DOI 10.1186/s12859-017-1698-5

BMC Bioinformatics

SOFTWARE Open Access

NaviSE: superenhancer navigator integrating epigenomics signal algebra

Alex M. Ascensión[1,2,3], Mikel Arrospide-Elgarresta[1], Ander Izeta[2*] and Marcos J. Araúzo-Bravo[1,4*]

Abstract

Background: Superenhancers are crucial structural genomic elements determining cell fate, and they are also involved in the determination of several diseases, such as cancer or neurodegeneration. Although there are pipelines which use independent pieces of software to predict the presence of superenhancers from genome-wide chromatin marks or DNA-interaction protein binding sites, there is not yet an integrated software tool that processes automatically algebra combinations of raw data sequencing into a comprehensive final annotated report of predicted superenhancers.

Results: We have developed NaviSE, a user-friendly streamlined tool which performs a fully-automated parallel processing of genome-wide epigenomics data from sequencing files into a final report, built with a comprehensive set of annotated files that are navigated through a graphic user interface dynamically generated by NaviSE. NaviSE also implements an 'epigenomics signal algebra' that allows the combination of multiple activation and repression epigenomics signals. NaviSE provides an interactive chromosomal landscaping of the locations of superenhancers, which can be navigated to obtain annotated information about superenhancer signal profile, associated genes, gene ontology enrichment analysis, motifs of transcription factor binding sites enriched in superenhancers, graphs of the metrics evaluating the superenhancers quality, protein-protein interaction networks and enriched metabolic pathways among other features. We have parallelised the most time-consuming tasks achieving a reduction up to 30% for a 15 CPUs machine. We have optimized the default parameters of NaviSE to facilitate its use. NaviSE allows different entry levels of data processing, from sra-fastq files to bed files; and unifies the processing of multiple replicates. NaviSE outperforms the more time-consuming processes required in a non-integrated pipeline. Alongside its high performance, NaviSE is able to provide biological insights, predicting cell type specific markers, such as *SOX2* and *ZIC3* in embryonic stem cells, *CDK5R1* and *REST* in neurons and *CD86* and *TLR2* in monocytes.

Conclusions: NaviSE is a user-friendly streamlined solution for superenhancer analysis, annotation and navigation, requiring only basic computer and next generation sequencing knowledge. NaviSE binaries and documentation are available at: https://sourceforge.net/projects/navise-superenhancer/.

Keywords: Superenhancers, Next-generation sequencing, Parallel computing, Epigenomics, Computational biology, Graphics user interface, Signal algebra

*Correspondence: ander.izeta@biodonostia.org; mararabra@yahoo.co.uk
[1] Computational Biology and Systems Biomedicine, Biodonostia Health Research Institute, 20014 San Sebastián, Spain
[2] Tissue Engineering Laboratory, Bioengineering Area, Biodonostia Health Research Institute, 20014 San Sebastián, Spain
[4] IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain
Full list of author information is available at the end of the article

Background

Superenhancers (SEs) are a novel class of transcription regulatory DNA regions with unusually strong enrichment for binding of transcriptional coactivators such as Mediator of RNA polymerase II transcription subunit 1 (MED1), activation histone marks such as H3K27ac, or cell and tissue-specific transcription factors (TFs) [1]. As a result, SEs represent large clusters of transcriptional enhancers that drive the expression of 'master control' genes that define cell identity. SEs differ from typical enhancers (TEs) for enclosing higher TF binding density and number of TF binding sites (TFBSs), which correlate with a much higher expression of their target genes [2]. Since SEs determine cell fate and gene expression regulation [3], they are related to altered expression of genes contributing to diseases such as Alzheimer or systemic lupus erythematosus [4]. Aberrant DNA methylation patterns in SEs, as well as SE-associated gene sets, have also been found to be altered in cancer [5–7].

Although protocols for computational prediction of SEs already exist [4], there is yet no tool that integrates all the processing stages from the raw data reads generated by the sequencer, through quality control and reads alignment, to peak estimation and peak stitching, ending with a fully annotated and interactive documentation of the results.

Furthermore, although SEs were initially predicted with MED1 [4] and activation histone marks such as H3K27ac, which has been proposed as a proxy for their estimation [2], the combination of several activation and repression epigenomics marks could help sharpen SE predictions. Therefore, we have designed NaviSE to use data with a wide range of chromatin status information, being able to process raw data from Assay for Transposase Accessible Chromatin (ATAC-seq) and DNase I hypersensitive sites (DHSs) experiments, apart from the usual ChIP-seq signals. In the case of other signals such as DNA methylation, NaviSE is prepared to integrate their information to perform SE predictions with the only condition that the user provides such data in bed or bam files, such as the bam files produced by the Parallel Processing Pipeline software for automatic analysis of Bisulfite Sequencing data (P3BSseq) [8].

On the other hand, there are no computational tools neither integrating several epigenomics signals simultaneously, nor performing signal algebra. Moreover, CPU resources and running-time are crucial for the high quantity of data produced by Next Generation Sequencing (NGS) technologies, hence another of the main demands in NGS software development is the parallelisation of the most time-consuming processes.

To meet all these demands, we have developed NaviSE, a user-friendly tool which automatically processes and integrates multiple genome-wide NGS epigenomics signals from various input file formats into an interactive HTML report, built with annotations about SEs, such as associated genes, gene ontology (GO), graphs with metrics and statistical analysis, integrating all the data into the Graphical User Interface (GUI) to navigate through all the results. NaviSE parallelises the most relevant and time-consuming processes to optimise them, running multiple analysis in a significantly reduced amount of time. Finally, NaviSE is developed for users with working knowledge in informatics.

Implementation

Preprocessing of NGS files

Before the determination of SEs, NaviSE prepares the raw data, allowing multiple replicates and controls at once. The main steps for such preprocessing are as follows:

1. *Input format file recognition and file processing*: NaviSE recognizes multiple file formats, e.g., .sra, .fastq, .sam, .bam and .bed, and transforms an *upstream* format (.sra, .fastq, .sam) into a .bam file. In the absence of *upstream* files, *downstream* .bed files are also processed to .bam files.
2. *Alignments*: Performed by default with bowtie2 [9], .sam files are processed to .bam files by samtools. NaviSE also allows read alignment with MOSAIK [10], STAR [11] and BWA [12] aligners. Furthermore, users may generate their own .sam or .bam files with other aligners, and NaviSE will recognize these files for further processing.
3. *Quality control with FastQC*: NaviSE performs the quality analysis of the reads from the .fastq files using FastQC to create a report including several quality parameters, such as per base quality, GC content, *k*-mers distribution or presence of adapters.
4. *Combination of replicates and peak calling with MACS*: If there is more than one replicate or control, NaviSE will combine all the associated .bam files into one, and calculate the signal peaks with MACS (Model-based Analysis for ChIP-Seq) [13]. If control files are introduced for background correction, NaviSE configures MACS to use the control signal to calculate the peaks from the sample. Conversely, if no control is introduced, NaviSE configures MACS to use a pre-calculated background.

SE prediction and annotation

Once the data is preprocessed, a SE prediction and ranking is performed. SEs then are further analysed in search of SE related genes, DNA sequence motifs, GO terms or statistical estimations.

1. *Epigenomics signal algebra*: In case more than one epigenomic signal was used to predict SEs, NaviSE integrates all the signals to improve the SE prediction.

The way in which different epigenomic signals are combined is defined by the names of the signal data files $Sig \in$ {H3K27ac, H3K4me1, H3K4me3, H3K9me3, H3K27me3, ATAC-seq, DHS,...} separated by signal operators $Ope \in$ {AND, OR, NOT, XOR, +, - SYM}.

The way these algebra operators have been adapted to operate over pairs of genomic signals is illustrated in Fig. 1. To invoke this algebra, NaviSE is called writing these signal and operators as additional arguments in the command line:

$$Sig_1\ Ope_{1,2}\ Sig_2\ Ope_{2,3}\ Sig_3\ Ope_{3,4}\ Sig_4$$

where Sig_i is the name of the file containing the epigenomics data of a type of signal i, and $Ope_{i,i+1}$ is the pairwise signal operator applied to combine i and $i+1$ signals.

When performing 'epigenomics signal algebra', NaviSE picks the first pair of signals separated by each operator starting from the left ($Sig_1\ Ope_{1,2}\ Sig_2$). Once an operation is processed, its results are combined with the next signal using the next operator (($Sig_1\ Ope_{1,2}\ Sig_2$) $Ope_{2,3}\ Sig_3$). This process continues from left to right side recursively until the last signal identifier is reached. To speed up the process, for each pair of signals NaviSE searches first all their overlapping regions and performs the signal operator only over these regions.

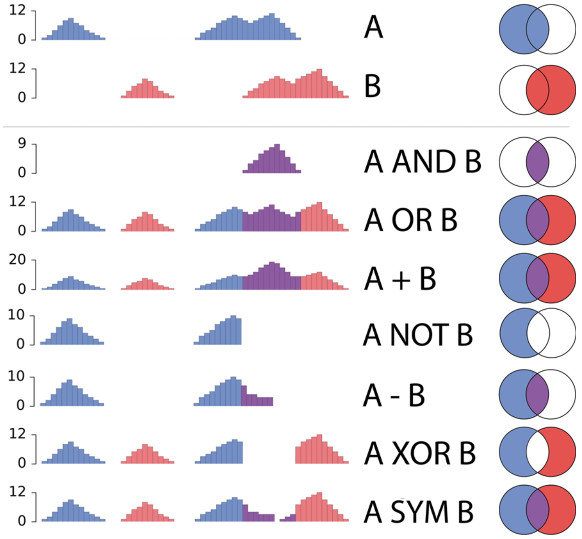


Fig. 1 Epigenomics signal algebra. Schemes of the different pairwise operations implemented in NaviSE. The top two rows depict an example of the two epigenomic signals to be combined, and the remaining rows illustrate the signal profile after applying the respective operator. Simplified Euler-Venn diagrams given in the rightmost column illustrate the set operations. AND, OR, NOT and XOR are Boolean operations which do not change the signal pileup; whereas +, -, and SYM are arithmetic operations which can change the signal pileup. In the case of - and SYM, negative pileups are transformed into zeros

2. *SE prediction*: To predict the SEs in a sample, NaviSE performs the *stitching* of MACS peaks which fall within a threshold distance, using our own implementation of the algorithm developed by Young's lab [2], in which MACS peaks (inferred as enhancers) are stitched according to a constant distance (12.5 kb by default) criterion algorithm, in case the distance between the end of one MACS peak and the start of the following peak is less than the established threshold, they are *stitched* as a single peak. Then, NaviSE ranks the *stitched* peaks with a score based on the measured signal level within the *stitched* region.

NaviSE assigns a score to each *stitched enhancer*, considering that a *stitched enhancer* with a higher number of bam reads has a higher SE predictive value. Thus, to build the SE ranking, NaviSE takes the raw reads from the .bam files, and for each *stitched enhancer* it collects all the reads over the *stitched enhancer* support. This support is defined by the DNA sequence lying between the *stitched enhancer* start, $STIT_{start}$, and the *stitched enhancer* end, $STIT_{end}$, nucleotide positions. Then, we define the *stitched enhancer* count, $Count_{STIT}$, as the cumulative sum of the bam reads throughout the *stitched enhancer* support:

$$Count_{STIT} = \sum_{i=1}^{N_{reads}} \sum_{j=STIT_{start}}^{STIT_{end}} read_i(j) \qquad (1)$$

where $read_i(j)$ indicates whether a bam read i, from the set of N_{reads}, lies at the position j of the *stitched enhancer* within the support $[STIT_{start}, STIT_{end}]$. Therefore, $read_i(j) = 1$ if a nucleotide of the bam read i is mapped to the location j of the *stitched enhancer*, and $read_i(j) = 0$ otherwise.

Then, the SE ranking, r, is defined as the sorted list of $Count_{STIT}$ in descending order:

$$r = \text{sort}_{\blacktriangledown}\{Count_{STIT}\} \qquad (2)$$

thus, Eq. 2 assigns position one in the ranking to the *stitched enhancer* with the highest $Count_{STIT}$, position two in the ranking to the *stitched enhancer* with second highest $Count_{STIT}$, etc. until we reach the *stitched enhancer* with the lowest $Count_{STIT}$.

