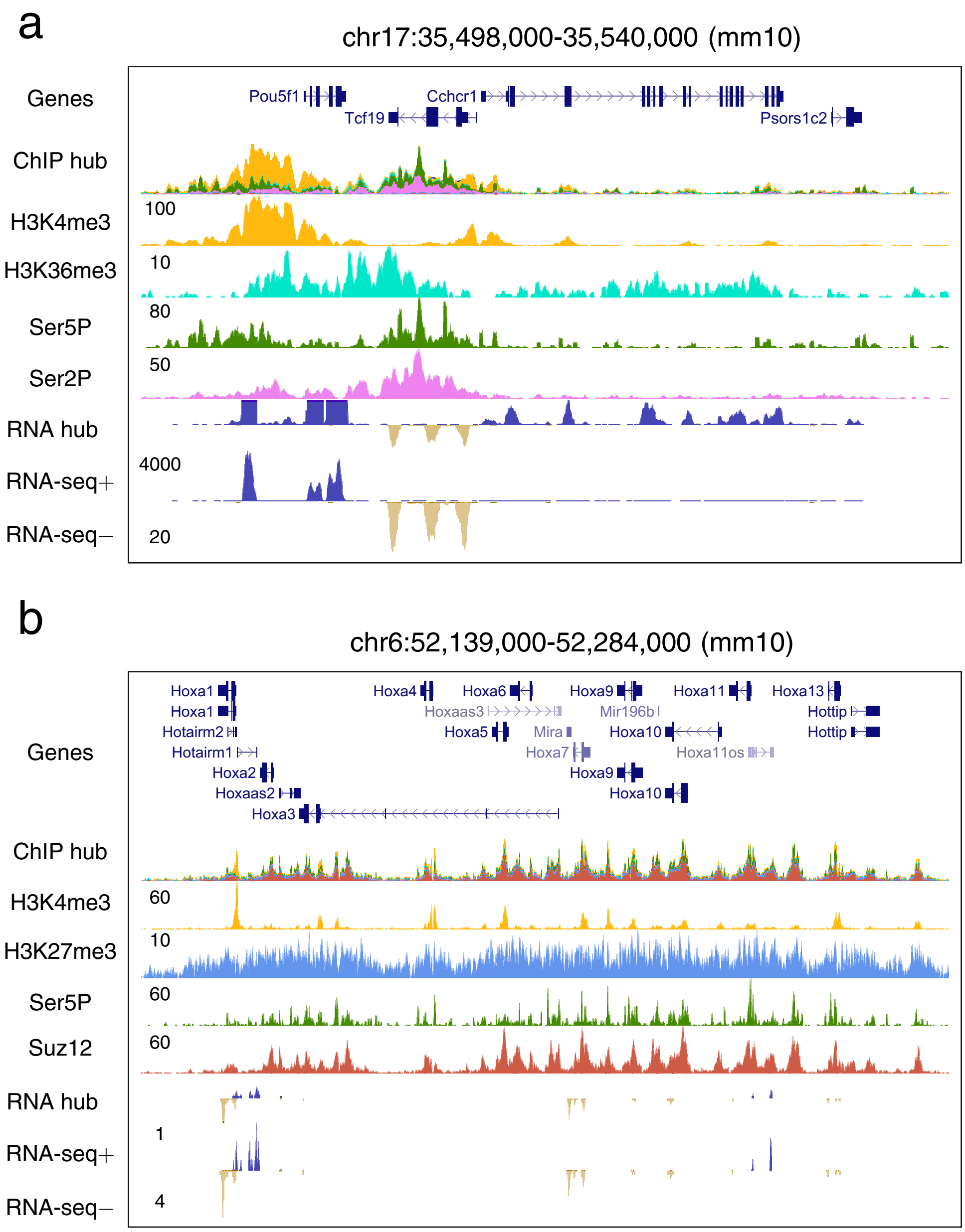


Figure 2. SeqCode ChIP-seq and RNA-seq profiles in mESCs for visualizing in genome browsers. **(a)** Actively transcribed region containing the *Oct4/Pou5f1* pluripotency gene. **(b)** Region repressed for transcription by PcG proteins containing the HoxA complex. Raw data were retrieved from[32,33,38,75]. The SeqCode `buildChIPprofile` function (window size = 10) was used to generate each custom track from resulting BAM files. The SeqCode `combineChIPprofiles` function was used to generate the RNA-seq strand-specific profiles. Composite ChIP-seq and RNA-seq supertracks integrate all information from each individual track shown below.

overlap. Moreover, SeqCode offers a generic version of the same command, called `matchpeaks`, to compare two lists of genomic intervals in BED format. Overlapping regions and peaks found only in one set are formatted as custom tracks, which can be graphically visualized in genome browsers (see our website for further information).

Evaluation of the evolutionary conservation of genomic regions. Functional sequences of the genome exhibit a strong degree of conservation along evolution. By comparing the sequence of the orthologous region from multiple species, it is possible to score the potential of a genomic interval as containing a conserved regulatory block. The `scorePhastCons` function takes advantage of the PhastCons score[29] to rank the members of a peak collection in BED format (note that the PhastCons data files must have been previously downloaded from the UCSC genome browser) (Fig. 4).

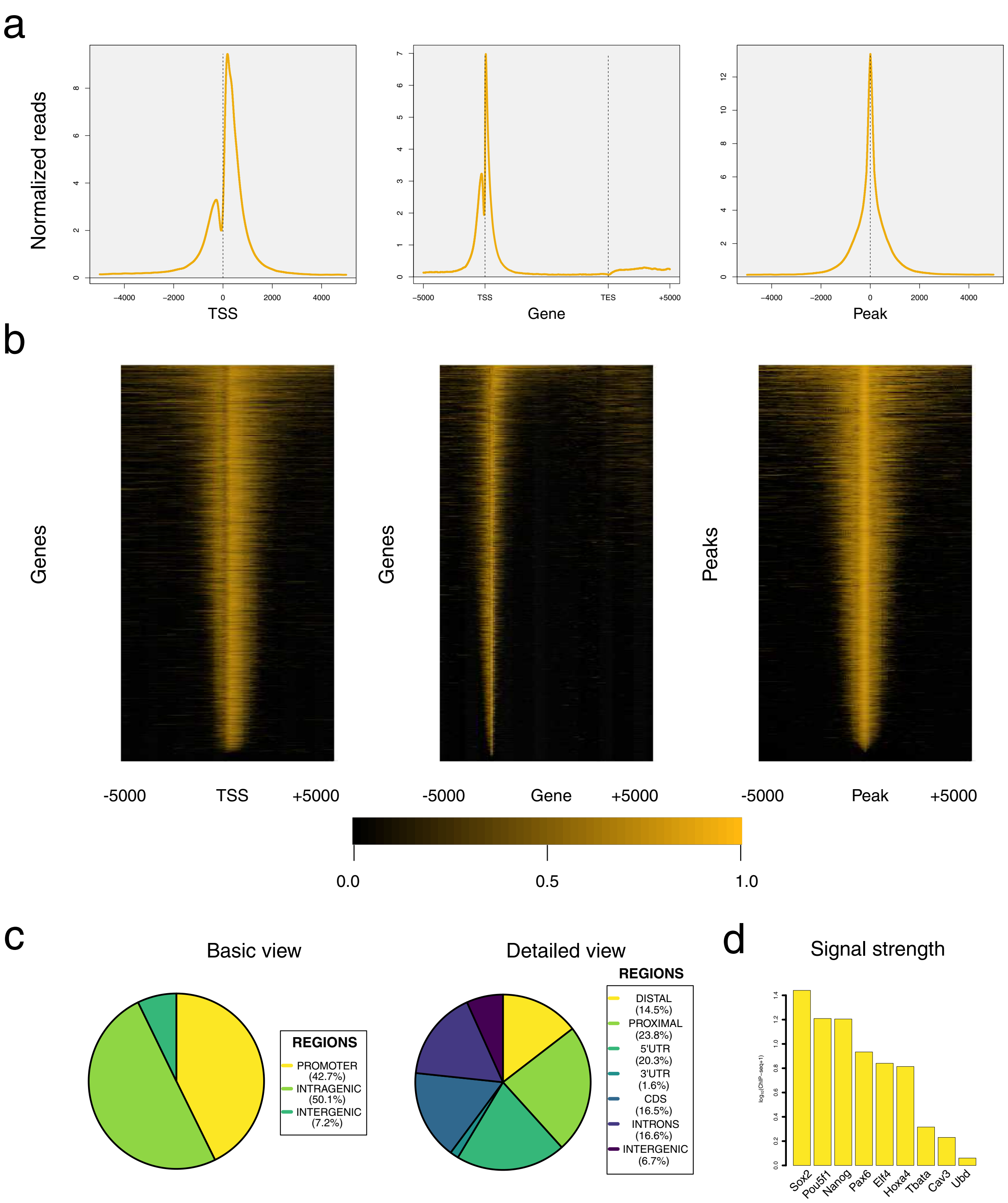


Figure 3. Basic analysis of the H3K4me3 ChIP-seq sample in mESCs using SeqCode. (**a**) From left to right: average distributions around the TSS of H3K4me3 target genes (`produceTSSplots`), along the gene body of the same genes (`produceGENEplots`), and around the center of H3K4me3 peaks (`producePEAKplots`). (**b**) From left to right: heatmap around the TSS of H3K4me3 target genes (`produceTSSmaps`), along the gene body of the same genes (`produceGENEmaps`), and around the center of H3K4me3 peaks (`producePEAKmaps`). (**c**) Basic and detailed genome distribution of H3K4me3 peaks (`genomeDistribution`). (**d**) ChIP-seq signal levels of selected genes (`recoverChIPlevels`). Raw data were retrieved from[32].

Evaluating the SeqCode soundness. Gene expression is regulated in part by genome organization. For instance, promoters and enhancers associated to transcriptionally active genes must be physically accessible for transcription factors to bind, while silent genes are embedded in genomic regions with a higher level of chromatin compaction[30]. Chromatin remodeling complexes, such as those involving the Polycomb group (PcG) and Trithorax group (TrxG) proteins, modulate nucleosomes by introducing post-translational modifications to the histone tails, which ultimately affect the chromatin conformation[31]. To determine the consistency of SeqCode, we selected the ChIP-seq experiments in mouse embryonic stem cells (mESCs) of two histone modifications: histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 36 trimethylation (H3K36me3). While H3K4me3 exhibits a characteristic pattern of sharp peaks around TSSs of expressed genes[32], H3K36me3 is distributed in broad domains entirely covering the bodies of transcriptionally active genes[33].

We first produced the genome-wide profile of each ChIP-seq experiment using the `buildChIPprofile` function (Supp. Figure 4) and then generated the average occupancy plot of a single gene comprising both histone modifications using the `produceGENEplots` function. Custom tracks focused on the same gene, and meta-gene plots coincided well (Supp. Figure 4a). We successfully contrasted the occupancy plot around the TSS of another gene produced by the `produceTSSplots` function, with the corresponding custom track and its density heatmap generated with the `produceTSSmaps` function (Supp. Figure 4b). To evaluate the effects of the sequencing depth, we generated genome-wide profiles of both ChIP-seq experiments with the `buildChIPprofile` function, using different down-sampling sizes. The effect of saturation in both marks is different: one-million reads was sufficient to visualize the H3K4me3 sharp patterning, while additional sequencing depth of eight-million reads was necessary to distinguish the broader domains of H3K36me3 (Supp. Figure 5).

