

Package 'HiCeekR'



Type: Package

Title: HiCeekR GUI for Hi-C data analysis

Version: 0.99.0

Description: HiCeekR is a novel Shiny based R package for Hi-C data analysis. In particular, HiCeekR combines several R/Bioconductor packages widely used for Hi-C data analysis and visualization. It starts from already aligned sequence files obtained from Hi-C experiments, then proceeds through a series of steps from pre-processing and filtering, to the evaluation and normalization of the contact matrices. Once the contact matrices are available, HiCeekR allows the users to perform several downstream analyses. Moreover, HiCeekR produces several interactive graphics that allow exploring the results by the usage of the mouse pointer.

Thanks to its GUI, HiCeekR friendly guides users during the entire analysis process, allowing them to perform a complete data analysis pipeline (i.e., pre-processing, filtering, binning, normalization, identification of compartments and TADs) and to integrate Hi-C data with other omic datasets such as ChIP-seq and RNA-seq.

License: GPL-2

Encoding: UTF-8

LazyData: true

RoxygenNote: 6.1.1

Imports: shiny,
shinyjs,
shinyFiles,
Matrix,
ggplot2,
plotly,
heatmaply,
corrplot,
networkD3,
hwriter,
gProfileR,
InteractionSet,
diffHic,
SummarizedExperiment,
Rsamtools,
rhdf5,
ReportingTools,
Haarfisz,
TopDom,
HiCseg

Contents

1. Introduction
2. Installation
3. To Get Start (configuration)
 - 3.1 Configuration
 - 3.2 Start new project or load an existing one
 - 3.3 Start new analysis or load an existing one
 - 3.4 GUI structure
 - 3.5 Summary
4. HiCeekR workflow
 - 4.1 Filtering
 - 4.2 Binning
 - 4.3 Normalization
 - 4.3.1 ICE normalization
 - 4.3.2 WavSis Normalization
 - 4.4 Post-Processing Analysis
 - 4.4.1 Identification of Compartments using PCA
 - 4.4.2 Identification of TADs by Directionality Index
 - 4.4.3 Identification of TADs by HiCsegTADs module
 - 4.4.4 Identification of TADs by TopDomTADs module
 - 4.4.5 Integration of ChIP-seq data using Epigenetic Features
 - 4.4.6 Import Bed file by bed2track module
 - 4.5 Results Visualization
 - 4.5.1 Chromosome Conformation & TADs (Heatmaps)
 - 4.5.2 Pathway Analysis & gene expression (Networks)
 - 4.5.2.1 Intersections Table
 - 4.5.2.2 Genes Table
 - 4.5.2.3 Enrichment Table
5. Input/Output structure & Folder organization

1 Introduction

HiCeekR analysis starts from already aligned sequence files obtained from Hi-C experiments, it proceeds through a series of steps from pre-processing and filtering to the evaluation and normalization of the contact matrices at the chosen resolution. Once the contact matrices have been computed, the user can choose among a series of functionalities for the downstream analysis. At the end of the process, HiCeekR leads to the identification of genome compartments and TADs and to the integration of Hi-C data with ChIP-Seq and/or RNA-Seq. Additionally, it allows to perform gene-enrichment pathway analysis and network visualization in order to better elucidate the interplay role between chromatin structure and gene regulation.

Thanks to its GUI, HiCeekR guides also non-expert users during the entire analysis process, allowing him/her to perform a complete data analysis pipeline. Moreover, HiCeekR provides several functions for integrating Hi-C data with other genome-wide omic data types and offers several visualization graphics (in interactive mode) to explore the results by the simple usage of mouse pointer

Through HiCeekR, each step/function can be executed sequentially for a step-by-step driven analysis. After each step is performed, the user can investigate intermediate results, such as summary statistics or graphical representations, before proceeding to the next step. However, each specific step or function can be re-executed by modifying the parameter settings and the results will be updated consequently. Intermediate and final results (as text files or figures) are stored in pre-organized data structures that can be retrieved for future investigations.

2 Installation

HiCeekR is freely available at <https://github.com/lucidif/HiCeekR>
It requires the R version 3.6.1 or above to be already installed

2.1 Via GitHub

The easiest way to install HiCeekR is via GitHub performing the following steps:

- i) Download and install *devtools* R package from CRAN web site.

Open R console and digit:

```
> install.packages("devtools")
```

- ii) Download and install *BiocManager* from Bioconductor website.
In the R console digit:

```
> if (!requireNamespace("BiocManager", quietly = TRUE))  
  install.packages("BiocManager")
```

- iii) Now install HiCeekR using the command

```
> library(devtools) ; install_github("HenrikBengtsson/TopDom") ;  
install_github("lucidif/HiCeekR", repos=BiocManager::repositories())
```

3 To get start

Once installed, launch HiCeekR Shiny app with the following commands:

```
> library("HiCeekR")
> HiCeekR()
```

3.1 Configuration

The first time HiCeekR is executed the user has to configure the computational environment as follows:

- i) Generate the SysVal.Robjfile. Such file is automatically generated by HiCeekR after its first execution in the current working directory. Note the first execution is the first time the user type **HiCeekR()** and the working directory is the current working directory of R.
- ii) Create configuration files (HCR.config, and HCRwd.info) which define the path where HiCeekR stores the required files and saves its results. Such files are created by the user using the following interface and saved in the current R working directory.