The next step is the determination of the SE threshold (θ_{SE}), the position of the ranking for which the *stitched enhancers* whose rank is below θ_{SE} will be considered as SEs, and TEs otherwise. To determine θ_{SE}, we scale both $Count_{STIT}$ and r between 0 and 1. Then, we determine θ_{SE} as the position of r whose slope is nearest to 45°.

3. *SE gene assignment*: Once the SE locations are determined, each SE is assigned a gene by proximity with the closest transcription start site (TSS). NaviSE also includes information about genes overlapping the SE or genes proximal to each SE.
4. *Subpeak annotation*: The SEs and TEs subpeaks have been shown to act synergistically within the SE despite being individual and independent structures [14]. To provide detailed information about the SE subpeaks structure and location, NaviSE performs a structural annotation of the subpeaks that represent each SE. The annotation contains the following parameters:

   - Number of subpeaks, *loci* and TSS locations.
   - Association to TSSs: Due to the TSS specific regulation role, a SE inside a TSS might not exert the role of SE itself. Thus, to understand the regulatory role of the SEs, it is important to resolve their association to TSSs. This analysis is portrayed by two related values: (i) the *Percentage OUTS*, which is the percentage of subpeaks outside the range of the user-defined distance within the TSS, and (ii) the *Enhancer Type*, a classification of the SE according to *Percentage OUTS*. The categories assigned to *Enhancer Type* are labeled as *Pure* if all the subpeaks are outside the TSS, *Only TSS* if all the subpeaks lay within the TSS, and *Mixed* if there are both types of subpeaks.

5. *Automatic generation of SE peak distribution profiles*: To visualize the SE peak distribution we have implemented in NaviSE our own Genome Viewer Tool (GVT). With this tool, two snapshots at *near* and *far* distances for each SE are portrayed, which are shown in the *SE table* window of the final report. NaviSE dynamically calculates the optimal range for each snapshot, based on the width of the SE. With *near* shot the user is able to determine the morphology of the SE, and with *far* the user is able to locate the SE in its genomics surroundings. In each snapshot both the location of the SE and the enhancer peaks determined by MACS are shown.
6. *HOMER motif finding*: SEs enclose high number of TFBSs [2]. Therefore, identifying such TFBSs is important for SE annotation. To find motifs of regulatory elements (mainly TFs) that are specifically enriched in the *loci* of SEs, relative to the *loci* of TEs, NaviSE uses the Hypergeometric Optimization of Motif Enrichment (HOMER). As a result, NaviSE generates in the final report a *HOMER table*, which includes motifs enriched in SEs, and a list of *de novo* motifs for which their respective binding elements are predicted by HOMER.
7. *Gene Ontology Enrichment Analysis (GOEA)*: To predict the functionality of the SEs, based on the closest gene of each SE determined by HOMER, NaviSE uses goatools [15].
8. *Pathways and protein-protein interaction annotation*: To obtain annotation of TFs and pathways related to SEs, NaviSE uses Enrichr [16]. To obtain protein-protein interaction (PPI) networks of SEs, NaviSE uses the database of PPIs String [17]. Results from Enrichr and String are processed and integrated into the final report to be navigated through NaviSE GUI for an easier interpretation for the user.
9. *NaviSE GUI*: To navigate throughout all the results, we have implemented an interactive chromosomal plot (Fig. 2) that represents the SE location in a karyotype; alongside with graphs that depict statistical values and properties related to SEs (shown in "Results" section), as well as information related to GOEA or Enrichr.

   Chromosomal plots are designed to include *hotspots* with links to the elements of the *SE table* from the final report, which are activated when the user navigates with the mouse over the gene names on the chromosomal plot. To enhance the usability of this feature, NaviSE generates three types of chromosomal plots:

   - Enrichment plot: it shows the *loci* location and the chromosome enrichment or depletion.
   - Rank plot: it shows *loci* coloured according to their rank in the *SE Table*. Several percentiles are represented based on the rank of the SE, and SEs falling within a percentile will be coloured correspondingly.
   - Closeness plot: it represents the proximity between SEs, according to which SEs will be coloured. This plot is highly useful to discern clusters of SEs that look overlapped. For the ordered list $\{SE_1, SE_2, \cdots, SE_{a-1}, SE_a, SE_{a+1}, \cdots, SE_{c-1}, SE_c\}$, of c SEs within a chromosome, for which each SE_k support is defined by its start ($SE_{k,start}$) and end ($SE_{k,end}$) *loci* positions, the closeness of a SE_k is its distance $C(SE_k)$ to the closest SE, determined by the following expression:

$$C(SE_k) = \begin{cases} SE_{2,start} - SE_{1,end} & \text{if } k = 1 \\ SE_{c,start} - SE_{c-1,end} & \text{if } k = c \\ \min(SE_{a+1,start} - SE_{a,end}, & \\ E_{a,start} - SE_{a-1,end}) & \text{otherwise} \end{cases} \quad (3)$$

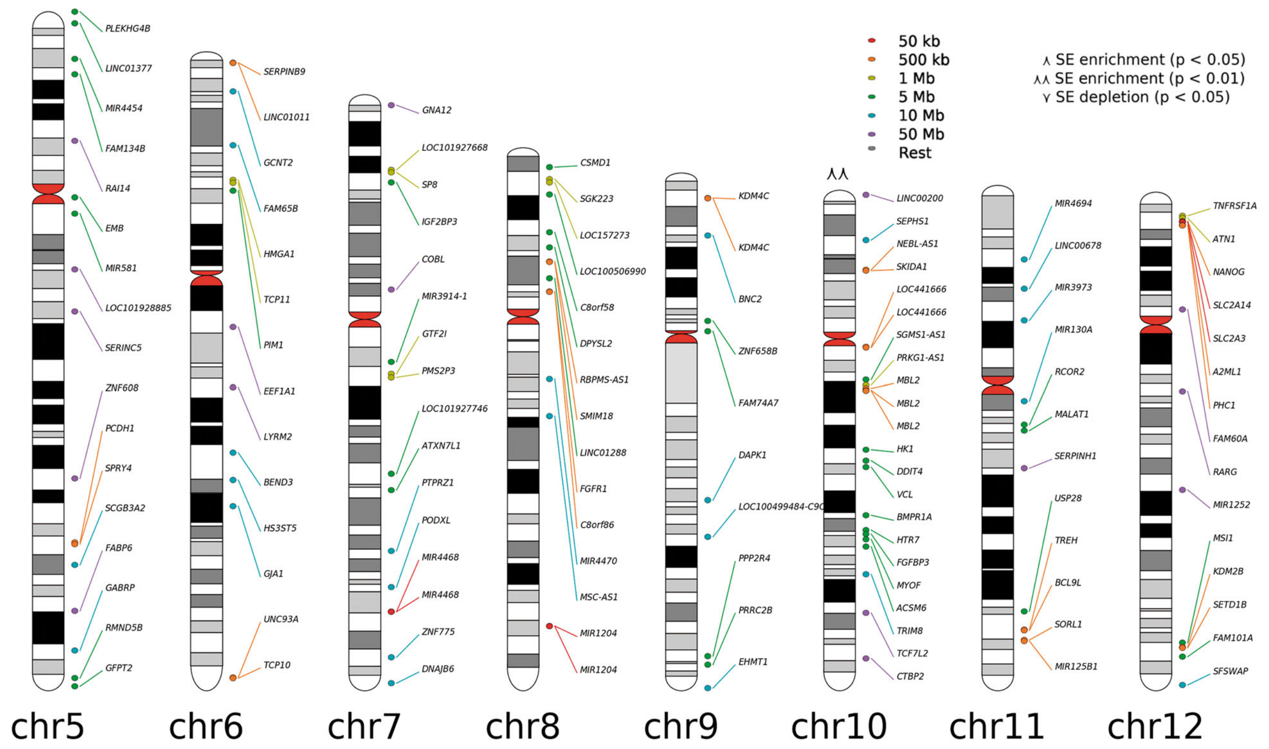


Fig. 2 Chromosomal plot. Partial snapshot of the chromosomal plot of SEs predicted with H3K27ac histone mark in ESCs. *Hot-spots* and *line colours* represent distances between two SEs (*red* to *violet* represents smaller to bigger distances), and ⅄/ ⅄⅄ and ϒ/ ϒϒ represent chromosome enrichment or depletion in genes with statistical significance p of 0.05 and 0.01, respectively

In all these chromosomal plots a probability p determining whether a chromosome is enriched (marked with ⅄ for $p < 0.05$ and ⅄⅄ for $p < 0.01$) or depleted (marked with ϒ for $p < 0.05$ and ϒϒ for $p < 0.01$) with SEs is calculated by a binomial approximation of the hypergeometric distribution ($h(k; K, n, N) \rightarrow b(k; K, z) \quad ; \quad z = \frac{n}{N}$) [18], where N is the number of genes in the whole genome, K is the number of SEs in all chromosomes, n is the number of genes in that chromosome, and k is the number of SEs in that chromosome.

10. *Gene Set Enrichment Analysis (GSEA)*: To obtain additional functional annotation of SEs, NaviSE performs the GSEA [19] from SE-associated genes, using gene sets from the Molecular Signatures Database (MSigDB).

Statistics of the comparison between TE and SE

Although both SEs and TEs derive from MACS peaks, they structurally differ for having higher peak density. To illustrate the differences between SEs and TEs, NaviSE shows in the final report a collection of metrics and plots depicting the differences between them. Among the most important plots are:

1. Ranking of SEs by the order of SE score: It is the plot of $Count_{STIT}$, given by Eq. 1 *vs* r, given by Eq. 2. It typically shows a *hockey stick* shape, with the inflection point marking the boundary between SEs and TEs, θ_{SE}.
2. INSs and OUTs: It shows statistics about the percentages of SEs and TEs that lay within a TSS or not. This might be interesting if a sample contains an elevated percentage of SEs within TSS, as some of these SEs might be misinterpreted as promoter signals.
3. SE *vs* TE length distribution: It shows the distribution of SE and TE length and pileup in a double histogram and a scatter plot. The histogram lying on the X-axis of the scatter corresponds to the length of SEs and TEs; and the histogram on the Y-axis corresponds to the pileup. This graph is complementary to the ranking of SEs by SE score, to shed light on the population of SEs and TEs.
4. SE *vs* TE subpeak length distribution: This graph contains the same elements than the SE *vs* TE length distribution graph, although showing the distribution of enhancers inferred by MACS.
5. Number of subpeaks (for number of bins $N = 10$ and $N = 20$): It shows the distribution of subpeaks each SE or TE has.

Generation of the NaviSE report

The final step of NaviSE is the generation of an HTML report, in which all the results from the analysis are gathered and presented within several windows, each of which contains interactive links both to external website which provide the user with further information about the SEs, as well as to other internal HTML pages created by NaviSE within the report. The content of this report is discussed in detail in the "Results" section.