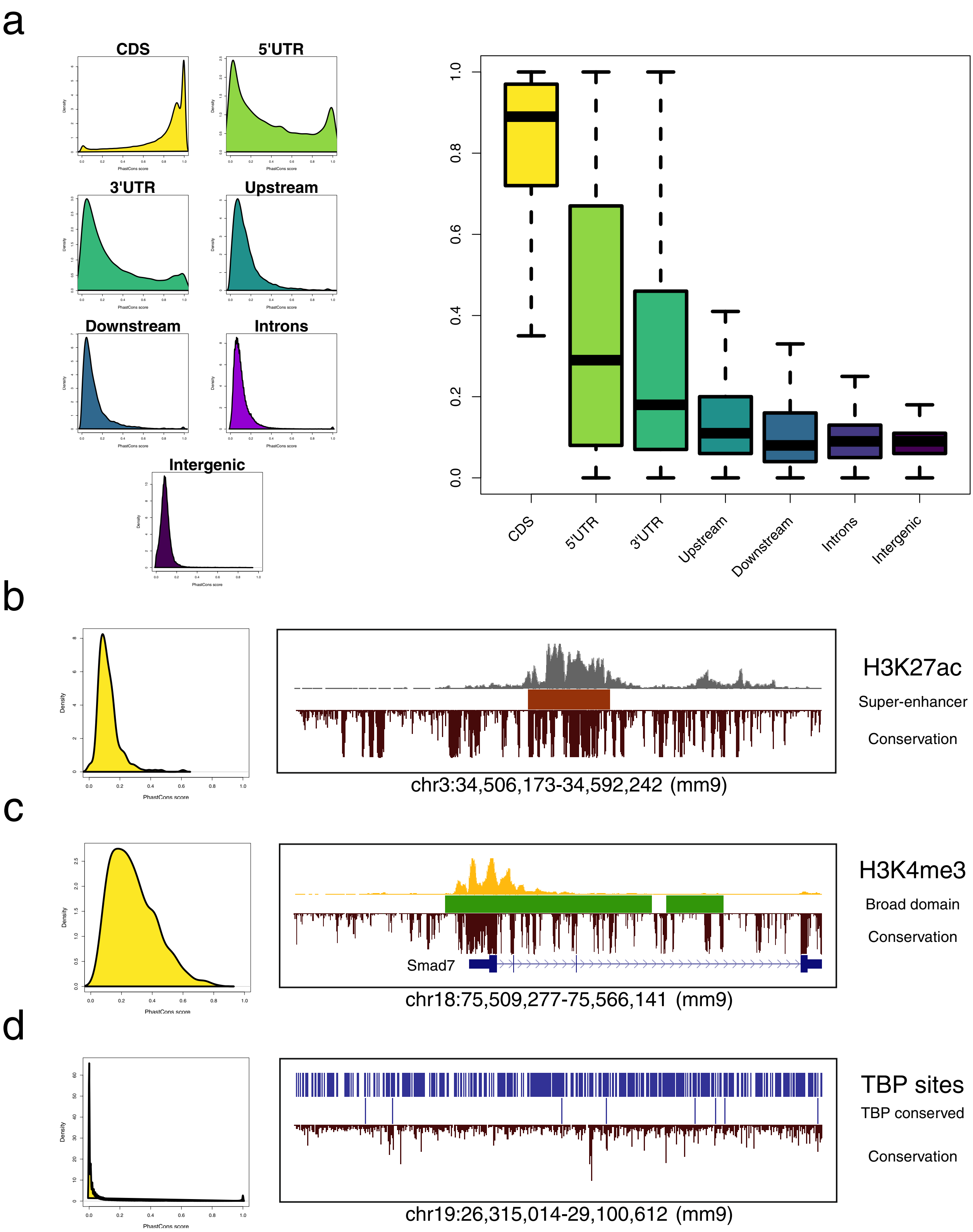
In our second test, we verified that SeqCode offers a good compromise between the quality of the biological results and the amount of available memory in the computer. Users can instruct SeqCode functions to use a particular window resolution in the binning of the genome (option –w), which affects the memory requirements. Although the quality of the final graphical representations was lower (as expected), the results generated with the `buildChIPprofile`, `produceTSSplots` and `produceTSSmaps` programs were consistent in all cases (Supp. Figure 6). Therefore, SeqCode can process a sequencing sample using almost any computer with minimal memory requirements, if necessary (Supp. Figure 7). On the other hand, users working on a powerful workstation can configure this software to obtain the maximum graphical quality.

Running SeqCode in realistic epigenetic scenarios. Histone marks are thought to be associated with positive and negative transcription events. So-called bivalent domains, first described in mESCs[34], are simultaneously decorated with a combination of opposing H3K4me3 and histone H3 lysine 27 trimethylation (H3K27me3) marks. Whereas H3K4me3 at the TSS of transcriptionally active genes is a key product of TrxG proteins, H3K27me3 is catalyzed by specific PcG complexes in gene silencing contexts. PcG proteins participate into two different classes of repressive complexes: Polycomb repressive complexes 1 and 2 (PRC1 and PRC2). Each complex comprises a set of core components—such as Ring1b for PRC1, and Suz12 for PRC2—together with other subunits that produce distinct variations of each complex[31]. Bivalent genes and enhancers are key components of the developmental regulatory circuitry, and intensive research in the past years is currently extending the bivalency paradigm to other cellular contexts[35]. Below, we show how to use SeqCode to characterize the epigenomic landscape of different groups of genes in mESCs, and during tumorigenesis, using three practical cases in which bivalency plays a role.

Case 1. Analysis of the epigenetic signature of the genome. First, we generated the genome-wide profiles of published ChIP-seq experiments of H3K4me3 and H3K27me3[36] using the `buildChIPprofile` routine (Fig. 5a). Next, we classified the ChIP-seq peaks available for each histone mark into three distinct classes (H3K4me3 + /H3K27me3–, H3K4me3 + /H3K27me3 + and H3K4me3–/H3K27me3 +) with the `matchpeaks` function (Fig. 5). Bivalent domains constitute most H3K27me3 peaks. Further ChIP-seq signal quantification of each class of peaks running the `recoverChIPlevels` program confirmed the composition of each peak class. Bivalency was preferentially found in gene promoters, and in intergenic regions to a minor extent (`genomeDistribution` pie charts; Fig. 5b). To study in-depth bivalency at the gene level, we ran the `matchpeaksgenes` application (which here associates a peak to one gene if it is located within 2500 bp from the TSS). Overlap between each set of target genes classified genes into active (H3K4me3 + /H3K27me3–), bivalent (H3K4me3 + /H3K27me3 +), or silent (H3K4me3–/H3K27me3 +). Gene ontology (GO) analyses confirmed that bivalent genes are related to development (Fig. 5c).

In addition to H3K4me3 and H3K27me3, we also characterized active and bivalent genes in terms of H3K36me3 and Suz12 occupancy using the `produceTSSplots`, `produceTSSmaps`, and `recoverChIPlevels` programs (Fig. 5d,e). As expected, H3K4me3 presented a sharp peak centred over the TSS of both classes of genes but which was higher for active genes. For H3K27me3, we observed a canonical broad peak pattern in the TSS of bivalent genes. With respect to H3K36me3, we reproduced the well-known broad domain over the gene body of transcriptionally active genes. Finally, Suz12 was present in the TSS of bivalent genes, which are precisely decorated by H3K27me3. We confirmed the differences at expression levels between both gene groups using RNA-seq data (Fig. 5e).

Case 2. Characterization of the panel of active and repressive actors in mESCs. We have reconstructed a global picture of this cellular context using ChIP-seq data[32,33,37] of histone modifications (H3K4me3, H3K36me3, H3K27me3, and ubiquitinated H2A [H2Aub]) and three PcG proteins (Suz12, Jarid2, and Ring1b). PcG proteins have key roles in preventing the transcription of many developmental regulators in mESCs, which would otherwise shift these cells from pluripotency to a more differentiated state[31]. We also integrated ChIP-seq


a
CDS
5'UTR
3'UTR
Upstream
Downstream
Introns
Intergenic
Density
PhastCons score
0.0
0.2
0.4
0.6
0.8
1.0
CDS
5'UTR
3'UTR
Upstream
Downstream
Introns
Intergenic
b
H3K27ac
Super-enhancer
Conservation
chr3:34,506,173-34,592,242 (mm9)
c
H3K4me3
Broad domain
Conservation
Smad7
chr18:75,509,277-75,566,141 (mm9)
d
TBP sites
TBP conserved
Conservation
chr19:26,315,014-29,100,612 (mm9)

◂**Figure 4.** Evolutionary conservation landscape of distinct features of the mouse genome calculated by SeqCode. **(a)** PhastCons average score distribution calculated by the SeqCode `scorePhastCons` function on distinct gene features: protein-coding regions (CDS), 5′ untranslated regions (5′UTR), 3′ untranslated regions (3′UTR), promoters 1000 bp upstream of the TSS (Upstream), regions 1000 bp downstream the TES (Downstream), intronic regions (Introns), and intergenic regions (Intergenic). RefSeq annotations were used to extract the coordinates of all instances of each feature in the mouse genome (mm9). Left, distribution of the score (from 0 to 1); right, boxplot summarizing differences between features. **(b–d)** PhastCons average score calculated by SeqCode on distinct regulatory features: **(b)** super-enhancers identified as significant concentrations of H3K27ac in mESCs[76], (red); **(c)** broad domains of H3K4me3 reported in mESCs[77] (green); and **(d)** computational predictions of TATA boxes using Jaspar[78], before and after the conservation filtering (TBP and TBP conserved), to highlight the significant drop in the number of ab initio predictions (blue). Left, the global distribution of each class of elements; right, an example of high conservation. The phastCons30way track of mouse was used to score each set of regions with the SeqCode `scorePhastCons` function. Raw data were retrieved from[32,79].

data of two different states of the RNA polymerase II[38]: paused (Ser5-phosphorylated [Ser5P]) and elongating (Ser2-phosphorylated [Ser2P]). Ser5P is a marker of poised promoters, which coincides in most cases with bivalency, while Ser2P is located in gene bodies of transcriptionally active genes[39]. Initially, for each ChIP-seq experiment analyzed here, we retrieved the target peak set from the original publication and used the `matchpeaksgenes` function to annotate the collection of target genes of each experiment. Next, we have used the `produceTSSplots`, `produceTSSmaps` and `produceGENEplots` functions of SeqCode to study the distribution of each element around the TSS of its target genes.

Analysis of meta-gene plots (Fig. 6) indicated that the highest ChIP-seq signal was found around the TSS, with the exception of H3K36me3 and Ser2P, which presented a broad pattern covering gene bodies. In line with this, similar differences were observed when analyzing the genome distribution of peaks of each experiment (Fig. 6): while H3K36me3 and Ser2P pie charts generated using the `genomeDistribution` command were enriched in intragenic categories, the remaining elements tended to be located around the TSS of genes. Intensity of ChIP-seq signals around the center of peaks was reported using the `producePEAKplots` and `producePEAKmaps` commands. Finally, we performed a random sampling of RefSeq genes, in three groups of 100 genes each according to RNA-seq expression, of highly expressed, moderately expressed, or silent (Fig. 6, lower panel). By calculating the ChIP-seq normalized counts of each experiment along the three gene classes with the `recoverChIPlevels` function, a clear pattern emerged: gene expression positively correlated with H3K4me3, H3K36me3, Ser5P, and Ser2P, while it negatively correlated with H3K27me3, Suz12, Jarid2, Ring1b, and H2Aub (Fig. 6). We successfully used the same approach to study the epigenetic landscape of the wing imaginal disc of the fruit fly, using distinct histone marks combined with transcription factor binding (Supp. Figure 8). Both examples illustrate how SeqCode can be executed to build comparative panels of a high number of sequencing samples, which can then be used to obtain novel knowledge for further analysis.