Welcome to HiCeekR

no config file found, please make a config file with panel below

.config

new

create new config file in:

select working folder:

explore

set

path selected: /home/

Parameters

- 1) **.config** : choose if you want to create a new config file or to load a pre-existed one.
- 2) **explore** : press this button to open the file manager to select the folder (if you selected "new") that you want use as main HiCeekR folder or to select pre-existent file (in case you selected "load").
- 3) **set** : press the "set" button when you selected the config folder/file

3.2 Start new project or load an existing one

HiCeekR data analysis is organized in user “projects” that are saved in the Project folders inside the **HiCeekRwd** working directory defined during the configuration step. Therefore, each time the user executes HiCeekR hi/she can choose either to start a new project (by typesetting the new project name) or continue working on an already existing project (by selecting the desired project name)

Welcome to HiCeekR

Select Project

1 new name: 2 make new 3

Show 25 entries Search:

4 list.files(path)

ooo

Parameters

- 1) **Select project:** choose if you want to make a *new* project or to *load* a previous one.
- 2) **name:** select the name of the project.
- 3) **make new/load:** make or load the project.
- 4) **list.files:** list of all existent analysis in HiCeekR main folder.

NOTE: Usually user projects are related to specific datasets. So that each user project corresponds to a given dataset consisting in one or more Hi-C experiments that are part of the same study. However, since the same Hi-C data can be analyzed at different resolution and/or using different parameter settings, HiceekR provides such possibility when setting the specific user Analysis within a given user project.

3.3 Start new analysis or load an existing one

Inside any project you can perform several data analyses. Usually different analyses are performed when the user wants to investigate results at different resolution or wants to change some parameters, without deleting previous results.

Therefore, in HiCeekR the user can create a new analysis by choosing the bam file, the restriction enzyme, the overhang parameter and the resolution, or can load previous performed analysis.

Then inside the corresponding project folder the user will find sub-folders corresponding to each executed analysis

The screenshot shows the HiCeekR web interface. At the top, a grey banner says "Welcome to HiCeekR". Below it, a section titled "Project Selected" contains a "reset project" button (callout 11). The main area is divided into "Analysis Settings" and "Analysis". In "Analysis Settings", there is a dropdown menu (callout 1) with "new" selected, and a "load" button (callout 3). Below this is the "input file type:" label and a dropdown menu (callout 4) with "BAM" selected. A "make new analysis" button (callout 10) is to the right. The "Analysis" section (callout 2) has a text input field. Below this, there are two buttons: "select BAM" (callout 5) and "select reference" (callout 6). Under "select files:", there are labels for "Bam:" and "Reference:". The "Set Enzyme" section has a "cut site" input field (callout 7) with "AAGCTT" and an "overhang" input field (callout 8) with "4". The "Other options" section has a "bin size" dropdown menu (callout 9) with "1000000".

Parameters

- 1) **Analysis Settings:** choose if you want to make a new analysis or to load a previous one.
- 2) **Analysis:** choose analysis name.
- 3) **load:** load previous analysis.
- 4) **Select input type:** select starting file type (only BAM type for now).
- 5) **Select BAM:** select the Hi-C BAM file to analyze.
- 6) **Select Reference:** select reference genome in FASTA format (use the same reference used for mapping)
- 7) **cut site:** specify the cut site of the enzyme used in wet protocol.
- 8) **overhang:** specify the overhang of the enzyme used in wet protocol.
- 9) **bin size:** specify the bins length (resolution) for the analysis (WARNING: the smaller the bin is, the more computational power it takes).
- 10) **make new analysis:** starts preliminary analysis.
- 11) **reset:** returns to begin.