Parallelisation implementation

The algorithm of parallelisation developed in NaviSE constitutes a significant improvement of performance in the analysis of NGS samples compared to non-parallelised pipelines. NaviSE determines the optimal number of processes, k, compatible with the computer resources as Luu et al. do in [8]. Such resources are the parallel processing capability of the computer measured as the number of cores, C, and the total main memory, M, in Gigabytes (GBs). NaviSE optimizes automatically, for each processing task i, the number of threads, k_i, in which the task i will be parallelised by the expression:

$$k_i = \min\left(C, C_u, \lfloor M/m_i \rfloor, l_i\right) \tag{4}$$

where C_u is the maximum number of cores reserved by the user to run NaviSE, m_i is the memory, measured in GBs, needed to run one process in task i, $\lfloor\ \rfloor$ is the floor operator and l_i is the cardinal of $D_i = \{d_1, d_2, \cdots, d_m\}$ which is the set of *chunks* of distributed data elements to be processed in task i. If $l_i > k_i$, the first k_i chunks are distributed to k_i threads. The distribution of information (number of chromosomes for stitching, SE peak distribution profiles for GVT, number of gene sets for GSEA) to be parallelised is based on a cyclic algorithm, implemented in Python, with the following outline: For the ordered set $S_i = \{s_1, s_2, \cdots, s_n\}$ of information elements, the set $P_i = \{1, \cdots, k_i\}$ of processes, and for the set D_i of data (chromosomes, positions on a list, gene sets) to be distributed across processors, we define D_{pi} as the *chunk* of data of the task i that is assigned to each processor p:

$$D_{pi} = \{d_j \mid \forall d \in D_i,\ p \in P_i,\ j \in \{1, \cdots, l_i\},\ \ j \bmod k_i = p\} \tag{5}$$

where mod is the module operator. Once the *chunk* D_{pi} is constructed, the subset of information elements $S_{D_{pi}} \subset S_i$ will be defined depending on the task i which is being parallelised. The list of parallelised tasks in NaviSE is $i = \{\mathrm{STIT}, \mathrm{GVT}, \mathrm{GSEA}, \mathrm{HOMER}\}$. In the case of parallelisation of SE prediction by stitching (STIT), the set of peak coordinates from MACS (S_{STIT}) is divided into k_{STIT} files, calculated with Eq. 4, with $m_{\mathrm{STIT}} = 3$ GBs. Here, $D_{\mathrm{STIT}} = \{1, 2, 3, \cdots, X, Y\}$ chromosomes, $D_{p,\mathrm{STIT}}$ represents the sets of chromosomes that will be processed in each $p \in P$ calculated by Eq. 5, $S_{D_{p,\mathrm{STIT}}}$ is the *chunk* of $s \in S_{\mathrm{STIT}}$ peaks which share the same chromosome from each set of chromosomes from $D_{p,\mathrm{STIT}}$. For a better understanding of the process, an example is developed in Fig. 3.

In the case of SE signal profile generation with GVT, $S_{\mathrm{GVT}} \equiv D_{\mathrm{GVT}}$, is the set of SE *loci*. Hence $D_{p,\mathrm{GVT}}$ contains all the *loci* that fulfill Eq. 5, based on k_{GVT} calculated with Eq. 4 with $m_{\mathrm{GVT}} = 2$ GBs.

In the case of GSEA parallelisation, D_{GSEA} is the set of combinations (GSEA signatures × GSEA cutoffs) and S_{GSEA} is the set genes associated to SEs and TEs up to the corresponding GSEA cutoff. Therefore, $D_{p,\mathrm{GSEA}}$ contains all the combinations that fulfil the Eq. 5, based on k_{GSEA} calculated with Eq. 4 with $m_{\mathrm{GSEA}} = 2$ GBs.

The parallelisation of all these tasks has been implemented with the *multiprocessing* module of Python. In the case of HOMER parallelisation, we took advantage of the capabilities already implemented in HOMER, with the number of processes k_{HOMER}, optimized by Eq. 4, with $m_{\mathrm{HOMER}} = 2$ GBs.

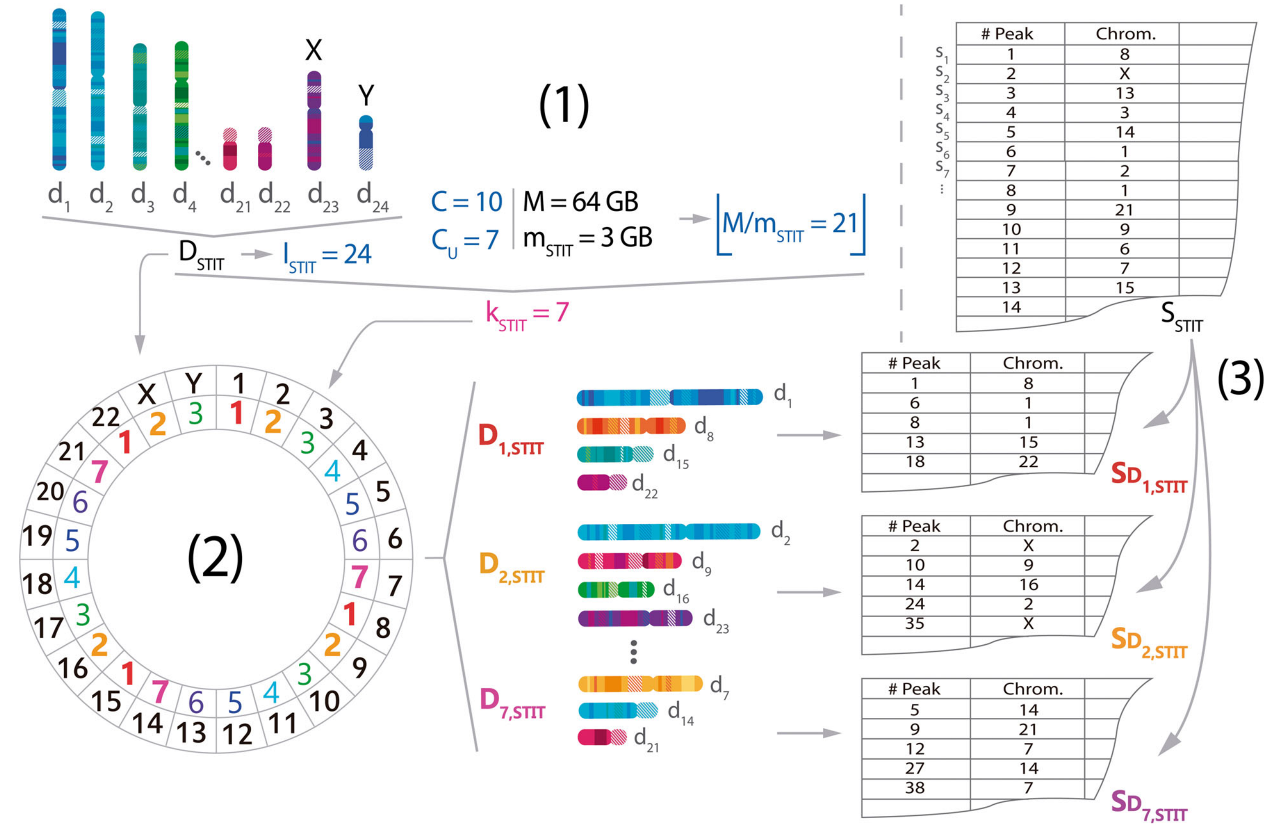
Results

To illustrate the performance of NaviSE, we have selected H3K27ac histone mark whose raw signal data has been downloaded from the GEO database [20] for three cell types: human Embryonic Stem Cells (ESC) (GSM663427, with control GSM605335), monocytes (MON) (GSM1003559 with control GSM1003475) and neurons (NEU) (GSM2072642, with control GSM2072639). For other analysis, we also used H3K4me1 (GSM409307) and H3K4me3 (GSM409308) from ESCs.

HTML report generation

The output of NaviSE for each experiment is a collection of HTML linked pages whose main page contains dynamic graphical elements, namely, a blue horizontal ribbon with links to all the HTML pages from the report, detailed below; a grey sidebar by which the user can access the different subsections; and a window in which the results are displayed.

The *main window* contains basic information about the analysis and different chromosomal plots, defined in the point 9 of "SE prediction and annotation" section, represented in the chromosomal plot snapshot of Fig. 2. The chromosomal plot includes links to the SEs in *SE Table* window of the final report, which includes general information about each SE (genomic location, number of subpeaks, SE score), alongside with a snapshot of the SE genomic signal profile, included for visual evaluation of the SE quality, together with the quantitative SE score. The *SE Table* columns referring to gene names and genomic location include, respectively, a link to GeneCards site [21] and UCSC Genome Browser [22], as shown in Fig. 4.


(1)
X
Y
d_1 d_2 d_3 d_4 d_{21} d_{22} d_{23} d_{24}
D_{STIT} → $l_{STIT} = 24$
$C = 10$ | $M = 64$ GB
$C_U = 7$ | $m_{STIT} = 3$ GB
$\lfloor M/m_{STIT} = 21 \rfloor$
$k_{STIT} = 7$
(2)
(3)
$D_{1,STIT}$
d_1 d_8 d_{15} d_{22}
$D_{2,STIT}$
d_2 d_9 d_{16} d_{23}
$D_{7,STIT}$
d_7 d_{14} d_{21}
s_1 s_2 s_3 s_4 s_5 s_6 s_7
S_{STIT}
$SD_{1,STIT}$
$SD_{2,STIT}$
$SD_{7,STIT}$

# Peak	Chrom.
1	8
2	X
3	13
4	3
5	14
6	1
7	2
8	1
9	21
10	9
11	6
12	7
13	15
14	

# Peak	Chrom.
1	8
6	1
8	1
13	15
18	22

# Peak	Chrom.
2	X
10	9
14	16
24	2
35	X

# Peak	Chrom.
5	14
9	21
12	7
27	14
38	7

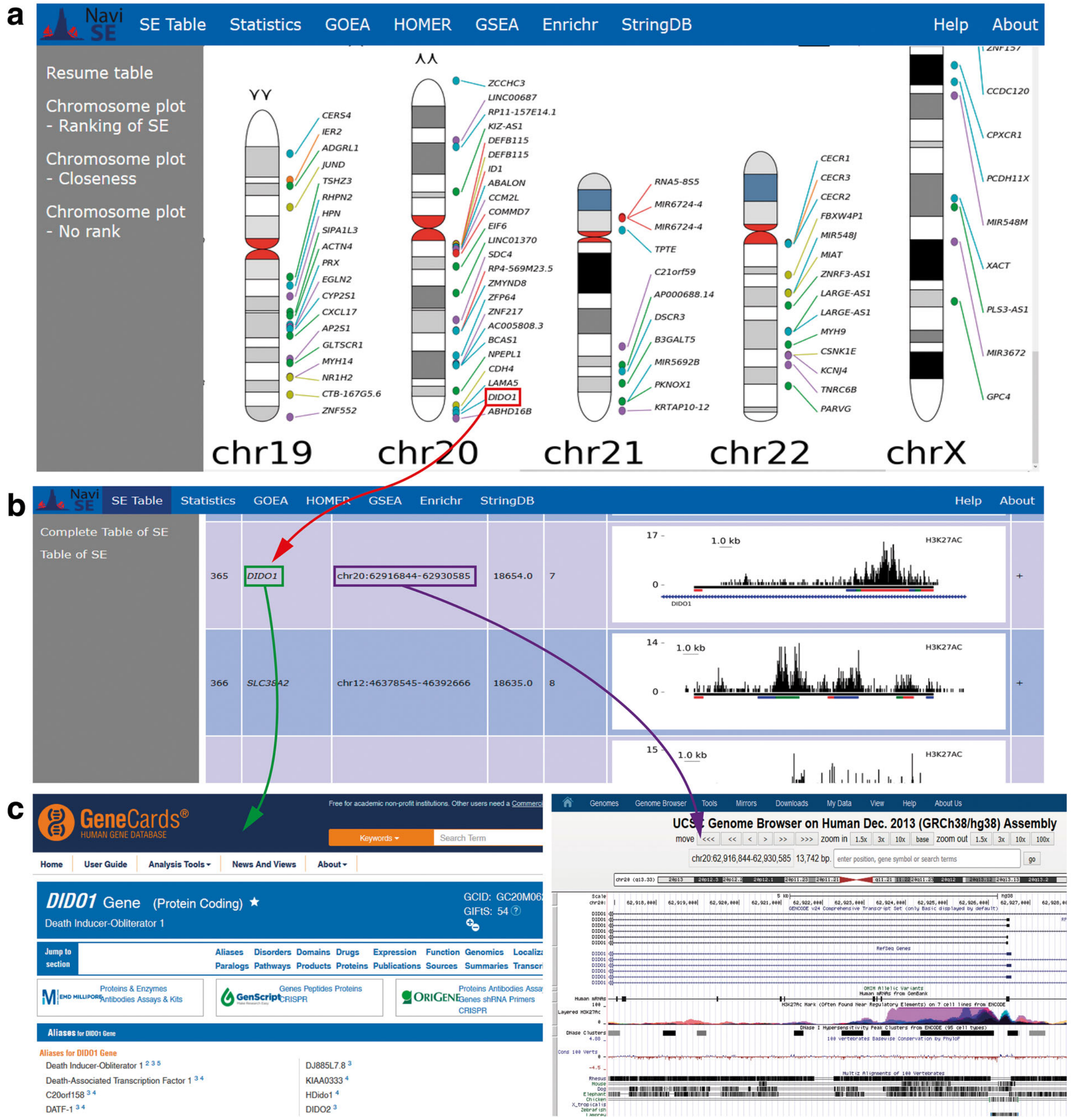
Fig. 3 Scheme of parallelization of stitching. **1** Determination of the number of processes (k_{STIT}) based on Eq. 4 for a case in which the number of available cores (*C*) is 10, the maximum number of cores reserved by the user (C_u) is 7, the memory main of the computer (*M*) is 64 GB, the memory allocated to stitching (m_{STIT}) is 3 GB and the cardinal (l_{STIT}) of the set of chromosomes ($D_{STIT} = \{d_1 = 1, d_2 = 2, \cdots, d_{22} = 22, d_{23} = X, d_{24} = Y\}$) is 24. The resulting number of allocated cores calculated by Eq. 4 is $k_{STIT} = C_u = 7$. **2** Construction of data *chunks* is calculated by Eq. 5. Since $k_{STIT} = 7$, the set of chromosomes, D_{STIT}, is divided into 7 subsets or *chunks*: $D_{1,STIT} = \{d_1, d_8, d_{15}, d_{22}\}$; $D_{2,STIT} = \{d_2, d_9, d_{16}, d_{23}\}$; $\cdots$; $D_{6,STIT} = \{d_6, d_{13}, d_{20}\}$ and $D_{7,STIT} = \{d_7, d_{14}, d_{21}\}$. **3** Assignment of information elements. In the case of stitching, assigned elements are MACS peaks (inferred as enhancers). After the assignment of the subsets $D_{1,STIT}$, $D_{2,STIT}$, etc., the set of MACS peaks, $S_{STIT} = \{s_1, s_2, \cdots\}$ is divided into 7 subsets of elements, $S_{D_{1,STIT}} = \{s_1, s_6, s_8, \cdots\}$, $S_{D_{2,STIT}} = \{s_2, s_{10}, s_{14}, \cdots\}$, $\cdots$, $S_{D_{7,STIT}} = \{s_5, s_9, s_{12}, \cdots\}$, based on the chromosome of each row. Finally, all the subsets of elements are simultaneously processed by NaviSE, combined into one file, and the SE ranks are calculated