Case 3. Comparative analysis after genetic perturbation of epigenetic actors.

Although ChIP-seq experiments are semiquantitative approaches to evaluate the presence of a certain histone modification or protein on a genomic region, they can be relevant to compare the outcome for the same antibody in two different conditions. Genetic deletion of a gene can be used as validation for this scenario. Thus, the ChIP-seq pattern of a target protein in wild-type condition (WT) should be totally (or at least partially) erased when performing ChIP-seq analysis using the same antibody in knockout (KO) or knockdown (KD) cells. Moreover, once the effect of the perturbation has been performed, researchers often evaluate its effect, for example, over putative alterations of the ChIP-seq pattern of a second biological target.

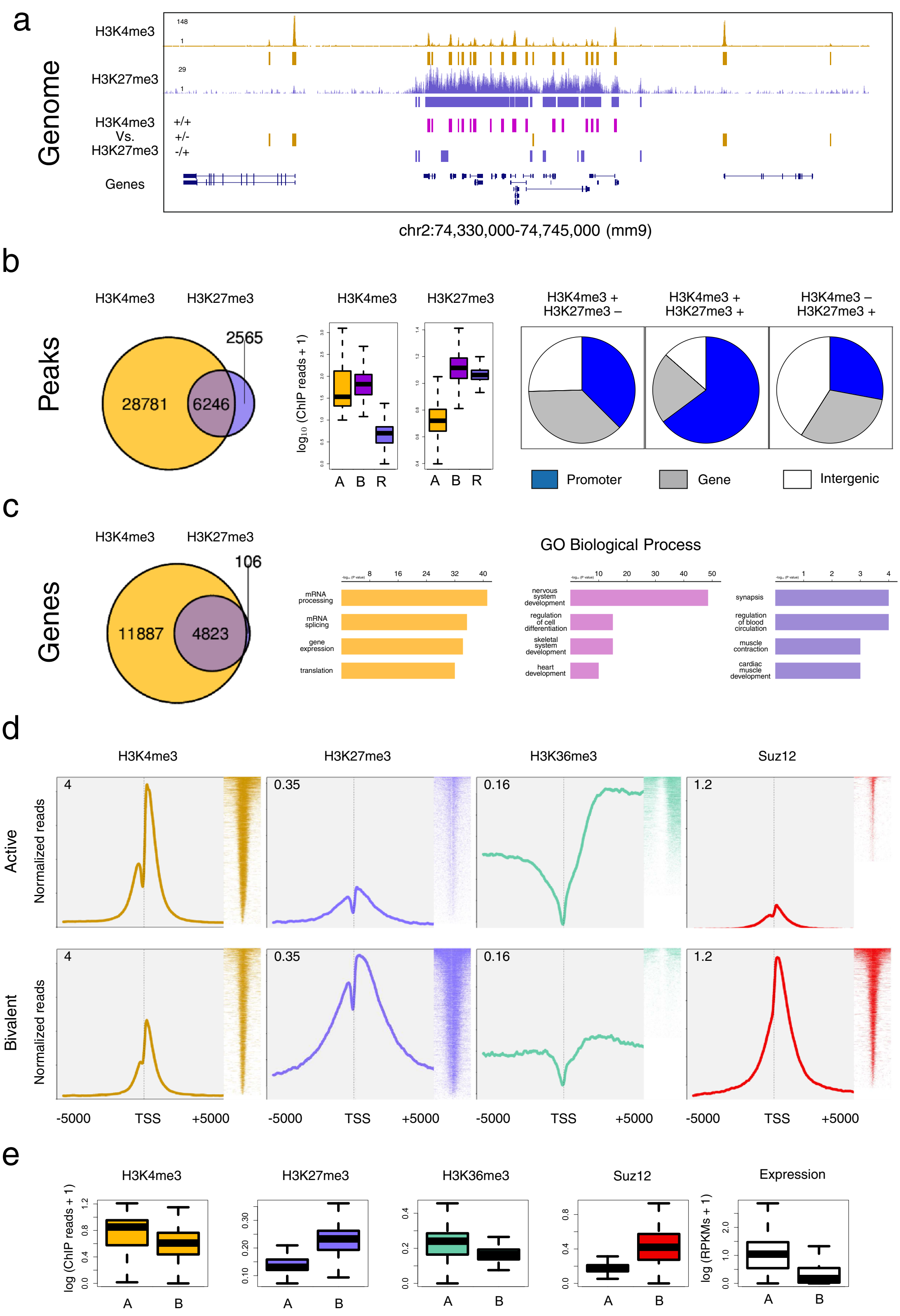
To illustrate both concepts, we have revisited three scenarios gathered from the literature. In first place, we depicted the intensity of ChIP-seq of MLL2[40], a TrxG protein required for H3K4me3 deposition associated to bivalent promoters in mESCs, under two conditions (WT and MLL2 KO) using the `recoverChIPlevels`, `producePEAKplots`, `producePEAKmaps`, and `buildChIPprofile` functions. In all the graphical representations, a substantial drop in the ChIP-seq signal of MLL2 upon KO of MLL2 is reported, as expected (Fig. 7a, left). Importantly, this event is accompanied by the corresponding decrease in the H3K4me3 product (Fig. 7a, right).

Secondly, we explored how the role of PcG components is crucial also in the context of cancer. In fact, when scrutinizing with SeqCode the ChIP-seq profiles of RING1B (a PRC1 subunit) for Ewing sarcoma (EWS)[41], we confirmed (i) the good performance of the RING1B KD, and (ii) a striking decrease in the binding of the EWSR1-FLI1 fusion protein, which is responsible for the EWS disease, upon KD of RING1B (Fig. 7b). Finally, in prostate cancer, we employed a similar strategy to evaluate the effects of the KD of PHF19, a PRC2 subunit, using SeqCode. In this case, we observed an increase of MTF2 (also a PRC2 subunit) upon PHF19 KD (Fig. 7c), suggesting the existence of a compensatory mechanism[42].

Offering SeqCode functions through a front-end web site.

This software can be potentially integrated as an external component of any bioinformatics pipeline. For that purpose, we have designed a user-friendly web interface to grant access to most functions of our software (http://ldicrocelab.crg.eu/). Our website homepage is structured into four classes of services (Fig. 8):

(1) NGS tools (Supp. Figure 9): sequencing analysis tools (ChIP-seq peak annotation and gene association, signal quantification, aggregated plots, and density heatmaps), for a selection of published ChIP-seq experiments;


a
Genome
H3K4me3
148
1
H3K27me3
29
1
H3K4me3
Vs.
H3K27me3
+/+
+/-
-/+
Genes
chr2:74,330,000-74,745,000 (mm9)
b
Peaks
H3K4me3
H3K27me3
2565
28781
6246
H3K4me3
H3K27me3
log10 (ChIP reads + 1)
A B R
A B R
H3K4me3 +
H3K27me3 –
H3K4me3 +
H3K27me3 +
H3K4me3 –
H3K27me3 +
Promoter
Gene
Intergenic
c
Genes
H3K4me3
H3K27me3
106
11887
4823
GO Biological Process
-log10 (P value)
8 16 24 32 40
mRNA processing
mRNA splicing
gene expression
translation
10 20 30 40 50
nervous system development
regulation of cell differentiation
skeletal system development
heart development
1 2 3 4
synapsis
regulation of blood circulation
muscle contraction
cardiac muscle development
d
H3K4me3
H3K27me3
H3K36me3
Suz12
Active
Normalized reads
4
0.35
0.16
1.2
Bivalent
Normalized reads
4
0.35
0.16
1.2
-5000 TSS +5000
-5000 TSS +5000
-5000 TSS +5000
-5000 TSS +5000
e
H3K4me3
H3K27me3
H3K36me3
Suz12
Expression
log (ChIP reads + 1)
0.0 0.4 0.8 1.2
0.10 0.20 0.30
0.0 0.2 0.4
0.0 0.4 0.8
log (RPKMs + 1)
0.0 1.0 2.0
A B
A B
A B
A B
A B

◂**Figure 5.** Epigenetic signature of active and bivalent genes from mESCs reconstructed with SeqCode. **(a)** Bivalent region repressed for transcription by PcG proteins containing the HoxD complex. Genome-wide profiles and peaks of H3K4me3 and H3K27me3 are shown along this locus. **(b)** Overlap of H3K4me3 and H3K27me3 peaks produces three classes of sites: active (H3K4me3 + /H3K27me3–; orange), bivalent (H3K4me3 + /H3K27me3 +; violet), and silent (H3K4me3–/H3K27me3 +; blue). Determination of signal strength and genomic distribution of each class of peaks. **(c)** Overlap of H3K4me3 and H3K27me3 target genes, showing the functional analysis of active genes (orange), bivalent genes (violet), and silent genes (blue). **(d)** Average distribution and heatmap of the ChIP-seq signal around the TSS of active and bivalent genes for H3K4me3, H3K27me3, H3K36me3, and Suz12. **(e)** ChIP-seq levels of each experiment are shown for active and bivalent genes (left and right boxes in the boxplots, respectively). For comparison, gene expression (in RPKMs) is shown for both gene sets. Raw data of ChIP-seq and RNA-seq experiments were retrieved from[36]. The SeqCode `buildChIPprofile` function was used to generate each custom track from the resulting BAM files after mapping; the `matchpeaks` application compared the H3K4me3 and H3K27me3 peaks; the `recoverChIPlevels` application determined the strength of the ChIP-seq signal at each subset of peaks; the `genomeDistribution` program calculated the genomic composition of each collection of peaks, according to RefSeq annotations; the `matchpeaksgenes` routine associated ChIP-seq peaks and target genes; and the `produceTSSplots` and `produceTSSmaps` programs generated the average distribution meta-plots and heatmaps of each ChIP-seq sample for signal around the TSS of active and bivalent genes, respectively. GO term enrichment was analyzed with Enrichr[80]. For boxplots in **(b,e)**, ChIP-seq counts were normalized by the total number of mapped reads, and RNA-seq expression values were calculated as RPKMs.

(2) Datasets (Supp. Figure 10): functions to graphically represent user data distributions also in form of correlations or clusters (e.g. boxplots, PCA plots, heatmaps and scatter plots);

(3) Gene sets (Supp. Figure 11): plots to compare the overlap between members of distinct user datasets (e.g. Venn diagrams, alluvial charts and UpSet plots[27]);

(4) List Ops: operations between lists of values provided as plain text files. Each line of information contains a record, and each column contains the value for a particular feature (e.g. joining two files, filtering in one file of specific records, etc.).

In most cases, users will find options for appropriate labelling and customization of most graphical parameters of the corresponding SeqCode web service. Each application generates the images in PNG and PDF formats and provides links to the source files and R scripts. Users will find abundant on-line documentation about each service and a comprehensive set of tutorials. Moreover, we have integrated a complete manual of the command-line version of SeqCode, furnished with multiple examples of analysis, into our website.