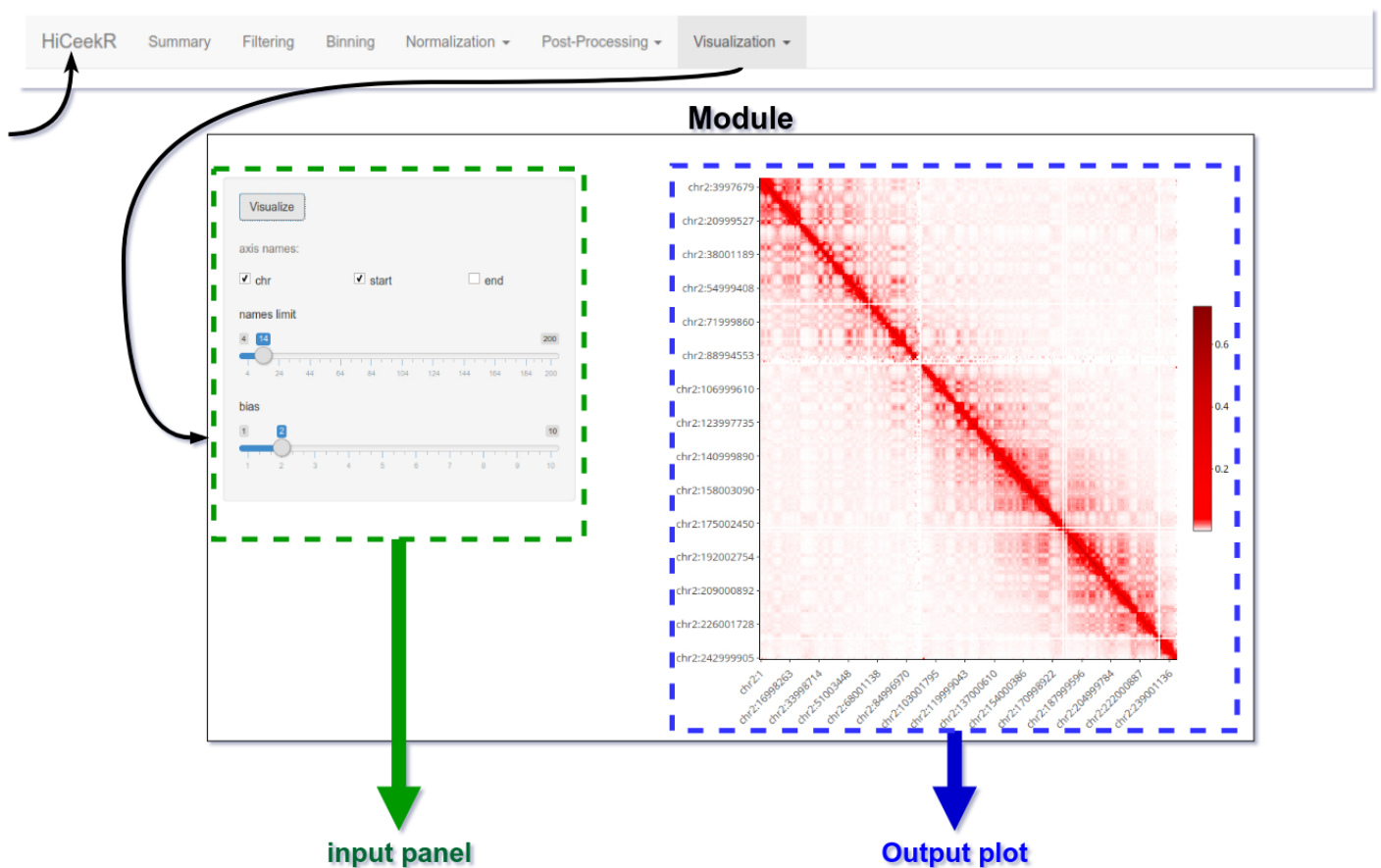
3.4 GUI structure

The upper part of HiCeekR interface displays the *navigation bar* illustrating all the analysis steps in sequential order. Each analysis step panel contains one or more specific functions.

By selecting one of them, HiCeekR renders the “*Function panel*” where input data files, function parameters and/or options (default values are suggested whenever possible) can be set before executing the function (the left-side of the interface allows the user to chose all parameters/options).

Results are shown in the “*Result panel*”, that is displayed in right-side of the interface, as plots or tables and/or automatically saved in a pre-structured way.

The graphical representations are interactive and allow exploring the results through point&click and dragging&dropping approach. Through this approach each step/function can be executed sequentially for a step-by-step driven analysis. After each step, the user can investigate intermediate results such as summary statistics or other graphics before proceeding to the next step.



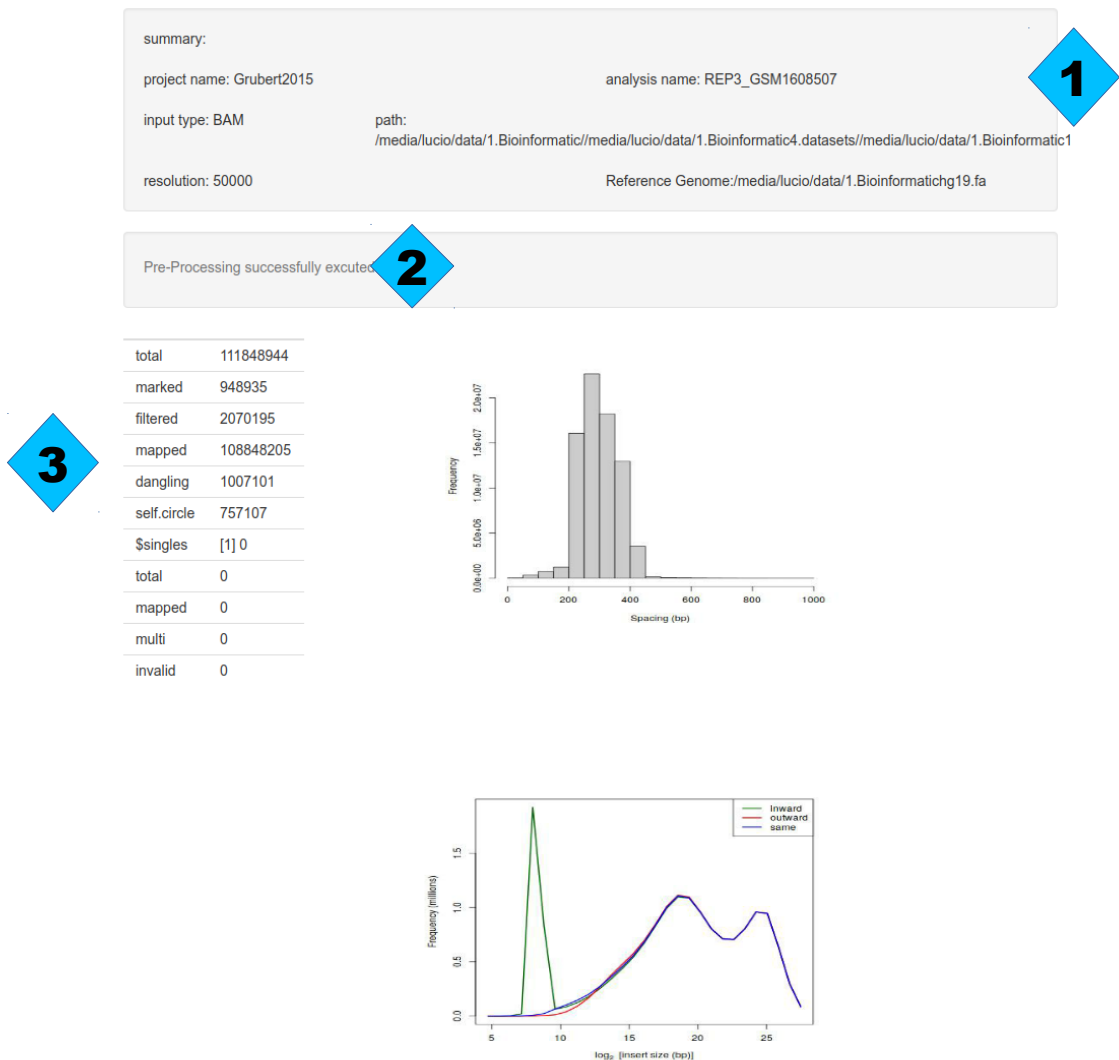
Current version of HiCeekR navigation bar includes the following choices:

- 1) Summary
- 2) Filtering
- 3) Binning
- 4) Normalization
- 5) Post-processing
- 6) Visualization

3.5 Summary

When an analysis is executed, the aligned bam file and the reference genome undergo to a pre-processing step that consists of a series of fundamental operations required for the proper execution of HiCeekR. Such operations allow HiCeekR to easily access the information in the subsequent steps and are aimed to reduce the overall execution time. After such pre-processing step basic information concerning the dataset under analysis are summarized in the Summary panel.

Moreover, after each step of the pipeline, all parameters and intermediate results are stored in the corresponding Project/analysis folder. Few additional summary information are showed in Summary table.



Results

- 1) **General Report:** general project information.
- 2) **Module general Report:** general module execution report.
- 3) **Module parameters & Results:** Module set parameters and output reports.

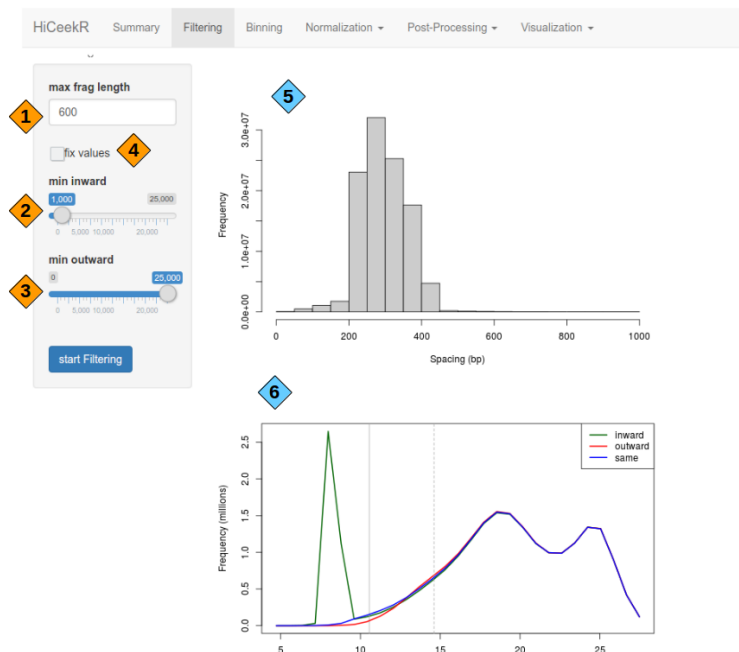
4 HiCeekR workflow

When the analysis parameters are defined (see 3.3 section), HiCeekR performs the pre-processing step. In such phase, the coordinates of each detected restriction fragment site are stored in an index-file (HDF5 file) and are associated with one or more mapped read. Such process is required in order to speed up the remaining steps and is not of direct interest for the user. Indeed, the produced intermediate files are used by HiCeekR, however they does not contain information in a user friendly form.

After the pre-processing HiCeekR performs: i) Filtering, ii) Binning, iii) Normalization, which are described in the following.

4.1 Filtering

The read-pair association process brings to a two case scenario, one where the read-pairs are associated to different restriction fragments and another where they are indexed to the same restriction fragment. The former case constitutes the set of valid reads, the latter occurs when un-ligated dangling-end or circularized self-circle fragments are present into the library preparation. The fragment-level filtering step is aimed to remove not only self-circle and dangling-end above mentioned events, but also the PCR-artifacts produced during the library preparation^{1,2,3,4}. HiCeekR automatically removes the provided duplicates, in case they have been already marked in the BAM files.



Parameters

- 1) The max.frag argument removes read pairs where the inferred length of the sequencing fragment (i.e., the ligation product) is greater than a specified value¹.
- 2 & 3) Min distance thresholds of inward and outward reads should be chosen with min.inward min.outward parameters¹
- 4) specify the numeric values for min.inward and min.outward parameters¹

Results

- 5) Fragments length distribution¹
- 6) Lower insert sizes distribution, spikes are observed in the outward- and inward-facing distributions due to self-circularization and dangling ends, respectively¹.