Statistics window implements a series of graphs which allow the user to obtain information related to the SEs in the sample. Some of those graphs are analysed thoroughly in the corresponding *Analysis of different cell lineages* "Results" section.

GOEA window includes the results from the GOEA. At first, a barplot shows the significant terms from GO categories (biological process, cellular component and molecular function) which, upon clicking, will lead to a Directed Acyclic Graph (DAG) of the GO terms associated with the significant term, each of which contains the related genes associated to that term. Below the barplot, there is a table that leads to the DAG for the corresponding GO term, which includes values such as enrichment ratio of the predicted cell population, and the False Discovery Rate (FDR) for each term.

Similarly, the *GSEA* window (Fig. 5) contains several graphs depicting the GSEA profile of the significance of the analysis, for each signature (group of gene sets) and threshold. Clicking on a graph leads to its corresponding information element on a table below, which contains several related values, such as the significant GSEA term, related SE genes, and statistical values linked to the GSEA term such as Enrichment Score (ES), Normalized Enrichment Score (NES), FDR and *p*-values provided by GSEA, which are further described in Additional file 1.

HOMER window shows the results from the motif analysis by the HOMER tool, which includes two ranked tables, one for known motifs and another one for *de novo* motifs. 'Known motifs' table contains a LOGO image for each motif and the name of the TF or binding protein using such binding motif. It also includes the percentage of SE and TE sequences that has such motif, and a *p*-value



365	DIDO1	chr20:62916844-62930585	18654.0	7
366	SLC38A2	chr12:46378545-46392666	18635.0	8

Fig. 4 NaviSE GUI. All NaviSE windows contain a navigation bar on the top with links to all the results windows. On the *left side* there is a side *menu* bar with links to subsections of the active window. **a** The *main window* of NaviSE depicting the chromosomal plot in which the positions of all predicted SEs are mapped into a karyotype. Each SE in this window is a *hot-spot* with a link to the SE table. **b** Amongst other features, *SE table* contains the ranking of SEs, the names in the SE table linked to GeneCards (**c**), the chromosomal locations linked to UCSC Genome Browser (**d**), the SE score, the number of subpeaks, and, in the last column, the SE signal profile drawn with our GVT module

that measures the statistical significance of the association of the SE with such motif. The '*de novo*' table includes motifs predicted by HOMER to bind elements differentially in SEs and TEs. Upon clicking on each element in the '*de novo*' table, NaviSE redirects to a HOMER-generated page that includes more information about the motif.

Finally, *StringDB* and *Enrichr* windows show, respectively, PPI networks from SEs at different confidence values; and results from Enrichr website including TFs related to SEs, cell or tissue specification or metabolic pathways linked to the SE population. Each subsection includes a barplot of the significant terms which link to

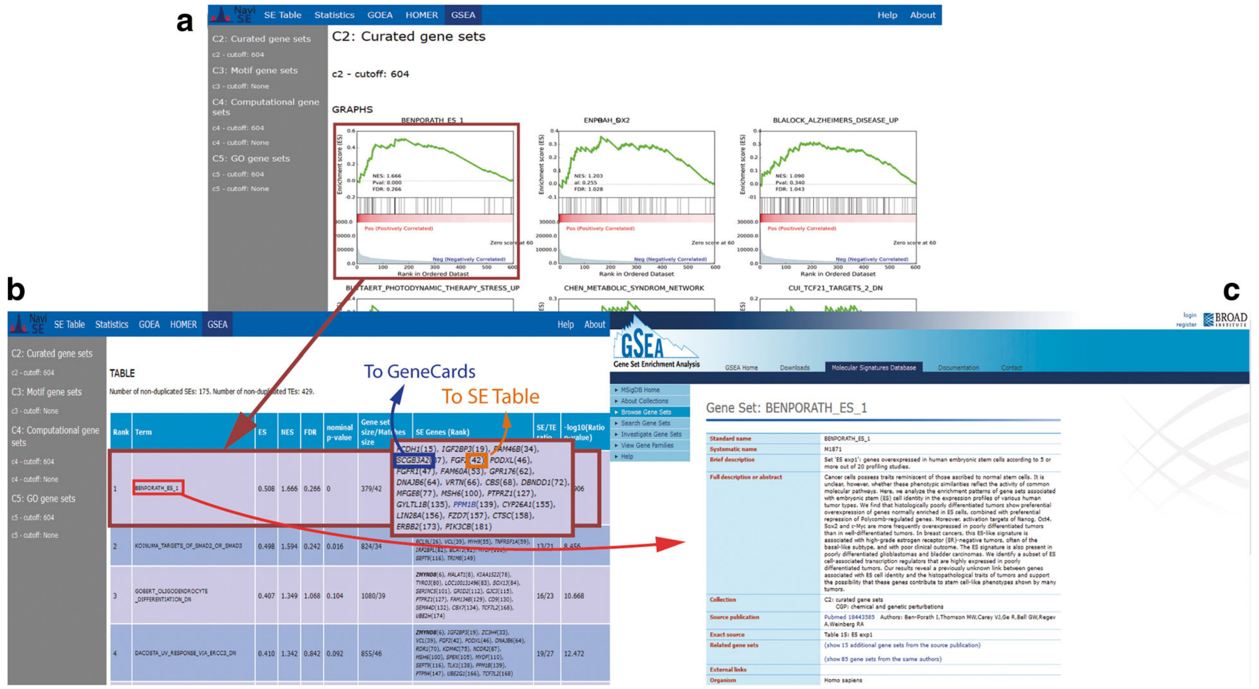


Fig. 5 GSEA NaviSE GUI. **a** Window with interactive links to the corresponding GSEA terms in the *GSEA table*. **b** Table with the ranking of GSEA terms; each GSEA term, linked to GSEA website (**c**); statistical values such as Enrichment Score (ES), Normalized Enrichment Score (NES), FDR and *p*-values provided by GSEA; list of associated genes, linked to GeneCards with SE ranking value (in parentheses) linked to *SE table* window, described in Fig. 4b

the elements in a specific table. This window is described in detail in Additional file 1.

Process parallelisation

The parallelisation of NaviSE is fundamental to save time during the data processing, more so when the analysis is performed simultaneously with numerous cell types or marks. The computing time optimization achieved upon NaviSE parallelisation is shown in Fig. 6a.

Most time-consuming processes show a considerable decrease in running time: in SE prediction up to a 30% of the original time, in gene annotation up to 10%, and in GVT up to 8.5%. TF prediction by HOMER, and GSEA, are also parallelised, although, interestingly, their optimal processing time achieved is obtained using between 4 and 7 processors, probably due to limits in main memory usage or difficulties of the Python-operating system interface for managing the optimal access to all the CPUs. In short, the overall amount of time is reduced up to a 40% between 1 and 19 processors, and the optimal difference is achieved at 15 processors, with a reduction up to 30%.

Hence, NaviSE shows a considerable reduction of processing time even with small processing capability, below 6 CPUs, which may allow conducting research with mid-range computers. As for the non-parallelisable processes, they involve very fast computing tasks that do not require parallelisation or in which the algorithm shown in Eq. 5 cannot be efficiently implemented (such as reading files, alignment of reads or processing of some tables).

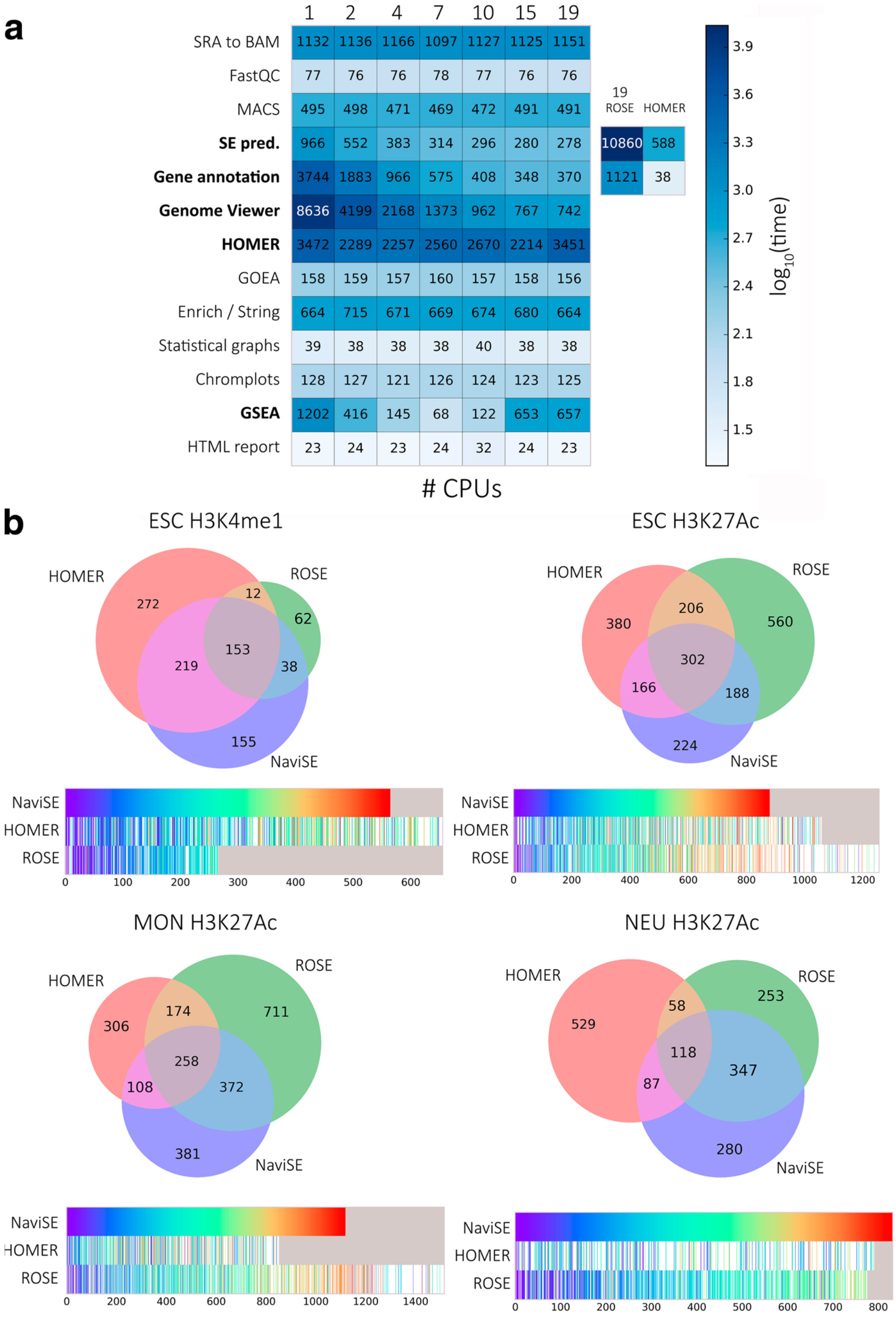
Comparison of SE predictions among different software

As previously mentioned, NaviSE performs the whole processing from raw data files to comprehensive annotations of SEs. However, there are alternative software packages that can perform the SE prediction task as well. Here, we compare NaviSE SE predictions with those obtained with ROSE and HOMER, in ESCs epigenomics data.