Comparison with similar tools.

Assessing the available software for graphical annotation of large-scale sequencing information is complex due to the heterogeneity of the applications and their potential audiences[43–45]. Thus, certain tools such as SAMtools[18], BEDTools[19], and FAST[46]—which are designed to manage NGS information at a low level—will be excluded from this comparative study. On the other hand, instead of GALAXY[17], the major hub of web interfaces towards other NGS applications, we opted for evaluating Cistrome[44], a GALAXY-based server, and deepTools[47], a visualization resource also available inside GALAXY. Here, we evaluated 18 different visualization tools, which to our knowledge constitutes the most comprehensive selection to date and provides a faithful portrait of the current state of art in this field (Table 2). We propose to focus on eight elementary features that characterize every program (Fig. 9a): (i) the inventory of genomes that can be analyzed, (ii) the class of genomic element that can be interrogated, (iii) the type of sequencing data to be processed, (iv) the family of graphical representations that can be generated, (v) the available options for determining signal strength of samples, (vi) the flexibility of the working interface for analysis, (vii) the computational requirements to be satisfied to run the software, and (viii) the state of completion of the final documentation provided to the users. The final scores of each program according to these attributes are summarized in Fig. 9b and Table 3, and examples of output from several tools reproducing our analysis of active genes, bivalent genes, or repressed genes in mESCs (see Fig. 5) are shown in Supp. Figure 12–14.

(i) Which inventory of genomes can be analyzed?

Most programs work only for a small subset of species, mostly human and mouse (GeneProf, BindDB) and perhaps also other model organisms, such as the fruit fly (ChIP-Seq, CisGenome). Certain programs provide access only to a subset of genome assemblies for these species (Cistrome, ngs.plot). To circumvent these issues, EaSeq provides a connection to link available annotations in the UCSC genome browser to elements into the analysis pipeline. Other tools allow externally annotations for other genomes to be introduced using GTF files that must be previously customized by the user (CGAT, HOMER). In contrast to these, SeqCode adopts the standard annotation files provided by the RefSeq consortium for almost any genome. This strategy permits users to easily work with different releases of the same genome by simply switching to the required RefSeq gene annotation file.

(ii) Which class of genomic information can be interrogated?

Most resources perform actions exclusively focusing on a single class of element as the principal viewpoint. Thus, users can interrogate sequencing data using gene names (GeneProf and Cistrome) or genomic intervals defining ChIP-seq peaks (ChIP-Seq, ChIPseeker, deepTools, EaSeq, genomation, seqMINER and Spark). However, depending on the context, dual access through both classes of reference points (genes and peaks) is frequently convenient throughout most bio-

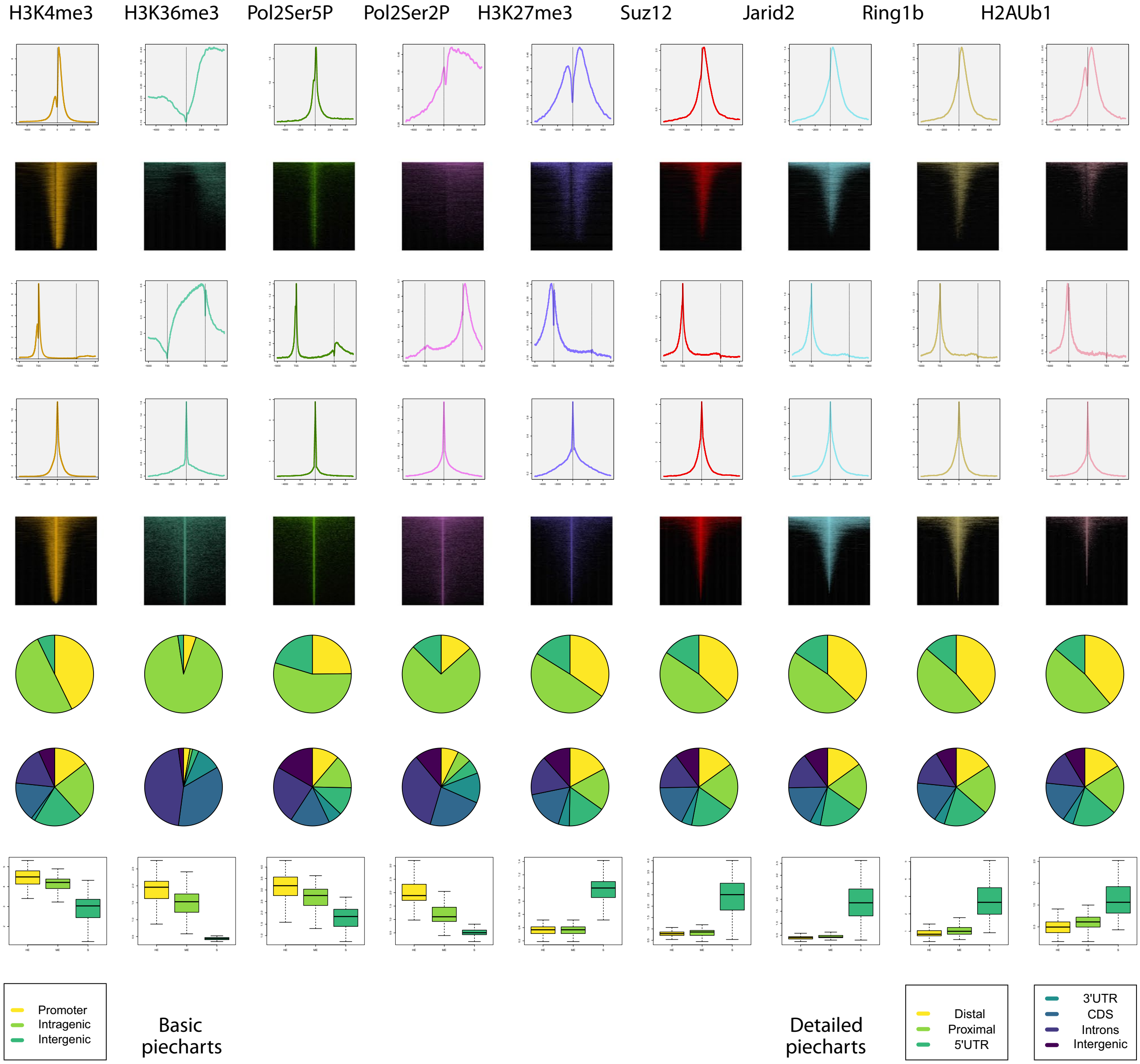


Figure 6. Panel of active and repressive actors in mESCs generated by SeqCode. For each ChIP-seq signal and corresponding set of target genes, the average distribution (`produceTSSplots`) and heatmap (`produceTSSmaps`) around the TSS, average distribution along the gene body (`produceGENEplots`), average distribution (`producePEAKplots`) and heatmap (`producePEAKmaps`) around the center of the peaks, genome distribution of each set of peaks (`genomeDistribution`), and ChIP-seq levels normalized by the total number of reads (`recoverChIPlevels`) are shown for 100 highly expressed genes (more than 1000 RPKMs; red), 100 moderately expressed genes (100–500 RPKMs; yellow) and 100 silenced genes (0–1 RPKMs; blue). Raw data were retrieved from[32,33,37,38]. Target genes of each ChIP-seq were identified with the `matchpeaksgenes` routine.

informatics analysis, which usually combine ChIP-seq binding data with RNA-seq expression data. Therefore, although it is not yet a universal feature, SeqCode and other tools such as BindDB, CGAT and ngs.plot allow the user to characterize sets of genomic intervals (e.g. enhancers) together with gene regions based on their functions.

(iii) Which class of sequencing experiments can be processed? Many applications are uniquely designed to analyze ChIP-seq data. However, other NGS methods, including RNA-seq, ATAC-seq, and MeDIP-seq, are widely accessible to the scientific community. We found a variety of combinations in the type of experiment that can be analyzed: GeneProf and ngs.plot, ChIP-seq and RNA-seq; Cistrome, ChIP-seq and ATAC-seq; and ChAsE, ChIP-seq and MeDIP-seq. There is an emerging group of general-purpose programs that can manage all class of NGS information: SeqCode (presented here), CGAT, deepTools, EaSeq, Genomation, and