4.2 Binning

The binning step is aimed to perform all those operations needed to evaluate the raw contact matrix. The genome is partitioned into contiguous non-overlapping bins of constant size. Each interaction is defined as a pair of these bins.

This approach avoids the need for prior knowledge of the loci of interest when summarizing Hi-C counts. Counting of read pairs between paired bins is performed using binning module. Bin pairs can also be filtered to remove those with a count sum below a filter parameter value¹.

The screenshot shows the HiCekR web interface with the 'Binning' tab selected. The interface includes a top navigation bar with tabs: HiCekR, Summary, Filtering, Binning, Normalization, Post-Processing, and Visualization. On the left, there is a 'start' button (callout 1) and an 'Export' button (callout 6). The main area has two dropdown menus for 'first chromosome' (callout 2) and 'second chromosome' (callout 3), both set to 'chr2'. Below these is a 'show' button (callout 4). At the bottom, there is a 'Show 25 entries' dropdown (callout 5) and a search bar. The results are displayed in a table with 8 columns: T_Ch, T_start, T_end, T_width, A_Ch, A_start, A_end, and A_width. The table contains 7 rows of interaction data for chromosome chr2.

T_Ch	T_start	T_end	T_width	A_Ch	A_start	A_end	A_width
chr2	1	51591	51591	chr2	1	51591	51591
chr2	51588	97033	45446	chr2	1	51591	51591
chr2	51588	97033	45446	chr2	51588	97033	45446
chr2	97030	150161	53132	chr2	1	51591	51591
chr2	97030	150161	53132	chr2	51588	97033	45446
chr2	97030	150161	53132	chr2	97030	150161	53132
chr2	150158	199560	49403	chr2	1	51591	51591

Parameters

- 1) **Start:** start the binning step on all reference genome (all chromosomes VS all chromosomes)
- 2) **First Chromosome:** select exporting first chromosome of interest
- 3) **Second Chromosome:** select second exporting chromosome of interest, if you are interested to see inter-chromosome interactions, or the same chromosome, if you are interested to see intra-chromosome interactions.
- 4) **Export:** export results for selected chromosomes as tsv file
- 6) **show:** press "show" button after chromosomes selection to refresh table

Results

5) **Interaction Table:** tables that contain all interactions found for the selected chromosomes of interest (defined in points 2, 3); each row of the table contain the interacting bins in the first chromosome, chr, start position, end position and width and together with interacting bins in the second chromosome, chr, start position, end position and width, as illustrated in the above figures

4.3 Normalization

It is well recognized that several sequence-dependent artifacts can substantially bias Hi-C readouts. These include biases that are associated with sequencing platforms (such as GC content) and read alignment (such as mappability), and those that are specific to Hi-C (such as frequency of enzyme cutting sites).

The normalization step is aimed to remove technical biases from the raw contact matrix that could lead to false positive/negative findings. The output of such step is a normalized contact matrix (a symmetric square matrix) containing the strength of the interaction between two loci. Such matrix is important either for visualization purposes and for post-processing analysis.

Current version of HiCeekR implements the following normalization methods: ICE normalization and WavSis.

4.3.1 ICE Normalization

ICE is a correction method assuming that the bias in the interaction between two loci can be factorized as the product of the individual biases, affecting each of the two interacting loci⁵.

The screenshot shows the HiCeekR ICE Normalization web interface. It includes input fields for 'winsor.high' (set to 0), 'ignore.low' (set to 0), and a checkbox for 'set NA to min'. Below these are dropdown menus for 'first chromosome' and 'second chromosome', both set to 'chr1'. A 'Start Normalization' button is present. To the right, the 'chromosomes' field is set to 'chr1', 'sparsity' is 0.176557575962049, and the 'normalized matrix saved in:' field shows a file path. A '7' in a blue diamond highlights this path. Below the main form is an 'Export Data' section with a dropdown for 'Int Range' set to 'VS', checkboxes for 'add' and 'prefix', and an 'Export' button. A '6' in a blue diamond highlights the 'Export' button. Numbered callouts 1 through 7 are placed over various UI elements: 1 on the chromosome dropdowns, 2 on 'winsor.high', 3 on 'ignore.low', 4 on 'set NA to min', 5 on 'Start Normalization', 6 on the 'Export' button, and 7 on the file path.