The stitching calculation that we have implemented in NaviSE is much faster than the ones implemented in ROSE and in HOMER. For example, when running in 19 CPUs, the stitching of NaviSE takes 278 seconds, whereas the stitching of ROSE takes 10,860 s (39 times slower than NaviSE) and the stitching of HOMER 588 seconds (2.1 times slower than NaviSE) (Fig. 6a).

It has to be taken into account that NaviSE does not only annotate the SE peaks but also the subpeaks. This feature provides NaviSE with an important annotation feature to understand the SE structure that HOMER cannot provide.

To test the similarity between the SE predictions produced by the different software, we have used ESC, MON and NEU cell types. Although NaviSE, ROSE and HOMER predict different number of SEs, they share a significant number of predictions (see Euler-Venn diagrams in Fig. 6b). To analyse deeply the similarities among these



	1	2	4	7	10	15	19	ROSE (19)	HOMER (19)
SRA to BAM	1132	1136	1166	1097	1127	1125	1151		
FastQC	77	76	76	78	77	76	76		
MACS	495	498	471	469	472	491	491		
SE pred.	966	552	383	314	296	280	278	10860	588
Gene annotation	3744	1883	966	575	408	348	370	1121	38
Genome Viewer	8636	4199	2168	1373	962	767	742		
HOMER	3472	2289	2257	2560	2670	2214	3451		
GOEA	158	159	157	160	157	158	156		
Enrich / String	664	715	671	669	674	680	664		
Statistical graphs	39	38	38	38	40	38	38		
Chromplots	128	127	121	126	124	123	125		
GSEA	1202	416	145	68	122	653	657		
HTML report	23	24	23	24	32	24	23		

Fig. 6 NaviSE peformance comparisons. **a** NaviSE CPU running time for different numbers of CPUs. Heatmap of the processing time for each NaviSE process for different numbers of CPUs, writen on top. Tasks parallelised by NaviSE are highlighted in *bold* typeface. For the SE prediction and gene annotation, running times of ROSE and HOMER on 19 CPUs are also provided. **b** SE prediction similarities among different software. For each cell type and histone mark, the Euler-Venn diagram with the number of commonly predicted SEs is represented on top and the comparisons among the SE ranking generated by the different software at the bottom. The rank of each SE predicted by NaviSE is colour-codified (*bluer* colours indicate higher positions in the rank and redder colours lower positions). Each NaviSE SE is mapped onto HOMER and ROSE SE ranking tracks in the position predicted by HOMER and ROSE for such SE, with the colour codification corresponding to the ranking predicted by NaviSE. SEs predicted by other software that are not predicted by NaviSE appear in white. *Grey* boxes mark the indexes for which a rank in a predictor has exhausted its number of predicted SEs in comparison to the maximum rank predicted by the three software {HOMER, ROSE, NaviSE}

predictions we have designed a graphical representation that allows us to track the ranking of each SE predicted by each software in comparison with the ranking predicted by NaviSE. This representation shows that the rank of the score of the SEs is very similar among all of the predictors (ranking bars in Fig. 6b). A detailed explanation of prediction divergences between different software, as well as between epigenomic combinations, is provided with an example with ESCs at "NaviSE epigenomics signal algebra is able to predict SEs with sharper signals" "Results" section.

SE prediction of different cell lineages

To assess the capabilities and performance of NaviSE, we have run several real datasets from different species (human and mouse), histone marks (H3K27ac, H3K4me3 and H3K4me1), and cell types (ESC, MON and NEU), using the hg38 human genome version.

Main page, SE table, and Statistics

Using the same default parameters with H3K27ac histone mark, the NaviSE analysis for the different cell lines yielded a wide range of SEs (n_{ESC} : 664, n_{NEU} : 1073, n_{MON} : 1235). The signals of the most important SEs are shown in the Fig. 7 and the main statistics for each cell type are depicted in the Fig. 8.

The distribution of subpeaks varies considerably between SEs and TEs. TE subpeak distribution follows a Zipfian-like distribution in all the analysed cell lines, that is, most of the samples contain only 1 subpeak, and the number of samples that contain higher amount of subpeaks goes down at a rate of ~50% of the previous number of subpeaks; whereas the SE distribution might follow a χ^2 distribution or a normal distribution. In the case of ESCs, the maximum of subpeaks is between 5 and 7, whereas in NEU and MON the distribution is uniform between 6 and 14 subpeaks, with a considerable amount of SEs having more than 20 subpeaks.

The differences in length distribution between TEs and SEs are apparent in all samples. Interestingly, TEs usually show a bi or trimodal distribution with maxima at ~100, ~1000 or ~10,000 nt in all the analysed cell types, whereas SEs show a monomodal normal-like distribution with means around 25,000 - 50,000 nt. On the other hand, subpeak distribution shows no significant differences between SEs and TEs, both in length and pileup.

HOMER analysis

The results of the most relevant TFs revealed by HOMER are shown in Table 1. Although all three cell lines showed shared TFs such as *TCF3*, each cell type contained a set of cell-specific TFs. For instance, ESC contained *NKX2-2* and *NKX2-5*, involved in heart and nervous development; NEU contained *RXR*, involved in neural development, *NR5A2*, involved in embryonic development and, interestingly, *RUNX1*, thought to be involved in hematopoiesis. Finally, MON contained *GATA2* and *GATA1*, the first closely related to hematopoiesis and the second involved in the switch of fetal hemoglobin to adult hemoglobin.

GOEA and GSEA results

GOEA and GSEA are represented in Figs. 9 and 10 respectively. Both results are related, as the signatures used for GSEA contain sets of genes related to GO sets. Both analysis show correlation of functions to each cell type.

For ESC, the most relevant GO terms are related to protein expression (*positive regulation of transcription*), rearrangement of cellular morphology (*focal adhesion, lamellipodium*) or pluripotency (*somatic stem cell population maintenance*). As for GSEA, significant terms are related to master TFs of ESCs, such as *NANOG*,

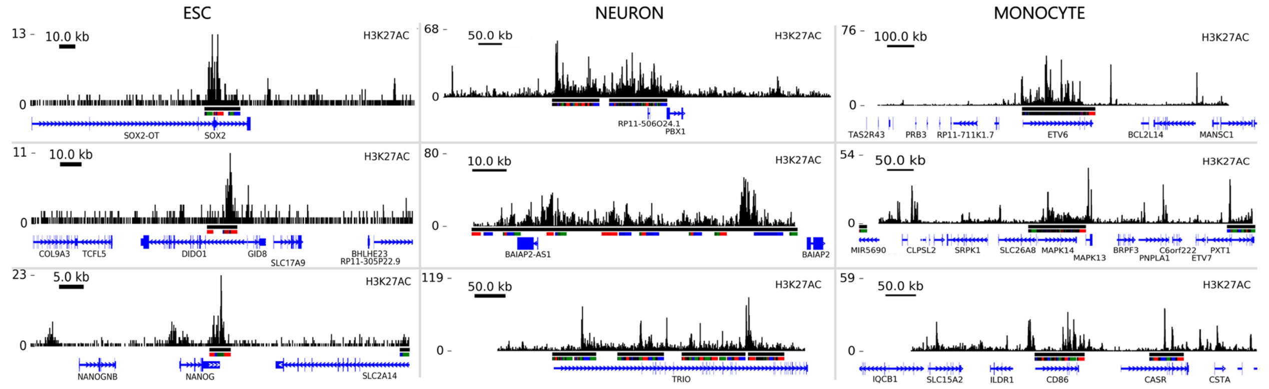


Fig. 7 SE ChIP-seq peak distribution. Box plot of peak distribution for each SE obtained with our GVT module, for some representative SE predictions. Each column represents SEs from each cell type. *Black lines* below the signal represent the SE supports at each sample, and bars in alternating colours below SE bar show the supports of the SE subpeak composition

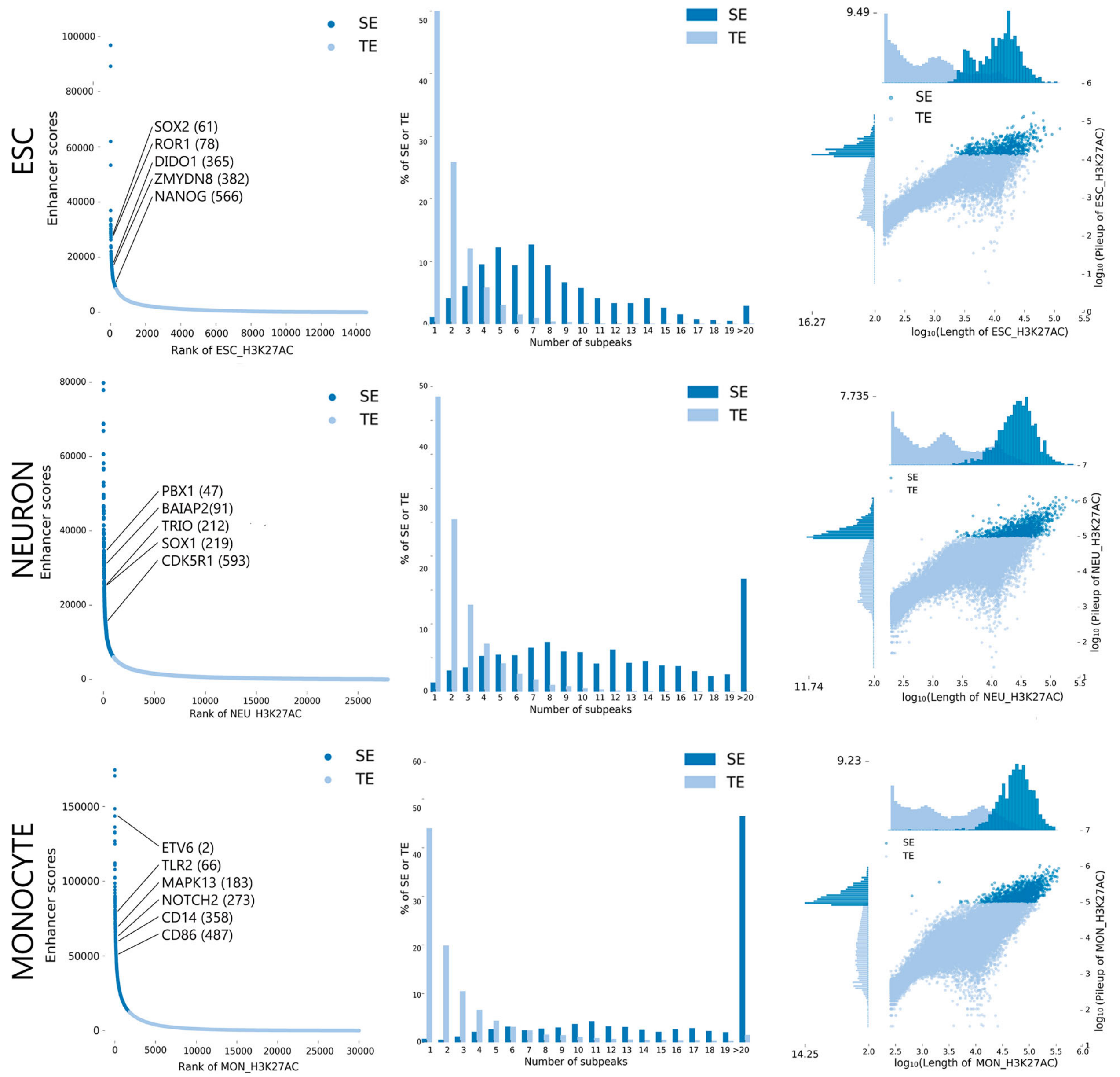
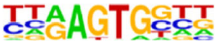
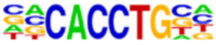
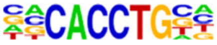
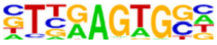
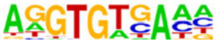
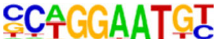
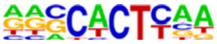
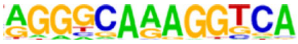
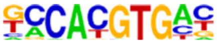
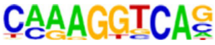
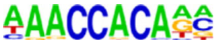
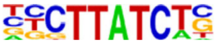
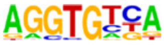
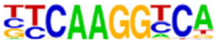
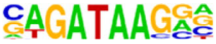