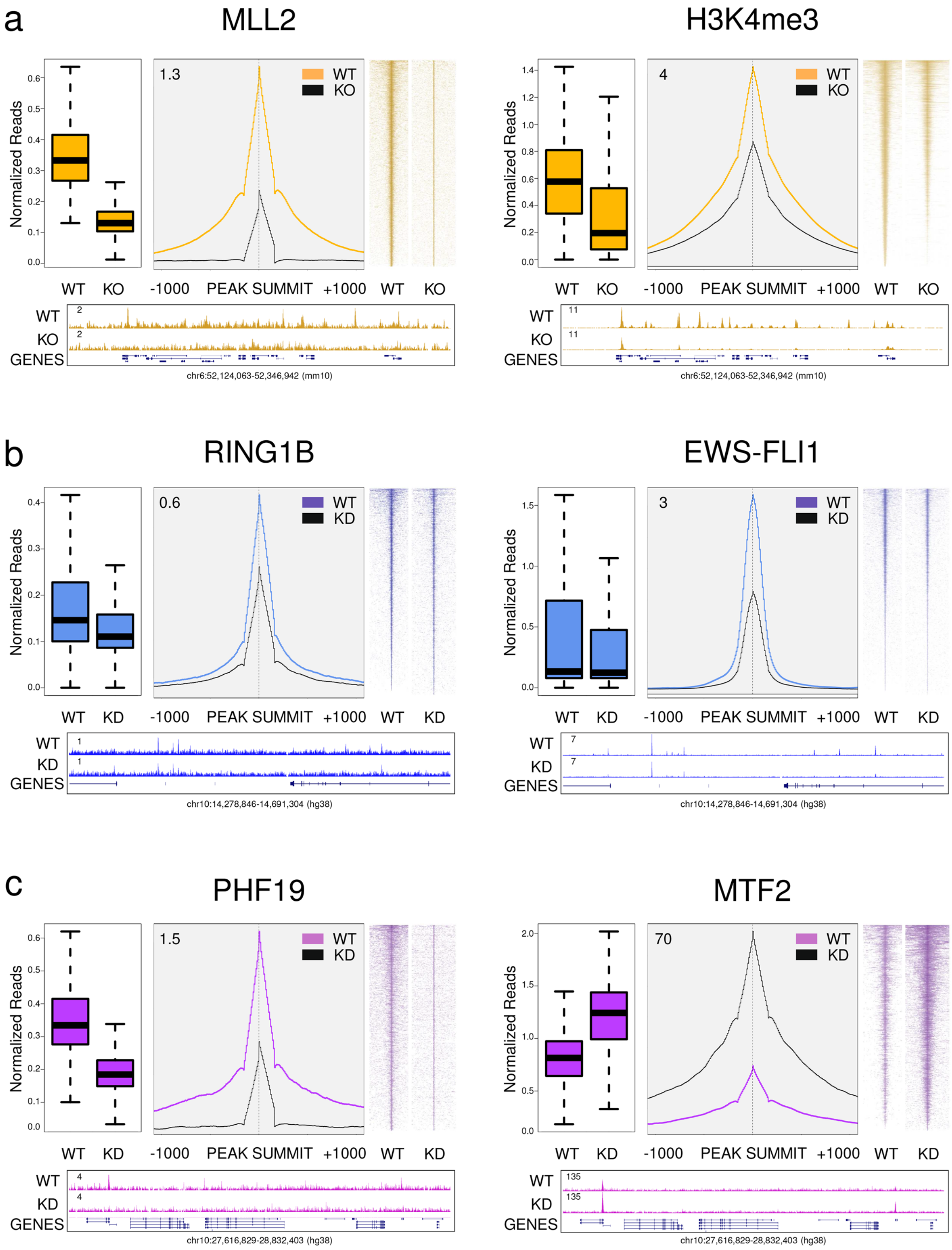
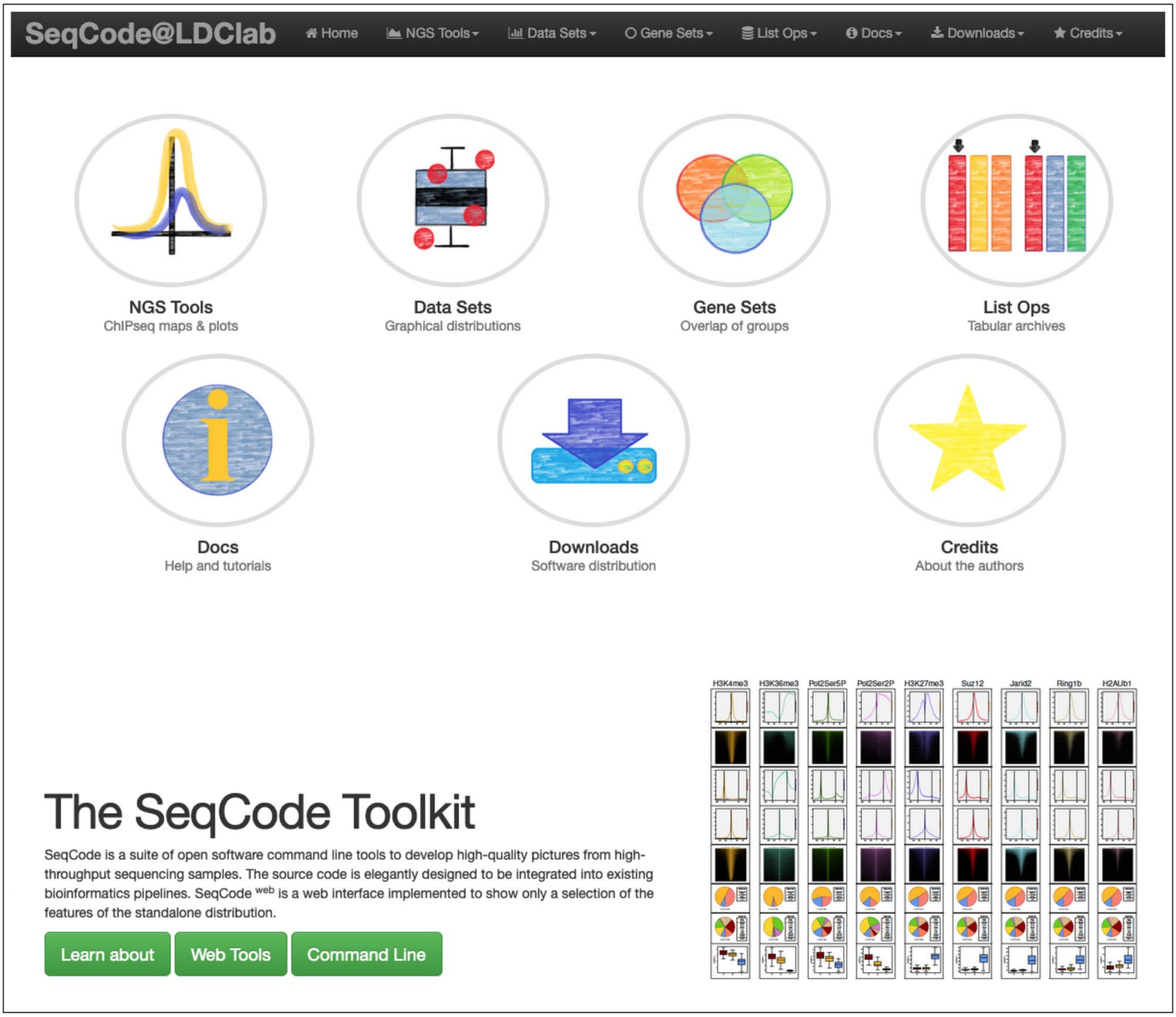


Figure 7. Evaluation of ChIP-seq experiments after genetic perturbation with SeqCode. **(a)** Study of the efficiency of MLL2 knockout (KO) in mESCs (left) and its effect over the occupancy pattern of H3K4me3 (right). **(b)** Study of the efficiency of RING1B knockdown (KD) in the Ewing sarcoma A673 cell line (left) and its effect over the binding pattern of EWS-FLI1 oncogenic protein (right). **(c)** Study of the efficiency of PHF19 KD in the prostatic cancer DU145 cell line (left) and its effect over the pattern of MTF2 (right). Peaks called in the wild-type (WT) condition of the factor being analyzed on each case were used as a reference in both conditions. For every pair of WT-KO/KD experiments, ChIP-seq normalized levels (`recoverChIPlevels`), and the average distribution (`producePEAKplots`) and heatmap (`producePEAKmaps`) around the peaks are shown. Below, genome-wide screenshots of the WT and KO/KD profiles on a locus containing peaks of this class were generated with the SeqCode `buildChIPprofile` function. Boxplots of ChIP-seq values are shown in log scale. All samples were normalized by the total number of reads of each experiment, except MTF2 ChIP-seq samples that were normalized by the total number of spike-in fruit-fly reads. Raw data were retrieved from[40–42].



The SeqCode Toolkit

SeqCode is a suite of open software command line tools to develop high-quality pictures from high-throughput sequencing samples. The source code is elegantly designed to be integrated into existing bioinformatics pipelines. SeqCode ^{web} is a web interface implemented to show only a selection of the features of the standalone distribution.

Figure 8. Homepage of the SeqCode website. The main menu of the web server is divided into four categories: NGS Tools (sequencing plots), Datasets (distribution plots), Gene Sets (intersection plots) and List Ops (operations with lists of identifiers). Additional help documentation and a section with downloading links are also provided here.

HOMER. To our knowledge, SeqCode is the only software in which all the routines can be applied over the reads aligned in a BAM file irrespectively of the class of sequencing experiment (e.g. paired-end or single-end). This allows users to easily add spike-in corrections to the data normalization of every class of experiment.

(iv) What families of graphical representations can be generated? The vast majority of applications (CGAT, ChAsE, Cistrome, Epidaurus, genomation, HOMER, and seqMINER) generate at least basic occupancy aggregated plots or heatmaps of ChIP-seq samples. However, the viewpoint is usually anchored to a single point (e.g. TSS or peak center), and representations along meta-gene bodies and genomic intervals are rarely available. Notably, certain tools (SeqCode, deepTools, EaSeq, and ngs.plot) are also able to generate heatmaps along gene bodies mimicking meta-gene aggregated plots. Finally, while annotation of genomic intervals into different features of the genes is available in most approaches, SeqCode provides highly customized graphics following multiple rules of genomic association, with different degrees of detail.

(v) What methods are available to evaluate the experimental signal strength? Functions for read counting are extremely useful for bioinformaticians. Visualization of genome-wide signal profiles reveals at first sight the outcome of fresh NGS experiments. Determination of the amount of signal of an experiment within particular intervals of genomic regions can be used to calculate how well two or more sets of peaks, genes, or experiments are correlated. Only a handful of tools provide options for quantification (SeqCode, deepTools, EaSeq, and HOMER), while SeqCode is the only one to provide options to normalize the results (e.g. profiles

Software	References
SeqCode	This article
BindDB	Livyatan et al.[81]
CGAT	Sims et al.[82]
ChAsE	Younesy et al.[83]
ChIPseeker	Yu et al.[84]
ChIP-seq tools	Ambrosini et al.[85]
CisGenome	Ji et al.[86]
Cistrome	Liu et al.[44], Zheng et al.[87]
deepTools	Ramirez et al.[47], Ramirez et al.[88]
EaSeq	Lerdrup et al.[43]
Epidaurus	Wang et al.[89]
GeneProf	Halbritter et al.[90]
Genomation	Akalin et al.[91]
HOMER	Heinz et al.[92]
ngs.plot	Shen et al.[45]
PyBedGraph	Zhang et al.[93]
seqMINER	Ye et al.[94]
Spark	Nielsen et al.[95]

Table 2. List of bioinformatics tools for graphical analysis of sequencing information.

and values) and to allow for spike-in correction of all routines. SeqCode (along with ChIP-seq, Cistrome and HOMER) also supports phylogenetic footprinting assessment of evolutionary information on sequencing data.

(vi) How efficient is the access to the information through the user interface? User-friendly websites (BindDB, ChIP-seq, Cistrome and Epidaurus) typically work without installing additional software. This means that elevated transfer time of samples through the Internet is a limiting factor, reducing the analysis over a static repository. In contrast, while local-based tools (such as SeqCode, CGAT, ChAsE, EaSeq, HOMER, ngs.plot and seqMINER) require initial setup, they are able to dynamically process user sequencing experiments with a much shorter time of response. Graphical interfaces are accessible, but command-line environments are more flexible for customizing each function and automatize batch processing. Hybrid software combining web servers and local interfaces is less frequent (ChIP-seq, Cistrome, and deepTools). We have designed the web interface of SeqCode with the aim of providing a graphical interface to command-line functions but also to build a solid educational platform to learn the basic analytical procedures of NGS data.

(vii) Which are the computational requirements to run this software? Web resources (BindDB, ChIP-seq, and Cistrome) appear to facilitate users to ignore processing power and memory capacity, but their use is limited in realistic scenarios. Standalone programs for professional sequencing analysis, although more efficient, exhibit distinct throughput, depending on their internal implementation. As a rule of thumb[48], executable binaries (SeqCode and EaSeq) are less time-consuming than Java-based systems (CisGenome, seqMINER, ChAsE), scripting approaches (CGAT, deepTools, HOMER), and libraries developed for R (genomation, ngs.plot, ChIPseeker). Strikingly, only a few methods pay attention to customizing the usage of memory (SeqCode, ChAsE, deepTools, and seqMINER), allowing the analysis to be performed in every computational configuration without affecting the biological conclusions of the experiments. Virtualization of these services and portability are key properties to allowing users to launch and test the software over any platform. Of note, only SeqCode and deepTools offer this service, through virtual machines and docker applications.

(viii) How easy is accessing information and getting assistance for users? Inexpert users could be interested in learning the basic use for each function, while more experienced users might wish to generate more complex analyses. However, the majority of the available programs provide minimal documentation addressing these issues (such as README files, manual of functions, working case tutorials, videos). In addition to SeqCode, we only found two other programs (deepTools and EaSeq) that provide comprehensive descriptions of programs and examples. SeqCode and EaSeq also integrate textual descriptions and additional pieces of information included with the final PDF files, which can be very useful for interpreting the final results.