Parameters

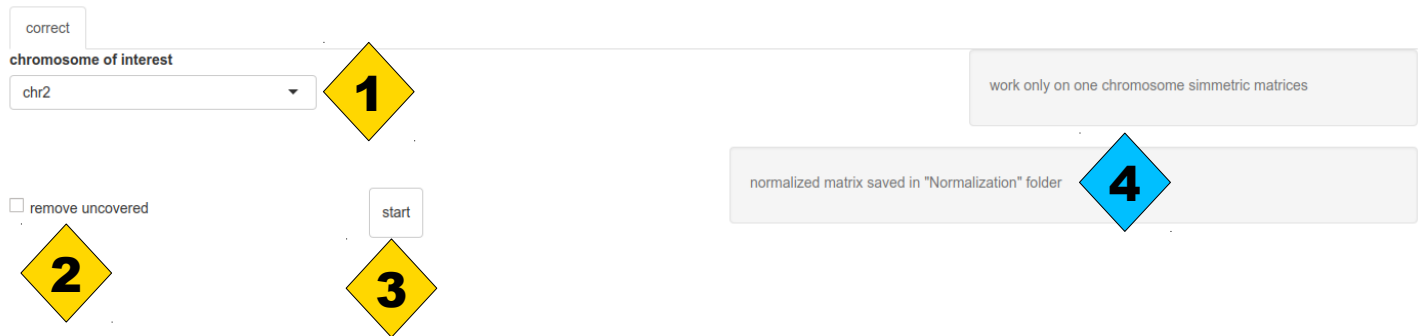
- 1) **Chromosome Box:** select the chromosomes of interest.
- 2) **Winsor.high:** a numeric scalar indicating the proportion of high-abundance bin pairs to winsorize¹.
- 3) **Ignore.low:** indicating the proportion of low-abundance bins to ignore¹.
- 4) **set NA to min:** Set all NA produced by ICE model as min matrix value, this can be make more noise (require for apply PCA).

Results

- 6) **export panel:** export results, can be set a prefix for outputfile by checking "add prefix" box and define prefix text by "add prefix" box.
- 7) **summary:** generated after the normalization it shows the matrix sparsity on the selected chromosomes and the path of the saved file.

4.3.2 WavSis Normalization

WavSis normalization combines wavelet methods and a Bayesian approach for correction (bias and noise) and comparison (detecting significant changes between Hi-C maps) of Hi-C contact maps. WavSis aims to remove noise by inspecting the variance distribution of the coverage across different physical scales, stabilizing the variance and applying a wavelet denoising strategy^{6,7}.



The diagram illustrates the WavSis Normalization interface with four numbered steps:

- 1**: Selecting the chromosome of interest (chr2) and clicking the 'correct' button.
- 2**: Clicking the 'remove uncovered' checkbox.
- 3**: Clicking the 'start' button.
- 4**: The normalized matrix is saved in the 'Normalization' folder.

Additional interface elements include a 'work only on one chromosome symmetric matrices' warning box and a 'normalized matrix saved in "Normalization" folder' confirmation box.

Parameters

- 1) Chromosome:** select chromosome of interest.
- 2) Remove uncovered:** specifying whether regions that are uncovered (i.e. no interaction with other regions) should be removed before correction⁵.
- 3) start:** start normalization.

Results

- 4) help text:** text that indicate the end normalization.

4.4 Post-Processing

The downstream analysis step implements several functionalities for extracting chromatin structures, interpreting the normalized contact matrix produced during previous steps, and integrating Hi-C data with other data types such as ChIP-seq data.

Current version of HiCeekR implements the PCA for the identification of Compartments, and three methods (directionality index, TopDom and HiCseg) for the identification of TADs. Moreover, it includes methods for preprocessing and integrating ChIP-seq data as additional tracks, and for processing any BED track of interest.

The above mentioned features are available in the modules following modules

- 1) PCA
- 2) Directionality index
- 3) TopDomTADs
- 4) HiCsegTADs
- 5) Epigenetic Features
- 6) bed2tracks

4.4.1 Identification of Compartments using PCA

The principal component (eigenvector) correlates with the distribution of genes and with features of open and close chromatin. With this module the user can execute the PCA starting from the contact matrix and from epigenetic markers^{7,8}.

The screenshot shows the 'PCA' module interface. It includes a 'Contact Matrix' dropdown menu with the file 'REP5_GSM1608509_chr2VSchr2_iterative_matrix.tsv' selected (callout 1). To the right is a 'PCs' input field with the value '3' (callout 2). Further right is a 'chromo of interest' dropdown menu with 'chr2' selected (callout 3). A 'start PCA' button is located on the far right (callout 4).

Parameters

- 1) **Contact Matrix:** select the contact matrix
- 2) **PCs:** number of Principal Components showed in the results.
- 3) **Chromosome of interest:** Chromosome to compute the PCA.
- 4) **start PCA:** start module analysis.

4.4.2 Identification of TADs by Directionality Index

This module computes the Directionality Index as defined in (Dixon et al 2012). The directionality index allows to define the TADs as the regions between two sharp changes in the sign of the index (i.e., a positive/negative switch identifies the TADs boundaries).

The directionality index consists of a genome-wide track that can be visualized on the genome as coverage file.

The screenshot shows the 'Directionality Index' module interface. It features a 'Contact Matrix' dropdown menu with 'none' selected (callout 1). Below it is a 'find DI' button (callout 2) and a text label 'create the file that contains the directional indexes'.

Parameters

- 1) **Select binTable file:** select raw counts matrix
- 2) **find DI:** calculate directionality index as described from Dixon¹²

4.4.3 Identification of TADs by HiCsegTADs module

The HiCsegTADs module implements the HiCseg algorithm¹³. It defines a partition on the contact matrix with a block structure depending on the unknown TADs boundaries. The parameters of the distributions are estimated by a maximum likelihood approach assuming that the observed contact values, within the same TADs share the same parameters. A maximum number of TADs per chromosome has to be specified (Kmax parameter)

Then, maximum likelihood estimates are obtained using a dynamic programming algorithm.

In this context, Gaussian distributions have to be used for modelling normalized contact matrix, whereas Poisson or Negative binomial distributions for raw contact matrix.

The type of TADs can be tuned with parameter D.