Fig. 8 NaviSE GUI statistics. Diagrams of statistical parameters for each cell type, arranged in rows. SE ranking (Eq. 2) by ChIP-seq signal (*left column*) with the most relevant SEs of each cell type and their corresponding ranking, SE distribution of the number of subpeaks of the SEs (*center column*) and SE pileup *vs* length scatter plots in $\log_{10}$ scale with the respective distributions of SE pileups in ordinates and SE lengths in abscissas (*right column*)

or cytoskeletal reorganization. Among the predominant genes, most of them repeated in several functional terms, we remark *ROR1* (which modulates neurite growth and is highly expressed during early embryonic development [23]), *ZIC3/5* (involved in the formation of right/left axis during development, and direct activator of *NANOG* promoter in ESC [24]) or *SOX2* (one of the Yamanaka's reprograming TFs, used for the induction of pluripotency, as well as a core pluripotency factor in ESC [25]).

Regarding NEU, the most relevant GO terms are related to neural development (*ephrin signaling*, *Wnt signaling pathway*, *dendritic spine*, *axon guidance*). As for GSEA, three relevant terms are *generation of neurons*, *neuron differentiation*, and *neurite development*. Three highly ranked genes in these GSEAs are *CDK5R1* (neuron-specific activator of cyclin-dependent kinase 5, required for proper development of the central nervous system, also found essential for oligodendrocyte maturation and myelination [26]), *BAIAP2* (brain-specific angiogenesis inhibitor binding protein, might be related to neural growth-cone guidance, dendritic spine development and NMDA receptor regulation [27]) and *PBX1* (regulates

Table 1 The most relevant TFs, and their binding motifs for all cell types obtained from HOMER analysis. *p*-values are presented in their *integer* logarithmic form (p*P-val* $\equiv -\log_{10}$ *P-val*)

ESC			NEU			MON		
TF	Motif	p*P-val*	TF	Motif	p*P-val*	TF	Motif	p*P-val*
NKX3-2		121	*TCF3*		142	*TCF3*		142
NKX2-2		83	*TBX21*		132	*TEAD2*		105
NKX2-5		77	*RXR*		124	*NPAS2*		90
ESRRA		75	*RUNX1*		124	*GATA2*		85
TBX5		75	*NR5A2*		122	*GATA1*		83

differentiation and survival of certain neurons, and is impaired in Parkinson's disease [28, 29]).

Regarding MON, the most relevant GO terms are related to specific functions of monocytes involved in immune response (*phagocytosis, T cell receptor signaling pathway, MyD88-dependent toll-like receptor signaling pathway, lipopolysaccharide-mediated signaling pathway*). As for GSEA, three relevant terms are *T cell receptor signaling pathway, reactome immune system* and *immune system process*. Genes shared by several GO

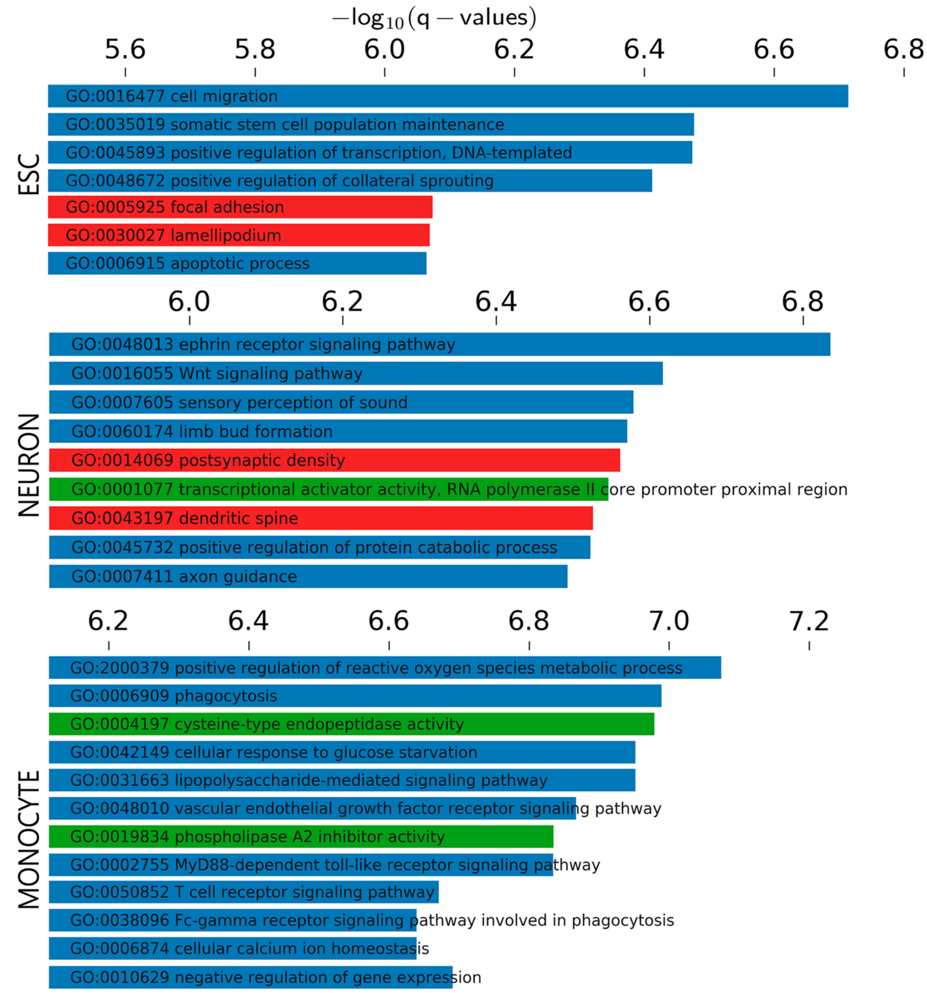


Fig. 9 NaviSE GUI GOEA significant terms. *Bar* plots for each cell type depicting the most relevant and statistically significant terms for GOEA of the genes associated with SE predicted for H3K27ac. *Red* - cellular component, *blue* - biological process, *green* - molecular function

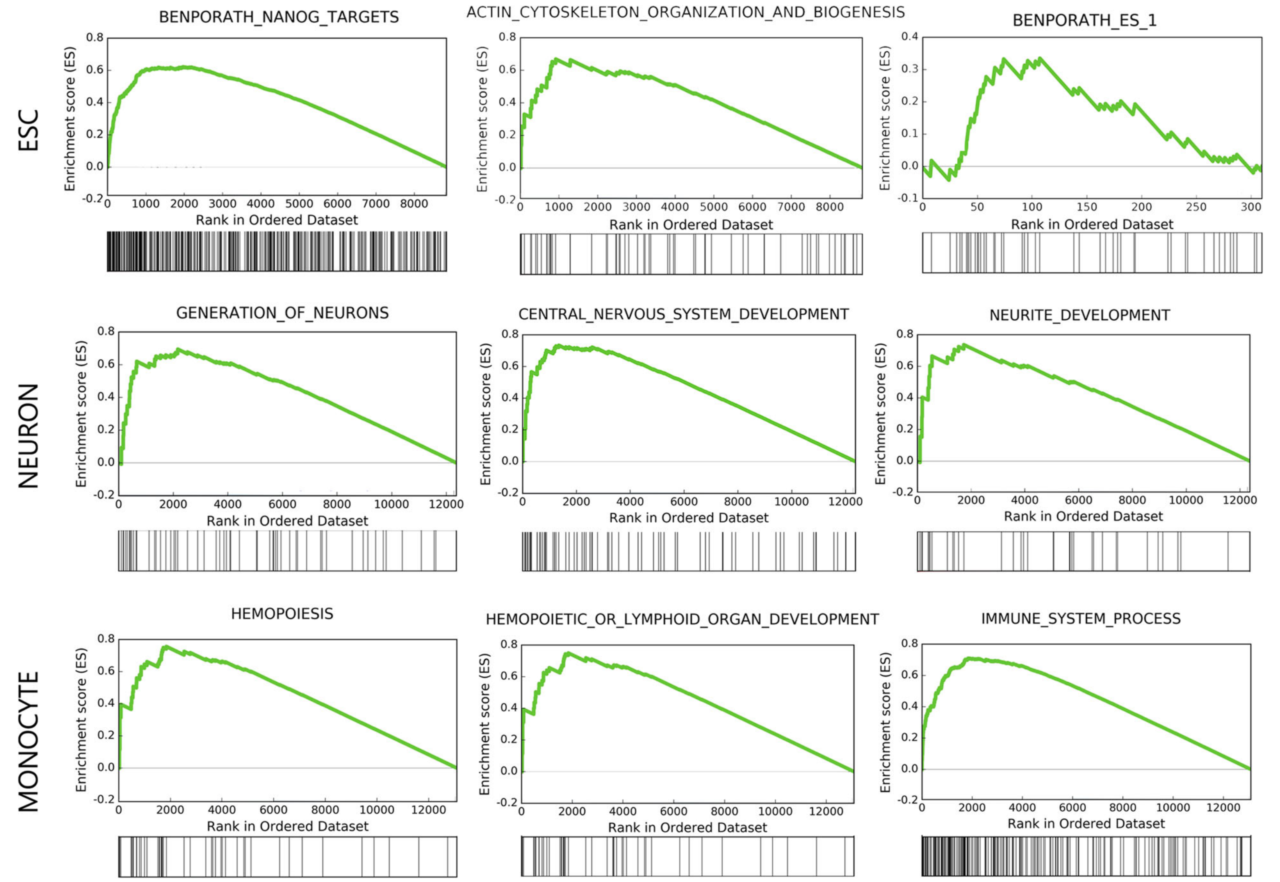


Fig. 10 NaviSE GUI GSEA most significant terms. GSEA profiles depicting three significant GSEA sets, from MSigDB, for each cell type for genes associated with SE predicted for H3K27ac. Each graph contains the typical GSEA profile alongside its positive matches in the bar below

terms are *NOTCH2* (related to hematopoiesis), *CD14* (one of the main markers of monocytes), *TLR2* (Toll-like receptor 2, which plays a fundamental role in pathogen recognition and activation of innate immunity [30]), *MAPK13* (is activated by proinflammatory cytokines and cellular stress [31]) or *LYN* (might be involved in the regulation of mast cell degranulation, and erythroid differentiation [32, 33]). Interestingly, *NOTCH1* gene, which is essential for hematopoiesis [34], does not appear in the list of SEs predicted by NaviSE for this dataset.