Discussion

High-throughput experiments are fundamental to current research in molecular biology in the data-driven hypothesis paradigm. However, certain aspects of the analysis of sequencing still creates a barrier for the democratization of access to such technologies, to all level of users. For instance, computational performance requirements and storage capacities that are necessary to run most applications are still being improved[49–51]. Standardization of graphical data representation is also important[52]: most bioinformatics pipelines mainly focus on

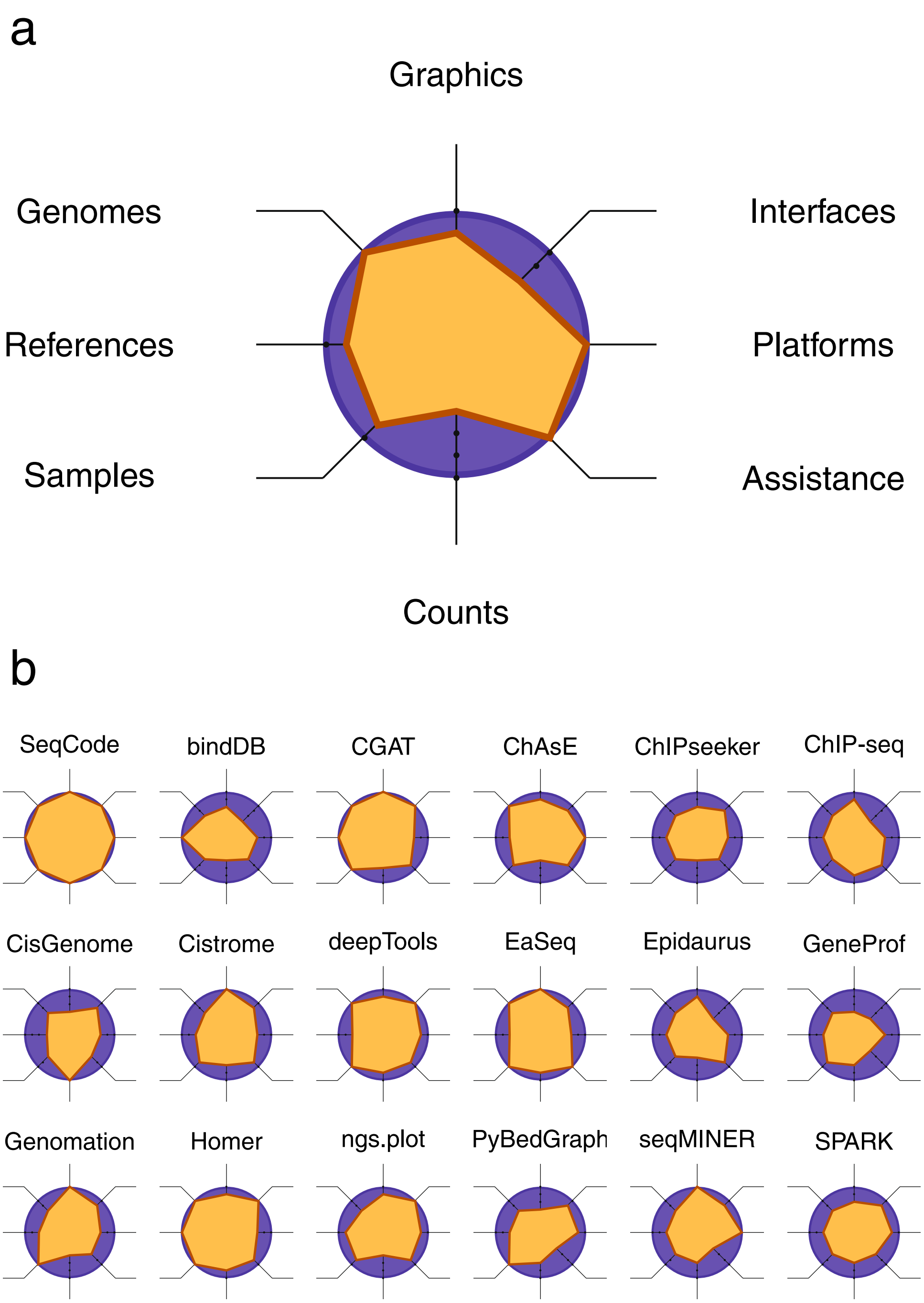


Figure 9. Evaluation of tools for the visualization of sequencing data. **(a)** Radar chart displaying the eight attributes that have been used to describe the characteristics of the existing applications. **(b)** Radar charts of each selected bioinformatics software for graphical summarizing sequencing data (see Table 3 for the values that are graphically displayed for each program here).

the mapping and counting stages, with moderate interest for visualization of the resulting information. It is very difficult for the bioinformatician to choose the appropriate software once the primary analysis of the sequencing experiment is done, as existing methods to generate graphical representations of the results do not completely solve the problem. Limited reproducibility among methods and the complex technical set-up of such tools further compromise their use. Indeed, when trying to recreate the bivalency analysis (Fig. 5) using other programs, we noted that certain functions are available in one program but not in another, and detected that most graphical representations of the same data are similar but not identical (Supp. Figure 12–14). Of note, recent efforts on formally defining a uniform framework for genomic data visualization promise to revolutionize this area of research. Thus, Gehlenborg and colleagues[53] classified genomic information into three distinct taxonomies: data (sparse/contiguous nature and interconnections), visualization (coordinate system, tracks, and views), and tasks

Feature	SeqCode	BindDB	CGAT	ChAsE	ChIPseeker	ChIP-seq	CisGenome	Cistrome	deepTools	EaSeq	Epidaurus	GeneProf	genomation	HOMER	ngs.plot	PyBedGraph	seqMINER	Spark
Genomes																		
1. Selection	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
2. All genomes	X		X	X					X	X				X				
3. All assemblies	X		X	X					X	X				X		X		
Reference																		
1. Gene	X	X	X					X			X	X		X				
2. Enhancer	X	X	X											X	X			
3. Genomic	X	X	X	X	X	X			X	X			X	X	X	X	X	X
Samples																		
1. ChIP-seq	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2. RNA-seq	X		X						X	X		X	X	X	X	X		
3. ATAC-seq	X		X	X				X	X	X			X	X		X		
Graphics																		
1. Aggregated	X		X	X		X		X	X	X	X		X	X	X		X	
2. Heatmaps	X	X	X	X				X	X	X	X		X		X		X	X
3. Distributions	X		X		X	X		X		X			X	X			X	
Counts																		
1. Custom tracks	X		X			X	X		X	X		X		X				X
2. Signal strength	X						X		X	X				X		X	X	
3. Conservation	X					X	X	X										
Interface																		
1. Local	X		X	X	X		X		`X	X			X	X	X	X	X	X
2. Flexible	X		X					X	X					X	X			
3. Dynamic	X		X	X	X		X	X	X	X			X	X	X	X	X	X
Platform																		
1. Executable	X			X						X							X	X
2. Parametric	X			X					X						X	X	X	
3. Portable	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X
Documents																		
1. Web site	X	X	X	X	X	X	X	X	X	X	X		X	X	X			X
2. Tutorials	X		X	X		X		X	X	X	X			X	X			
3. Self-docs	X									X								

Table 3. Characteristic features of SeqCode and other similar resources.

(searches and queries). Interestingly, such principles of design have been successfully applied to build interactive and scalable genomics visualizations[54].

We introduce SeqCode here to bridge this gap in visualizing processed data in a thorough manner. Our main purpose has been to build a software that performs most graphical operations on sequencing data with a quick response time and that is functional in most computational environments. One could argue that, biased for the type of representations constructed here (e.g. meta-plots, heatmaps, boxplots, etc.), SeqCode is more appropriate for analysis of local genomic information, which is produced in form of static high-resolution plots for publication. However, we consider that through the combination of genome-wide profiles and lists of biological features ranked by signal strength and gene associations generated by SeqCode, users will efficiently perform global analytical tasks on their NGS datasets too. Ultimately, the selection of a particular tool to generate a certain data visualization mostly depends on the biological question to address (Supp. Figure 15). Although several commands can be suitable to give response to a particular need, the pros and cons of each visualization, usually due the nature of the data graphically represented, must be rigorously considered. Thus, while genome-wide profiles displayed along chromosomes are a simple way to visually inspect experiments, systematic approaches such as meta-plots and heatmaps provide a more effective way to identify characteristic trends at a glance. On the other hand, average values reported in meta-plots can be seriously affected by the existence of genomic loci with an aberrant number of sequencing events[55]. Similarly, although heatmaps graphically represent the whole map of signal values and are extremely useful to highlight clusters of elements presenting distinct patterns, boxplots are necessary to study the distribution of signal strength along the full collection of elements (e.g. genes, enhancers). Pie charts, which are typically used to represent the abundance of a particular feature for different genomic regions, are recently under discussion and often substituted by bar plots[56]. We recommend to integrate each available visualization that might address a biological question to generate a consistent interpretation of the data.

The unit of information in SeqCode is the RefSeq annotation file, which is provided by most genome browsers; therefore, our applications are usable over every existing genome assembly. Notably, and in contrast to complex graphical and command-line interfaces based in configuration files, SeqCode provides simple but effective options to customize the final results. How to choose the appropriate palette of colors to generate the final graphical representation becomes a fundamental question[57]. We have also implemented options in our software to address the latest findings on analysis of NGS data. For instance, although the relationship between ChIP-seq signal strength and the functionality of a biological feature is still under investigation, significant progress has been based on performing quantitative comparisons with spike-in methods[58,59]. Thus, we allow users of SeqCode to correct the normalization of sequencing samples by introducing the spike-in factor in our applications. This option has recently proven to be useful to successfully evaluate different strategies of data normalization[60]. Indeed, proper graphical comparison of multiple sets of NGS samples corresponding to distinct biological scenarios (Fig. 7), and evaluation of consistency in the replicates of the same experiment coupled with quality control metrics[61] are increasingly becoming fundamental analytical tasks. Thus, we envisage that future developments in SeqCode will integrate tools to systematically automatize such set of comparative procedures.

SeqCode is entirely written in standard C, which makes this tool the best choice for fast visualization of the results, as compared to other programs designed with scripting languages that require more time to process huge volumes of information. Despite our efforts to cover the majority of graphical summary classes, we cannot discard, though, that we have omitted other graphical visualization modes or data formats. Generation of genome-wide profiles in bigwig format[62,63], for instance, should be considered in future versions. However, we believe that the open architectural design of SeqCode favors the easy incorporation of new components while preserving its common interface and performance. In fact, as the SeqCode source code is open and freely distributed through GitHub, other developers can introduce improvements and new functions in a simple manner to adapt it to specific issues. We have comprehensively scrutinized the portability of our software under distinct UNIX-like platforms (i.e. macOS and Linux). However, to approach users with different degrees of expertise and prevent potential issues due to different compiler versions in certain platforms, we have distributed our software in other running environments, such as Docker containers or Linux virtual machines loaded with a pre-installation of SeqCode inside. We have also designed a web interface to cover most SeqCode functions that incorporates additional useful tools for data plotting. We would like to stress that the interoperability provided through both SeqCode standalone and web platforms, will fulfil most needs of our potential audience. Therefore, SeqCode can be easily adopted as another component of a comprehensive computational pipeline for analyzing sequencing data, such as ChIP-seq, RNA-seq, or ATAC-seq, once quality control, read mapping, and counting have been performed. In sum, after this thorough evaluation of existing tools, we consider SeqCode to be the most interesting option for analysts of NGS data with different backgrounds. We have already successfully used SeqCode in multiple research projects[36,40,64–70].