The module returns a BED tracks with TADs boundaries that can be visualized with the heatmap modules together with the other omic tracks.

The screenshot shows the HiCsegTADs module interface. It includes a 'Contact Matrix' section with a file input (chr2_WavSis_Norm.tsv.tsv), a 'Show' dropdown (25), and an 'entries' field. A 'distribution' dropdown is set to 'G'. A 'model' dropdown is set to 'D'. A 'Kmax' dropdown is set to '10'. A 'chromo of interest' dropdown is set to 'chr2'. A 'bin' dropdown is set to '1'. A 'HICsegJ' dropdown is set to '1'. A search bar is on the right. A table of results is displayed below the controls.

bin	HICsegJ
chr2:63995032-64048690	-332936192.23445
chr2:65049673-65105526	-329468020.197107
chr2:143549540-143597871	-326357981.749059
chr2:150401350-150449022	-323655697.134084
chr2:170151642-170200287	-321109419.393759
chr2:178802044-178850039	-318553007.643339
chr2:190298698-190349689	-316006729.903013
chr2:212152438-212203143	-313621548.065033

Parameters

- 1) **Contact Matrix:** select normalized matrix.
- 2) **distribution:** Distribution of the data: "B" is for Negative Binomial distribution, "P" is for the Poisson distribution and "G" is for the Gaussian distribution.
- 3) **model:** Type of model: "D" for block-diagonal and "Dplus" for the extended blockdiagonal model.
- 4) **Kmax:** max segmentations.
- 5) **Chromosome:** select chromosome of interest.

Results

- 6) **Table:** TADs boundaries in BED format

4.4.4 Identification of TADs by TopDomTADs module

The TopDomTADs module implements TopDom algorithm, as proposed in Shin et al. (2015)¹⁴. In particular, it defines a segmentation of the genome based on a three steps procedure: it evaluates the contact frequency signal as the average contact frequency of each bin with its upstream or downstream regions, then selects potential TADs boundaries as the local minima of the contact frequency signal, finally it filters out potential false positive by using Wilcoxon Rank Sum test under the assumption that the expected contact frequencies of regions within a TADs should be higher than those of a bin in the TADs and a bin outside the TADs, and of those bins outside the TADs. The number of bins to be included in upstream or downstream regions can be controlled by the user-parameter Window Size , which constitutes the only tuning parameter of TopDom algorithm.

The module returns a BED tracks with TADs boundaries that can be visualized with the heatmap modules together with the other omic tracks.

1

2

3

Contact Matrix
chr2_WavSis_Norm.tsv.tsv

Window Size
7

chromo of interest
chr2

start

Show 25 entries

4

chrom	chromStart	chromEnd	name
chr2	1	653105	domain
chr2	653102	3846223	domain
chr2	3846220	6949838	domain
chr2	6949835	7247772	domain
chr2	7247769	8849215	domain
chr2	8849212	8998071	boundary

Parameters

- 1) Contact Matrix:** select normalized matrix.
- 2) Windows size:** non negative integer number used to compare binSignal. Recommended size from 5 to 20.
- 3) Chromosome:** select chromosome of interest.

Results

- 4) Chromosome:** border bins finded.

4.4.5 Integration of ChIP-seq data using Epigenetic Features

Chromatin Immunoprecipitation sequencing (ChIP-seq) and DNase-seq are high throughput experimental technologies that have been shown to be effective in defining a detailed map of transcription factor-binding sites (TFBSs), histone modifications and open chromatin regions⁹. HiCeekR can upload one or more aligned BAM files from ChIP-Seq experiments and computes the normalized coverage at the same bin-width resolution

The screenshot shows the 'EpigeneticFeatures' interface. It includes a 'data type' dropdown (1) set to 'bint.tsv', a 'select binTable file' dropdown (2) set to 'allRegions.bint.bed', a 'bam file path' dropdown (3) set to 'please select file', a 'col name' input field (4), an 'add' checkbox (5), a 'file name' input field (6), an 'export table' button, a 'Normalization' checkbox (7), and a 'regions' table (8) with 8 rows of genomic coordinates. A search bar is also present.

regions
chr1:1-52423
chr1:52420-101020
chr1:101017-155676
chr1:155673-175547
chr1:175544-248992
chr1:248989-317877
chr1:317874-348337
chr1:348334-401277

Parameters

- 1) Data Type:** input Data format can be used to choice between pre-existent epigenetic module out (epiCounts file) or previuos step output file (bint.tsv).
- 2) selected binTable file:** select specific starting file.
- 3) bam file path:** indexed bam file produced from Chip-seq pipeline.
- 4) col name:** the name of column produced from coverage function.
- 5) add:** un-check this box if you're going to replace latest column with new one.
- 6) file name:** select a name for output file.
- 7)normalization:** check this box if you want to normalize the date; two type of normalization are available: "RPM", "byInput".

Results

- 8)results table:** table of covered reads inside the bins

4.4.6 Import Bed file by bed2track module

The bed2tracks module allow to convert any BED track into a HiCeekR BED track , i.e., evaluated at the same resolution analysis of the other data. Such track can be later visualized with the heatmap Visualization (see Section 4.5.1).