Enrichr analysis

We performed an Enrichr analysis in order to search genes involved in cellular processes related to each cell type. Most of the found genes, if not mentioned previously, appeared in GSEA and GOEA as well.

For ESC, the Enrichr Reactome presents several terms such as *transcriptional regulation of pluripotent stem cells*; and *POU5F1, SOX2, NANOG genes related to proliferation*, widely related to embryogenesis. Predominant genes are *FGF2* (implicated in a multitude of physiologic and pathologic processes, including limb development, angiogenesis, wound healing, and tumour growth [35]), *SOX2* or *NANOG* (TF belonging to Homeobox proteins, critically involved with self-renewal of undifferentiated ESCs, which is also one of Thomson's reprogramming factors [36]). ENCODE and Chromatin Enrichment Analysis (ChEA) TFs includes TFs related to pluripotency (*TCF3*, *NANOG*, *SOX2*, *POU5F1* and *KLF4* as the most relevant) which share several genes, such as *ZMYDN8*, or *DIDO1* (involved in apoptosis, autophagy, and meiosis). Interestingly, and as described by Hnisz et al. [4], we found that the SEs predicted by NaviSE are capable of disclosing a crosstalk between TFs (for instance, all the aforementioned TFs interact with *SOX2* and *NANOG*, according to ENCODE).

As for NEU, Reactome includes significant terms such as *axon guidance* or *semaphorin interactions*, with genes such as *TRIO* or *CDK5R1*; which also appear as genes associated with several TFs such as *REST* (transcriptional repressor that represses neuron-specific genes, such as type II sodium channel gene [37, 38]), determined by ENCODE or TRANSFAC. A gene predicted to associate with *REST* is *SOX1*, a known neuronal marker.

Regarding MON, Reactome presents several terms such as *immune system*, *innate immune system*, *hemostasis* or *toll-like receptor 2 cascade*, widely related to monocytes, whose associated genes are *TLR2*, *FOS* (implicated as regulator of cell proliferation, differentiation, and transformation, associated with B lymphocyte differentiation and involved in lypopolisaccharide and low density lipoprotein response [39–41]) or *CD86*, expressed by antigen-presenting cells. Binding of this protein to CD28 antigen is a co-stimulatory signal for activation of the T-cell. TRANSFAC and ENCODE include genes associated with TFs like *GATA1*, *GATA2*, *SPI1* or *RUNX1*, among which are *IKZF1* or *JARID2*. Enrichr also determined markers for monocytes or lymphoid cells, such as *RIN3*, *CXCR4*, *TREM1* or *ETV6*.

NaviSE epigenomics signal algebra is able to predict SEs with sharper signals

To evaluate to which extent the use of the epigenomics algebra improves the SE predictions, we have selected combinations of activation and repression epigenetic signals and compared SE predictions of HOMER, ROSE and NaviSE in ESCs. We denote the set formed by a SE software predictor {HOMER, ROSE, NaviSE}, and the set of SEs and TEs derived from an algebra of single or combined epigenetic signals {H3K27ac, H3K4me1, H3K4me3, H3K27ac NOT H3K4me3, H3K27ac NOT H3K27me3, H3K27ac + H3K4me1 - H3K4me3, H3K27ac + H3K4me1 - H3K27me3} as $STIT_{pred-algebra}$. To quantify the results of the different $STIT_{pred-algebra}$, we collected a set of ESC core pluripotency markers from the literature [42] and built a metric of the global goodness of the $STIT_{pred-algebra}$ based on the SE ranking generated for each $STIT_{pred-algebra}$ over the set of ESC markers. As each $STIT_{pred-algebra}$ contains a different number of SEs (thus, producing ranks of different length), to make the different SE ranks comparable, we designed a transformation to rescale each SE rank, r, given by Eq. 2 into a scaled rank $s(r)$ as follows:

$$s(r) = \frac{r}{\left|STIT_{pred-algebra}\right|} \cdot 100 \tag{6}$$

where $\left|STIT_{pred-algebra}\right|$ is the number of SEs predicted by each $STIT_{pred-algebra}$. Thus, when we apply Eq. 6 to scale the rank r, it produces a $s(r)$ in the range $[0, 100]$ if the epigenomics signal algebra is predicted as a SE, and $s(r) > 100$ if the signal algebra is predicted as a TE or is not predicted at all. Better 1 $STIT_{pred-algebra}$ assigns lower $s(r)$s to the SEs associated to ESC gene markers.

To quantify the global performance of each $STIT_{pred-algebra}$, we calculated the average $\bar{s}$ of $s(r)$ over the list of all ESC markers. Therefore, the best $STIT_{pred-algebra}$ will produce the lowest $\bar{s}$. We depict the $s(r)$ for the list of ESC gene markers and the list of $STIT_{pred-algebra}$ in the heatmap of Fig. 11a.

We observe three main patterns of behaviour, a group I of genes (from *MED14* until *MYH9*) that has associated a majority of SEs predicted by almost all the $STIT_{pred-algebra}$, some of them not by HOMER, a group II (from *TPD52* until *LRRC2*) that has associated TEs predicted by ROSE and NaviSE $STIT_{pred-algebra}$ but not by HOMER, and a group III (from *RBM14* until *KLF2*) that has associated lower ranked TEs from some of the combined algebras of NaviSE.

Interestingly, no $STIT_{pred-algebra}$ predicts SEs associated with the master regulator of pluripotency *POU5F1/OCT4* (they appear as TEs with H3K4me3 and H3K27ac + H3K4me1 - H3K27me3 from NaviSE), suggesting that *POU5F1* has a subtle epigenomic regulation that hinders the discovery for upstream *POU5F1* regulators, as it has been observed in the computational attempts with unconstrained discovery algorithms to find *ab initio* motifs regulating the *POU5F1* promoter [43].

The plot in Fig. 11b depicts the normalised metric of global performance $\bar{s}$ of each $STIT_{pred-algebra}$, where the lowest values are associated to the best performance. We observe that HOMER-based predictors show the worse performance, NaviSE single epigenomic signal SE predictions are better than those of HOMER and ROSE, and NaviSE H3K27ac + H3K4me1 - H3K4me3 algebra is better than any other single epigenomic signal SE predictions, thus showing the advantage of using the NaviSE epigenomic signal algebra to perform SE predictions.

To illustrate how the profiles of the combined epigenetic signal algebras are developed, we selected the best performing algebra (H3K27ac+H3K4me1-H3K4me3) and depicted its resulting combination and component signals profiles H3K27ac, H3K27me1, H3K4me3 for *NANOG*, (Fig. 11c) and *FOXO1* (Fig. 11d). In both cases, the deletion of the H3K4me3 promoter signal upstream and over the first exon and intron shortens the SE support to focus the SE support upstream of these genes.

Therefore, although there might not be a 'gold standard' on what a real SE is, we can conclude that the SE predictions of NaviSE are better than other predictors', with the added advantage to be fast obtained, fully automatized and comprehensively annotated.

Conclusions

We designed NaviSE to perform automatic parallelised SE prediction from genome-wide epigenetic signals, or an algebra of them, due to an optimization that reduces the necessity of inputting most of the parameters, providing a comprehensive annotation of SEs. NaviSE SE annotation runs from the motifs of TFBSs enriched in SEs through

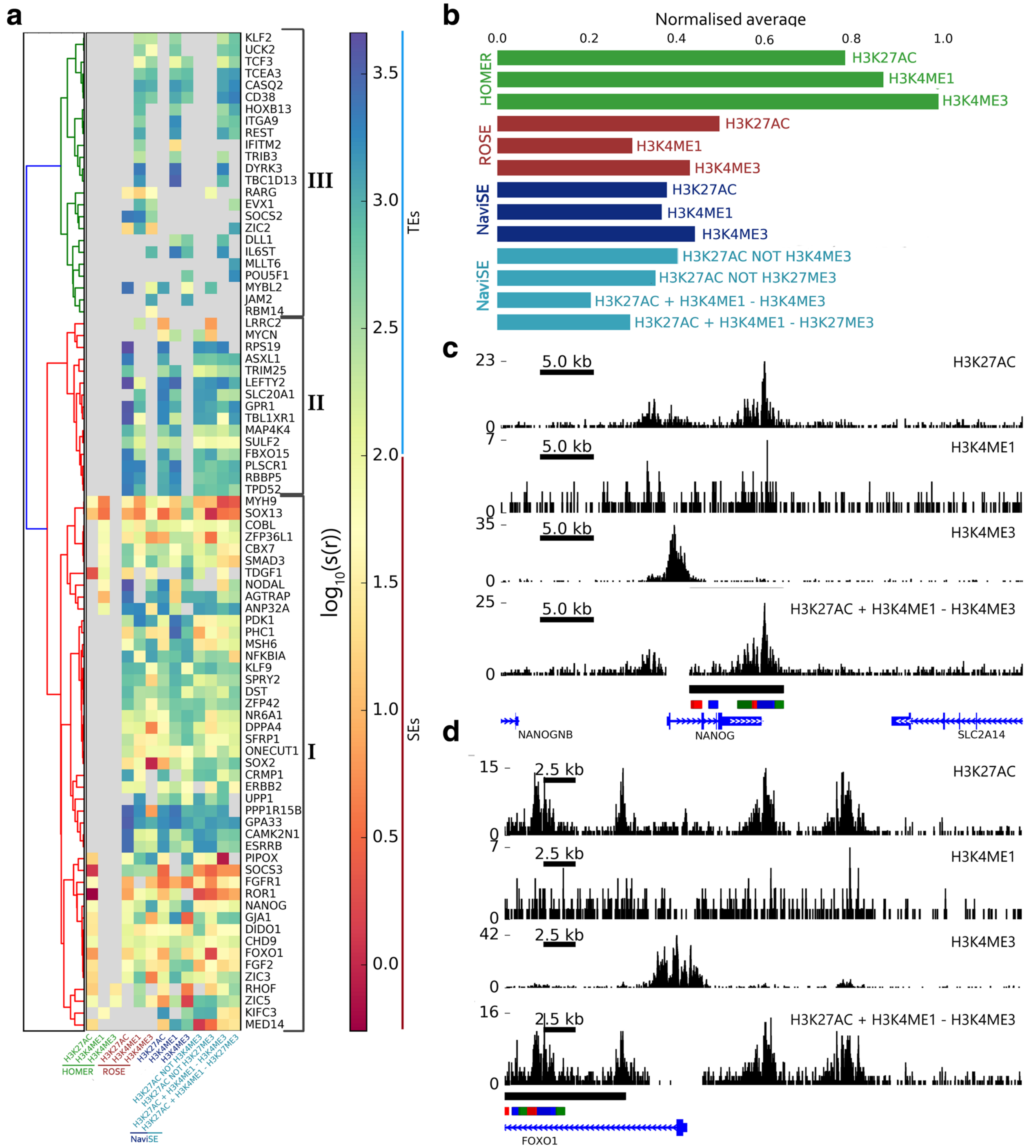


Fig. 11 Performance of the epigenomics algebra on ESC gene markers. **a** Heatmap of the scaled ranking $s(r)$ calculated by Eq. 6 for the SEs predicted by different $STIT_{pred-algebra}$ for ESC gene markers. The scaled ranking is colour coded in $\log_{10}$ scale, in *red* for good ranked (low ranking values) SEs, in *yellow* (from 2.0 onwards) for good ranked TEs, in *green* and *blue* for bad ranked TEs, and in *grey* for TEs without signal prediction. **b** Global goodness of each $STIT_{pred-algebra}$ over the whole set of ESC gene markers; normalised to the predictor of highest average (HOMER H3K4me3). Epigenomic algebra and single epigenomic signal box plot of peak distribution, depicted by our GVT module, for the SE associated to *NANOG* (**c**) and *FOXO1* (**d**). The *bottom row* contains the combination of epigenomic signals, and the rows above contain the original single signals. *Black lines* below the signal represent the SE supports at each sample, and bars in alternating colours below SE bar show the supports of the SE subpeak composition

functional analysis (GOEA, GSEA and enriched metabolic pathways) to PPI networks to a broad tissue prediction, thus, covering a wide range of valuable information. Such integrated annotation is of paramount importance due to the regulatory nature of the SEs, which have been described as key players in the determination of cell fate and in the involvement in the mechanisms of disease. Simultaneously, NaviSE performs all these tasks optimizing the use of the computer resources, identifying the available cores and main memory, and takes maximum advantage of them in function of the task requirements.