Last but not least, education of future bioinformaticians is increasingly becoming a hot topic in the field[71–73], and we would like to underscore that we specifically designed an open web and command-line platform that can be integrated into any educational program focused on the high-throughput analysis of genomic data. Feedback on the use of our software by teachers for such a topic is strongly welcome.

Conclusions

To sum up, we consider that SeqCode will become a fundamental tool to deal with NGS experiments that require a fast and complete analysis and will be the most valuable option for a wide range of users.

Methods

The full source code distribution of SeqCode can be downloaded from GitHub (https://github.com/eblancoga/seqcode). Our software during the execution initially loads the information from the chromosome size and RefSeq transcript files provided by the user to construct an image of each chromosome in memory. Chromosomes are divided into series of consecutive fragments of the same size, which are termed bins. The size of the bin is configurable by the user. This segmentation is therefore useful during the calculations of values for all the graphical representations generated by SeqCode. Our programs internally integrate the SamTools and HTS C libraries[18] in order to efficiently read BAM and SAM files. The SeqCode web site (http://ldicrocelab.crg.eu/) is implemented in PHP and the output of graphical services is provided in PNG and PDF formats. Basic R core packages are utilized to produce the resulting plots of all SeqCode functions.

Raw data of the examples introduced throughout the article (Table 4) were retrieved from the NCBI GEO repository[74]. Bowtie[7] was used to map the reads over the genome of mouse (mm9 and mm10), human (hg38), and the fruit fly (dm3). Unaligned reads must be excluded from the resulting BAM files that will be processed by SeqCode commands. To ensure a fair comparison among samples, SeqCode performed the normalization of values calculated within any finite element (e.g. regions, bins, etc.) for a given sequencing experiment by sequencing depth (i.e. total number of reads in the BAM file). When spike-in material was included into the ChIP-seq experiment (e.g. MTF2 samples in Fig. 7), the number of reads mapped over the spike-in genome was instead

GEO accession	Sample	Cell/Tissue	References
SRX367147	H3K4me3	mESCs	Tee et al.[32]
SRX336228	Jarid2		
GSM1019769	H3K36me3		Ballare et al.[33]
GSM1019772	H3K27me3		
GSM1019771	Suz12		
GSM850467	Ser5P		Brookes et al.[38]
GSM850470	Ser2P		
GSM850471	H2Aub1		
GSM1562339	RNA-seq		Jacinto et al.[75]
GSM1526287	H3K27ac		Ji et al.[79]
GSM2098958	H3K4me3		Beringer et al.[36]
GSM2098952	H3K27me3		
GSM1041372	Ring1b		Morey et al.[37]
GSM2645501	MLL2 WT		Mas et al.[40]
GSM2645502	MLL2 KO		
GSM2645495	H3K4me3 WT		
GSM2645496	H3K4me3 KO		
GSM4616798	RING1B WT	A673	Sánchez et al.[41]
GSM4616801	RING1B KD		
GSM4616799	EWSR1-FLI1 WT		
GSM4616802	EWSR1-FLI1 KD		
GSM4023634	PHF19 WT	DU145	Jain et al.[42]
GSM4023635	PHF19 KD		
GSM4023643	MTF2 WT		
GSM4023644	MTF2 KD		
GSM593408	H3K4me3	Wing imaginal disc (fly)	Perez-Lluch et al.[96]
GSM593409	H3K27me3		
GSM593410	H3K36me3		
GSM593412	Ser5P		
GSM593411	Ser2P		
GSM1182471	H3K4me1		Perez-Lluch et al.[97]
GSM1182472	H3K27ac		
GSM1363590	H3K9ac		
GSM1060715	RNA-seq		
GSM1005586	Cabut		Ruiz-Romero et al.[98]

Table 4. List of GEO accession codes of the raw data analyzed in this work.

used to normalize the data. MACS[12] was used to perform the peak calling of the ChIP-seq experiments. Custom tracks generated with the `buildChIPprofile` function of SeqCode follow the 0-start, half-open coordinate counting system. Screenshots were captured with the UCSC genome browser[14]. To speed up the generation of plots with `produceTSSplots`, `produceGENEplots`, `produceTSSmaps`, and `computeChIPlevels` web services for the genes provided by the user, visualization data of the whole catalog of genes were previously generated on each predefined ChIP-seq experiment with the option –g of such commands (Supplementary information).

Data availability

The SeqCode website is hosted at http://ldicrocelab.crg.eu. The SeqCode software is freely available under the GNU General Public License v3.0. Source code has been deposited at the GitHub repository (https://github.com/eblancoga/seqcode). To improve the portability of the software, we also provide Oracle VM virtualbox™ appliances and Docker containers built on Linux Ubuntu MATE at our web site (http://ldicrocelab.crg.eu/06_Downloads/). The datasets analyzed during the current study are available in the GEO repository (see Table 4).

Received: 14 March 2021; Accepted: 20 August 2021
Published online: 01 October 2021

References

1. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief Bioinform.* **18**, 279–290. https://doi.org/10.1093/bib/bbw023 (2017).

2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: The teenage years. *Nat. Rev. Genet.* **20**, 631–656. https://doi.org/10.1038/s41576-019-0150-2 (2019).
3. Sun, Y., Miao, N. & Sun, T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* **156**, 29. https://doi.org/10.1186/s41065-019-0105-9 (2019).
4. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
5. Chen, S. *et al.* AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinform.* **18**, 80. https://doi.org/10.1186/s12859-017-1469-3 (2017).
6. Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354 (2016).
7. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25. https://doi.org/10.1186/gb-2009-10-3-r25 (2009).
8. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595. https://doi.org/10.1093/bioinformatics/btp698 (2010).
9. Marco-Sola, S., Sammeth, M., Guigo, R. & Ribeca, P. The GEM mapper: Fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188. https://doi.org/10.1038/nmeth.2221 (2012).
10. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75. https://doi.org/10.1038/nbt.1518 (2009).
11. Zang, C. *et al.* A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958. https://doi.org/10.1093/bioinformatics/btp340 (2009).
12. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137. https://doi.org/10.1186/gb-2008-9-9-r137 (2008).
13. Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751. https://doi.org/10.1093/nar/gky1113 (2019).
14. Haeussler, M. *et al.* The UCSC genome browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858. https://doi.org/10.1093/nar/gky1095 (2019).
15. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU epigenome browser update 2019. *Nucleic Acids Res.* **47**, W158–W165. https://doi.org/10.1093/nar/gkz348 (2019).
16. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform.* **14**, 178–192. https://doi.org/10.1093/bib/bbs017 (2013).
17. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402. https://doi.org/10.1093/nar/gkaa434 (2020).
18. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
19. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. https://doi.org/10.1093/bioinformatics/btq033 (2010).
20. Zhao, H. *et al.* CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007. https://doi.org/10.1093/bioinformatics/btt730 (2014).
21. Li, P. E. *et al.* Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. *Nucleic Acids Res.* **45**, 67–80. https://doi.org/10.1093/nar/gkw1027 (2017).
22. Gong, H., Yang, Y., Zhang, S., Li, M. & Zhang, X. Application of Hi-C and other omics data analysis in human cancer and cell differentiation research. *Comput. Struct. Biotechnol. J.* **19**, 2070–2083. https://doi.org/10.1016/j.csbj.2021.04.016 (2021).
23. Goodstadt, M. & Marti-Renom, M. A. Challenges for visualizing three-dimensional data in genomic browsers. *FEBS Lett.* **591**, 2505–2519. https://doi.org/10.1002/1873-3468.12778 (2017).
24. Ing-Simmons, E. & Vaquerizas, J. M. Visualising three-dimensional genome organisation in two dimensions. *Development* https://doi.org/10.1242/dev.177162 (2019).
25. Yardimci, G. G. & Noble, W. S. Software tools for visualizing Hi-C data. *Genome Biol.* **18**, 26. https://doi.org/10.1186/s13059-017-1161-y (2017).
26. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745. https://doi.org/10.1093/nar/gkv1189 (2016).
27. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph* **20**, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248 (2014).
28. Feitelson, D. G. *Comparing Partitions with Spie Charts* (The Hebrew University of Jerusalem, 2003).
29. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050. https://doi.org/10.1101/gr.3715005 (2005).
30. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220. https://doi.org/10.1038/s41576-018-0089-8 (2019).
31. Schuettengruber, B., Bourbon, H. M., Di Croce, L. & Cavalli, G. Genome regulation by polycomb and trithorax: 70 years and counting. *Cell* **171**, 34–57. https://doi.org/10.1016/j.cell.2017.08.002 (2017).
32. Tee, W. W., Shen, S. S., Oksuz, O., Narendra, V. & Reinberg, D. Erk1/2 activity promotes chromatin features and RNAPII phosphorylation at developmental promoters in mouse ESCs. *Cell* **156**, 678–690. https://doi.org/10.1016/j.cell.2014.01.009 (2014).
33. Ballare, C. *et al.* Phf19 links methylated Lys36 of histone H3 to regulation of Polycomb activity. *Nat. Struct. Mol. Biol.* **19**, 1257–1265. https://doi.org/10.1038/nsmb.2434 (2012).
34. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326. https://doi.org/10.1016/j.cell.2006.02.041 (2006).
35. Blanco, E., Gonzalez-Ramirez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The bivalent genome: Characterization, structure, and regulation. *Trends Genet.* **36**, 118–131. https://doi.org/10.1016/j.tig.2019.11.004 (2020).
36. Beringer, M. *et al.* EPOP functionally links elongin and polycomb in pluripotent stem cells. *Mol. Cell* **64**, 645–658. https://doi.org/10.1016/j.molcel.2016.10.018 (2016).
37. Morey, L., Aloia, L., Cozzuto, L., Benitah, S. A. & Di Croce, L. RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. *Cell Rep.* **3**, 60–69. https://doi.org/10.1016/j.celrep.2012.11.026 (2013).
38. Brookes, E. *et al.* Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. *Cell Stem Cell* **10**, 157–170. https://doi.org/10.1016/j.stem.2011.12.017 (2012).
39. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: From the discovery of RNA polymerase pausing to its functional implications. *Curr. Opin. Cell Biol.* **46**, 72–80. https://doi.org/10.1016/j.ceb.2017.03.002 (2017).
40. Mas, G. *et al.* Promoter bivalency favors an open chromatin architecture in embryonic stem cells. *Nat. Genet.* **50**, 1452–1462. https://doi.org/10.1038/s41588-018-0218-5 (2018).
41. Sanchez-Molina, S. *et al.* RING1B recruits EWSR1-FLI1 and cooperates in the remodeling of chromatin necessary for Ewing sarcoma tumorigenesis. *Sci. Adv.* https://doi.org/10.1126/sciadv.aba3058 (2020).
42. Jain, P., Ballare, C., Blanco, E., Vizan, P. & Di Croce, L. PHF19 mediated regulation of proliferation and invasiveness in prostate cancer cells. *Elife* https://doi.org/10.7554/eLife.51373 (2020).