Furthermore, the automatic recognition of multiple file formats and the capability of working with replicates and controls, alongside with the possibility of integrating onto other pipelines or running multiple samples with multiple replicates and signal algebras at once with a simple script in Python, makes NaviSE a foremost tool for an efficient study of SEs. Due to all these capabilities, NaviSE is a time-saving and user-friendly tool for SE analysis.

To validate the biological performance of NaviSE, we applied it to predict the SEs on real data sets of several cell types with a different level of differentiation and commitment, and predicted in all cases SE-associated genes in agreement with the expected cell-specific markers. In the case of ESCs, NaviSE predicted SEs on the ESC markers *NANOG* and *SOX2*, in the case of neurons it predicted the *SOX1* and *CDK5R1* neuron markers, and in the case of monocytes, predicted the *CD86* and *CXCR4* monocyte markers.

The Additional file 1 provides a complete guide to the software installation and use instructions.

Availability and requirements

NaviSE. Project name: NaviSE. NaviSE is freely available at https://sourceforge.net/projects/navise-superenhancer/. **Operating system:** Linux 64bit (Ubuntu 11.04). **Programming language:** Python 3.5. License: GNU GPL.

Additional file

Additional file 1: Supplementary Information. Manual for installation, use and running examples of NaviSE. (pdf 33792 kb)

Abbreviations

ChIP-Seq: Chromatin immunoprecipitation sequencing; ChEA: ChIP enrichment analysis; DAG: Directed acyclic graph; ESC: Embryonic stem cell; FDR: False discovery rate; GB: Gigabyte; GO: Gene ontology; GSEA: Gene Set enrichment analysis; GOEA: Gene ontology enrichment analysis; GUI: Graphic user interface; GVT: Genomic viewer tool; HOMER: Hypergeometric optimization of motif enrichment; MACS: Model-based Analysis for ChIP-Seq; MSigDB: Molecular signatures database; MON: Monocyte; NEU: Neuron; NGS: Next generation sequencing; P3BSseq: Parallel processing pipeline software for automatic analysis of bisulfite sequencing data; PPI: Protein-protein interaction; ROSE: Ranking of superenhancers; SE: Superenhancer; TE: Typical enhancer; TF: Transcription factor; TFBS: Transcription factor binding site; TSS: Transcription start site

Acknowledgments

We thank Iñigo García de las Cuevas for providing the first model of the pipeline for file processing.

Funding

This work was supported by grants from the Ministry of Economy and Competitivity, Spain, MINECO grants PI16/01430 and BFU 2016-7798-P; DFG10/15, DFG15/15 and DFG141/16 from Diputación Foral de Gipuzkoa, Spain, and grant II14/00016 from I+D+I National Plan 2013−2016 of Carlos III Health Institute, Spain, FEDER funds and IKERBASQUE, Basque Foundation for Science, Spain; who have provided funding for the computational infrastructure to develop the project, of which this software is a part. AMA received a fellowship from the Ikasiker program (IkasC_2016_1_0017) of the Department of Education of the Basque Government, to support the development of the software as part of his final degree project. None of the funding bodies have played any part in the design of the study, in the collection, analysis, and interpretation of the data, or in the writing of the manuscript.

Authors' contributions

AMA developed and tested the software, obtained the results from biological samples, and wrote the manuscript. MAE contributed to the development of the installation software and the initial pipeline. AI designed and supervised the project. MJAB designed and supervised the project, and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Publisher's Note

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

Author details

[1]Computational Biology and Systems Biomedicine, Biodonostia Health Research Institute, 20014 San Sebastián, Spain. [2]Tissue Engineering Laboratory, Bioengineering Area, Biodonostia Health Research Institute, 20014 San Sebastián, Spain. [3]Department of Biochemistry and Molecular Biology, University of the Basque Country, 48940 Leioa, Spain. [4]IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain.

Received: 7 December 2016 Accepted: 18 May 2017
Published online: 06 June 2017

References

1. Pott S, Lieb JD. What are super-enhancers? Nature Gen. 2014;47(1):8–12.
2. Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI, Young RA. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell. 2013;153:307–19.
3. Adam RC, Yang H, Rockowitz S, Larsen SB, Nikolova M, Oristian DS, Polak L, Kadaja M, Asare A, Zheng D, Fuchs E. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. Nature. 2015;521(7552):366–70.
4. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-Enhancers in the Control of Cell Identity and Disease. Cell. 2013;155:934–47.
5. Heyn H, Vidal E, Ferreira HJ, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, Lin CY, Royo R, Sanchez-Mut JV, Martinez R, Gut M, Torrents D, Orozco M, Gut I, Young RA, Esteller M. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. Genome Biol. 2016;17(11):1–16.
6. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell. 2015;58(2):362–70.

7. Lovén J, Hoke HA, Lin CY, Lau A, Orlando DA, Vakoc CR, Bradner JE, Lee TI, Young RA. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell. 2013;153(2):320–34.
8. Luu PL, Gerovska D, Arrospide-Elgarresta M, Retegi-Carrión S, Schöler HR, Araúzo-Bravo MJ. P3BSSEQ: Parallel processing pipeline software for automatic analysis of bisulfite sequencing data. Bioinformatics. 2017;33(3):428–31.
9. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2,. Nat Methods. 2012;9:357–9.
10. Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A Hash-Based Algorithm for Accurate Next- Generation Sequencing Short-Read Mapping. PLoS ONE. 2014;9(3):1–11.
11. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754–60.
13. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):137.
14. Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE, Young RA. Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. Mol Cell. 2015;58:1–9.
15. Haibao T, Klopfenstein DV, Pedersen B, Ramirez F, Naldi A, Flick P, Yunes J, Sato K, Mungall C, Stupp G, DeTomaso D, Botvinnik O. GOATOOLS: Tools for Gene Ontology. Zenodo. 10.5281/zenodo.31628.
16. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinforma. 2013;14(128):1–14.
17. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43: 447–52.
18. Jaioun K, Teerapabolarn K. An improved binomial approximation for the hypergeometric distribution. Appl Math Sci. 2014;8(13):613–7.
19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005;102(43):15545–50.
20. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10.
21. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13(4):163.
22. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, Fischer CM, Gibson D, Gonzalez JN, Guruvadoo L, Haeussler M, Heitner S, Hinrichs AS, Karolchik D, Lee BT, Lee CM, Nejad P, Raney BJ, Rosenbloom KR, Speir ML, Villarreal C, Vivian J, Zweig AS, Haussler D, Kuhn RM, Kent WJ. The UCSC Genome Browser database: 2017 update. Nucleic Acids Res. 2017;4(45):626–34.
23. Afzal AR, Jeffery S. One gene, two phenotypes: ROR2 mutations in autosomal recessive Robinow syndrome and autosomal dominant brachydactyly type B. Hum Mutat. 2003;22(1):1–11.
24. Lim LS, Hong FH, Kunarso G, Stantonn LW. The pluripotency regulator Zic3 is a direct activator of the Nanog promoter in ESCs. Stem Cells. 2010;28(11):1961–9.
25. Takahashi K, Yamanaka S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell. 2006;126(4):663–76.
26. Luo F, Zhang J, Burke K, Miller RH, Yang Y. The Activators of Cyclin-Dependent Kinase 5 p35 and p39 Are Essential for Oligodendrocyte Maturation, Process Formation, and Myelination. J Neurosci. 2016;36(10):3024–37.
27. Kang J, Park H, Kim E. IRSp53/BAIAP2 in dendritic spine development, NMDA receptor regulation, and psychiatric disorders. Neuropharmacol. 2016;100:27–39.
28. Castro DS. One more factor joins the plot: Pbx1 regulates differentiation and survival of midbrain dopaminergic neurons. EMBO J. 2016;35(18): 1957–9.
29. Villaescusa JC, Li B, Toledo EM, Rivetti di Val Cervo P, Yang S, Stott SR, Kaiser K, Islam S, Gyllborg D, Laguna-Goya R, Landreh M, Lönnerberg P, Falk A, Bergman T, Barker RA, Linnarsson S, Selleri L, Arenas E. A PBX1 transcriptional network controls dopaminergic neuron development and is impaired in Parkinson's disease. EMBO J. 2016;35(18):1963–78.
30. Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik SG, Lee H, Lee JO. Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell. 2007;130:1071–82.
31. Hu MC, Wang YP, Mikhail A, Qiu WR, Tan TH. Murine p38-delta mitogen-activated protein kinase, a developmentally regulated protein kinase that is activated by stress and proinflammatory cytokines. J Biol Chem. 1999;274(11):7095–102.
32. Toubiana J, Rossi AL, Belaidouni N, Grimaldi D, Pene F, Chafey P, Comba B, Camoin L, Bismuth G, Claessens YE, Mira JP, Chiche JD. Src-family-tyrosine kinase Lyn is critical for TLR2-mediated NF-kB activation through the PI 3-kinase signaling pathway. Innate Immunol. 2015;21(7):685–97.
33. Parravicini V, Gadina M, Kovarova M, Odom S, Gonzalez-Espinosa C, Furumoto Y, Saitoh S, Samelson LE, O'Shea JJ, Rivera J. Fyn kinase initiates complementary signals required for IgE-dependent mast cell degranulation. Nat Immunol. 2002;3(8):741–8.
34. Kumano K, Chiba S, Kunisato A, Sata M, Saito T, Nakagami-Yamaguchi E, Yamaguchi T, Masuda S, Shimizu K, Takahashi T, Ogawa S, Hamada Y, Hirai H. Notch1 but Not Notch2 Is Essential for Generating Hematopoietic Stem Cells from Endothelial Cells. Immunity. 2003;18:699–711.
35. Ortega S, Ittmann M, Tsang SH, Ehrlich M, Basilico C. Neuronal defects and delayed wound healing in mice lacking fibroblast growth factor 2. PNAS. 1998;95(10):5672–77.
36. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, Slukvin II, Thomson JA. Induced pluripotent stem cell lines derived from human somatic cells. Science. 2007;318(5858):1917–20.
37. Chong JA, Tapia-Ramírez J, Kim S, Toledo-Aral JJ, Zheng Y, Boutros MC, Altshuller YM, Frohman MA, Kraner SD, Mandel G. REST: A Mammalian Silencer Protein That Restricts Sodium Channel Gene Expression to Neurons. Cell. 1995;80:949–57.
38. Schoenherr CJ, Anderson DJ. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. Science. 1995;267(5202):1360–3.
39. Phuchareon J, Tokuhisa T. Deregulated c-Fos/AP-1 accelerates cell cycle progression of B lymphocytes stimulated with lipopolysaccharide. Immunobiology. 1995;193(5):391–9.
40. Kang JG, Sung HJ, Jawed SI, Brenneman CL, Rao YN, Sher S, Facio FM, Biesecker LG, Quyyumi AA, Sachdev V, Hwang PM. FOS expression in blood as a LDL-independent marker of statin treatment. Atherosclerosis. 2010;212(2):567–70.
41. Ohkubo Y, Arima M, Arguni E, Okada S, Yamashita K, Asari S, Obata S, Sakamoto A, Hatano M, O-Wang J, Ebara M, Saisho H, Tokuhisa T. A role for c-fos/activator protein 1 in B lymphocyte terminal differentiation. J Immunol. 2005;174(12):7703–10.
42. Kim J, Woo AJ, Chu J, Snow JW, Fujiwara Y, Kim CG, Cantor AB, Orkin SH. A Myc Network Accounts for Similarities between Embryonic Stem and Cancer Cell Transcription Programs. Cell. 2010;143(2):313–24.
43. Müller-Molina AJ, Schöler HR, Araúzo-Bravo MJ. Comprehensive human transcription factor binding site map for combinatory binding motifs discovery. PLoS One. 2012;7(11):49086.