43. Lerdrup, M., Johansen, J. V., Agrawal-Singh, S. & Hansen, K. An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nat. Struct. Mol. Biol.* **23**, 349–357. https://doi.org/10.1038/nsmb.3180 (2016).
44. Liu, T. *et al.* Cistrome: An integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83. https://doi.org/10.1186/gb-2011-12-8-r83 (2011).
45. Shen, L., Shao, N., Liu, X. & Nestler, E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284. https://doi.org/10.1186/1471-2164-15-284 (2014).
46. Lawrence, T. J. *et al.* FAST: FAST analysis of sequences toolbox. *Front. Genet.* **6**, 172. https://doi.org/10.3389/fgene.2015.00172 (2015).
47. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187-191. https://doi.org/10.1093/nar/gku365 (2014).
48. Perkel, J. M. Julia: Come for the syntax, stay for the speed. *Nature* **572**, 141–142. https://doi.org/10.1038/d41586-019-02310-3 (2019).
49. Ravanmehr, V., Kim, M., Wang, Z. & Milenkovic, O. ChIPWig: A random access-enabling lossless and lossy compression method for ChIP-seq data. *Bioinformatics* **34**, 911–919. https://doi.org/10.1093/bioinformatics/btx685 (2018).
50. Roguski, L., Ochoa, I., Hernaez, M. & Deorowicz, S. FaStore: A space-saving solution for raw sequencing data. *Bioinformatics* **34**, 2748–2756. https://doi.org/10.1093/bioinformatics/bty205 (2018).
51. Wang, Z., Weissman, T. & Milenkovic, O. smallWig: Parallel compression of RNA-seq WIG files. *Bioinformatics* **32**, 173–180. https://doi.org/10.1093/bioinformatics/btv561 (2016).
52. Weissgerber, T. L. *et al.* Data visualization, bar naked: A free tool for creating interactive graphics. *J. Biol. Chem.* **292**, 20592–20598. https://doi.org/10.1074/jbc.RA117.000147 (2017).
53. Nusrat, S., Harbig, T. & Gehlenborg, N. Tasks, techniques, and tools for genomic data visualization. *Comput. Graph Forum* **38**, 781–805. https://doi.org/10.1111/cgf.13727 (2019).
54. Yi, S. L., Lekschas, F. & Gehlenborg, N. Gosling: A grammar-based toolkit for scalable and interactive genomics data visualization. *OSF Preprints*. https://doi.org/10.31219/osf.io/6evmb (2021).
55. Wreczycka, K. *et al.* HOT or not: Examining the basis of high-occupancy target regions. *Nucleic Acids Res.* **47**, 5735–5745. https://doi.org/10.1093/nar/gkz460 (2019).
56. Spence, I. No humble pie: The origins and usage of a statistical chart. *J. Educ. Behav. Stat.* **30**, 353–368 (2005).
57. Hattab, G., Rhyne, T. M. & Heider, D. T. simple rules to colorize biological data visualization. *PLoS Comput. Biol.* **16**, e1008259. https://doi.org/10.1371/journal.pcbi.1008259 (2020).
58. Bonhoure, N. *et al.* Quantifying ChIP-seq data: A spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.* **24**, 1157–1168. https://doi.org/10.1101/gr.168260.113 (2014).
59. Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.* **9**, 1163–1170. https://doi.org/10.1016/j.celrep.2014.10.018 (2014).
60. Blanco, E., Di Croce, L. & Aranda, S. SpikChIP: A novel computational methodology to compare multiple ChIP-seq using spike-in chromatin. *NAR Genom. Bioinform.* **3**, lqab064. https://doi.org/10.1093/nargab/lqab064 (2021).
61. Muller, L., Gerighausen, D., Farman, M. & Zeckzer, D. Sierra platinum: A fast and robust peak-caller for replicated ChIP-seq experiments with visual quality-control and -steering. *BMC Bioinform.* **17**, 377. https://doi.org/10.1186/s12859-016-1248-6 (2016).
62. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207. https://doi.org/10.1093/bioinformatics/btq351 (2010).
63. Pohl, A. & Beato, M. bwtool: A tool for bigWig files. *Bioinformatics* **30**, 1618–1619. https://doi.org/10.1093/bioinformatics/btu056 (2014).
64. Aranda, S. *et al.* Chromatin capture links the metabolic enzyme AHCY to stem cell proliferation. *Sci. Adv.* **5**, eaav2448. https://doi.org/10.1126/sciadv.aav2448 (2019).
65. Castano, J. *et al.* GATA2 promotes hematopoietic development and represses cardiac differentiation of human mesoderm. *Stem Cell Rep.* **13**, 515–529. https://doi.org/10.1016/j.stemcr.2019.07.009 (2019).
66. Jain, P., Ballare, C., Blanco, E., Vizan, P. & Di Croce, L. PHF19 mediated regulation of proliferation and invasiveness in prostate cancer cells. *Elife* **9**, e51373. https://doi.org/10.7554/eLife.51373 (2020).
67. Martone, J. *et al.* Trans-generational epigenetic regulation associated with the amelioration of Duchenne muscular dystrophy. *EMBO Mol. Med.* **11**, e12063. https://doi.org/10.15252/emmm.202012063 (2020).
68. Morey, L. *et al.* Polycomb regulates mesoderm cell fate-specification in embryonic stem cells through activation and repression mechanisms. *Cell Stem Cell* **17**, 300–315. https://doi.org/10.1016/j.stem.2015.08.009 (2015).
69. Pascual-Reguant, L. *et al.* Lamin B1 mapping reveals the existence of dynamic and functional euchromatin lamin B1 domains. *Nat. Commun.* **9**, 3420. https://doi.org/10.1038/s41467-018-05912-z (2018).
70. Santanach, A. *et al.* The Polycomb group protein CBX6 is an essential regulator of embryonic stem cell identity. *Nat. Commun.* **8**, 1235. https://doi.org/10.1038/s41467-017-01464-w (2017).
71. Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A. & Schneider, M. V. A global perspective on evolving bioinformatics and data science training needs. *Brief Bioinform.* **20**, 398–404. https://doi.org/10.1093/bib/bbx100 (2019).
72. Sims, D., Ponting, C. P. & Heger, A. CGAT: A model for immersive personalized training in computational genomics. *Brief Funct. Genomics* **15**, 32–37. https://doi.org/10.1093/bfgp/elv021 (2016).
73. Zhan, Y. A. *et al.* Fostering bioinformatics education through skill development of professors: Big genomic data skills training for professors. *PLoS Comput. Biol.* **15**, e1007026. https://doi.org/10.1371/journal.pcbi.1007026 (2019).
74. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets–update. *Nucleic Acids Res.* **41**, D991-995. https://doi.org/10.1093/nar/gks1193 (2013).
75. Jacinto, F. V., Benner, C. & Hetzer, M. W. The nucleoporin Nup153 regulates embryonic stem cell pluripotency through gene silencing. *Genes Dev.* **29**, 1224–1238. https://doi.org/10.1101/gad.260919.115 (2015).
76. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319. https://doi.org/10.1016/j.cell.2013.03.035 (2013).
77. Benayoun, B. A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673–688. https://doi.org/10.1016/j.cell.2014.06.027 (2014).
78. Khan, A. *et al.* JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266. https://doi.org/10.1093/nar/gkx1126 (2018).
79. Ji, X. *et al.* Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proc. Natl. Acad. Sci. U S A* **112**, 3841–3846. https://doi.org/10.1073/pnas.1502971112 (2015).
80. Kuleshov, M. V. *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-97. https://doi.org/10.1093/nar/gkw377 (2016).
81. Livyatan, I., Aaronson, Y., Gokhman, D., Ashkenazi, R. & Meshorer, E. BindDB: An integrated database and webtool platform for "reverse-ChIP" epigenomic analysis. *Cell Stem Cell* **17**, 647–648. https://doi.org/10.1016/j.stem.2015.11.015 (2015).
82. Sims, D. *et al.* CGAT: computational genomics analysis toolkit. *Bioinformatics* **30**, 1290–1291. https://doi.org/10.1093/bioinformatics/btt756 (2014).
83. Younesy, H. *et al.* ChAsE: Chromatin analysis and exploration tool. *Bioinformatics* **32**, 3324–3326. https://doi.org/10.1093/bioinformatics/btw382 (2016).

84. Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145 (2015).
85. Ambrosini, G., Dreos, R., Kumar, S. & Bucher, P. The ChIP-Seq tools and web server: A resource for analyzing ChIP-seq and other types of genomic data. *BMC Genomics* **17**, 938. https://doi.org/10.1186/s12864-016-3288-8 (2016).
86. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* **26**, 1293–1300. https://doi.org/10.1038/nbt.1505 (2008).
87. Zheng, R. *et al.* Cistrome data browser: Expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* **47**, D729–D735. https://doi.org/10.1093/nar/gky1094 (2019).
88. Ramirez, F. *et al.* deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165. https://doi.org/10.1093/nar/gkw257 (2016).
89. Wang, L. *et al.* Epidaurus: Aggregation and integration analysis of prostate cancer epigenome. *Nucleic Acids Res.* **43**, e7. https://doi.org/10.1093/nar/gku1079 (2015).
90. Halbritter, F., Kousa, A. I. & Tomlinson, S. R. GeneProf data: A resource of curated, integrated and reusable high-throughput genomics experiments. *Nucleic Acids Res.* **42**, D851-858. https://doi.org/10.1093/nar/gkt966 (2014).
91. Akalin, A., Franke, V., Vlahovicek, K., Mason, C. E. & Schubeler, D. Genomation: A toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics* **31**, 1127–1129. https://doi.org/10.1093/bioinformatics/btu775 (2015).
92. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 (2010).
93. Zhang, H. B., Kim, M., Chuang, J. H. & Ruan, Y. pyBedGraph: A python package for fast operations on 1D genomic signal tracks. *Bioinformatics* **36**, 3234–3235. https://doi.org/10.1093/bioinformatics/btaa061 (2020).
94. Ye, T. *et al.* seqMINER: An integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* **39**, e35. https://doi.org/10.1093/nar/gkq1287 (2011).
95. Nielsen, C. B. *et al.* Spark: A navigational paradigm for genomic data exploration. *Genome Res.* **22**, 2262–2269. https://doi.org/10.1101/gr.140665.112 (2012).
96. Perez-Lluch, S. *et al.* Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. *Nucleic Acids Res.* **39**, 4628–4639. https://doi.org/10.1093/nar/gkq1322 (2011).
97. Perez-Lluch, S. *et al.* Absence of canonical marks of active chromatin in developmentally regulated genes. *Nat. Genet.* **47**, 1158–1167. https://doi.org/10.1038/ng.3381 (2015).
98. Ruiz-Romero, M., Blanco, E., Paricio, N., Serras, F. & Corominas, M. Cabut/dTIEG associates with the transcription factor Yorkie for growth control. *EMBO Rep.* **16**, 362–369. https://doi.org/10.15252/embr.201439193 (2015).

Acknowledgements

We thank all the members of Di Croce laboratory for their invaluable feedback on the improvement of the software and helpful discussions, and V. A. Raker for scientific editing.

Author contributions

E.B. conceived, designed, and implemented the SeqCode software in ANSI C. E.B. and M.G.R. designed the website of SeqCode and the PHP interfaces of all functions. L.D.C. evaluated the software and provided valuable technical feedback and the computational platforms for analyses. E.B., M.G.R., and L.D.C. contributed in writing and revising the manuscript.

Funding

The work in the Di Croce laboratory is supported by grants from the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) (PID2019-108322 GB-I00), Fundación Vencer El Cancer (VEC), and the Catalan Agency for Management of University and Research Grants (AGAUR). We acknowledge support from the Spanish Ministry of Science and Innovation to the EMBL partnership Centro de Excelencia Severo Ochoa, and the CERCA Programme/Generalitat de Catalunya.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-98889-7.

Correspondence and requests for materials should be addressed to E.B. or L.D.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.