HiCeekR

Summary

Filtering

Binning

Normalization

Post-Processing

Visualization

File Name

CTCF

import bed file

Browse... HiCeekR ADs.bed

Upload complete

start

Show 25 entries

bin	CTCF
chr2:113550330-113594487	0
chr2:113594484-113649173	0
chr2:113649170-113697459	0
chr2:113697456-113750089	0
chr2:113750086-113798780	0
chr2:113798777-113848432	0
chr2:113848429-113899837	0
chr2:113899834-113959164	0
chr2:113959161-113999331	0
chr2:11398642-11451354	0
chr2:113999328-114051259	0
chr2:114051256-114099453	1
chr2:114099450-114147502	1
chr2:114147499-114198133	1
chr2:114198130-114250598	1

Parameters

- 1) **File Name:** name of output file.
- 2) **Import bed file:** select bed to import
- 3) **start:** export file in Downstream folder

Results

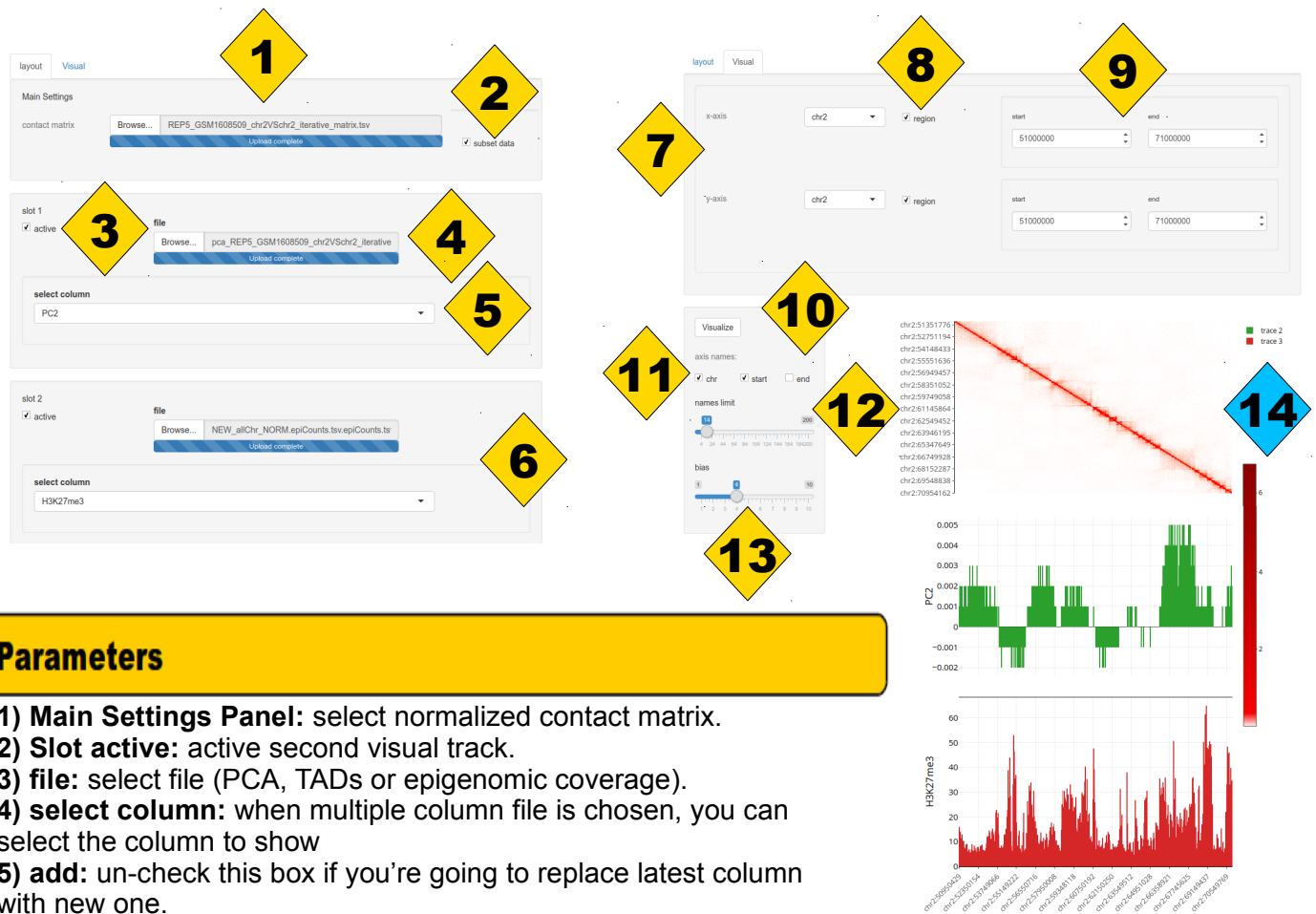
- 4) **results:** exported table overview.

4.5 Results Visualization

It is well known that the data visualization in a graphical form constitutes one of the most important data exploration tool. However, visualizing Hi-C data can be challenging due to the high-dimensionality of the files and the dimension of the genome. HiCeekR provides also different functions to visualize the results coming from the computational analysis without requiring additional software, moreover most of the HiCeekR plots are also interactive. In particular, the user can select two main representations: *Heatmap* and *Network*.

4.5.1 Chromosome Conformation & TADs (Heatmap)

With the Heatmap visualization HiCeekR allows to add and visualize multiple information such as the loadings of the PCA, the directionality index, as well as other genomic tracks, such as ChIP-Seq profiles or HiCeekR converted BED tracks.



Parameters

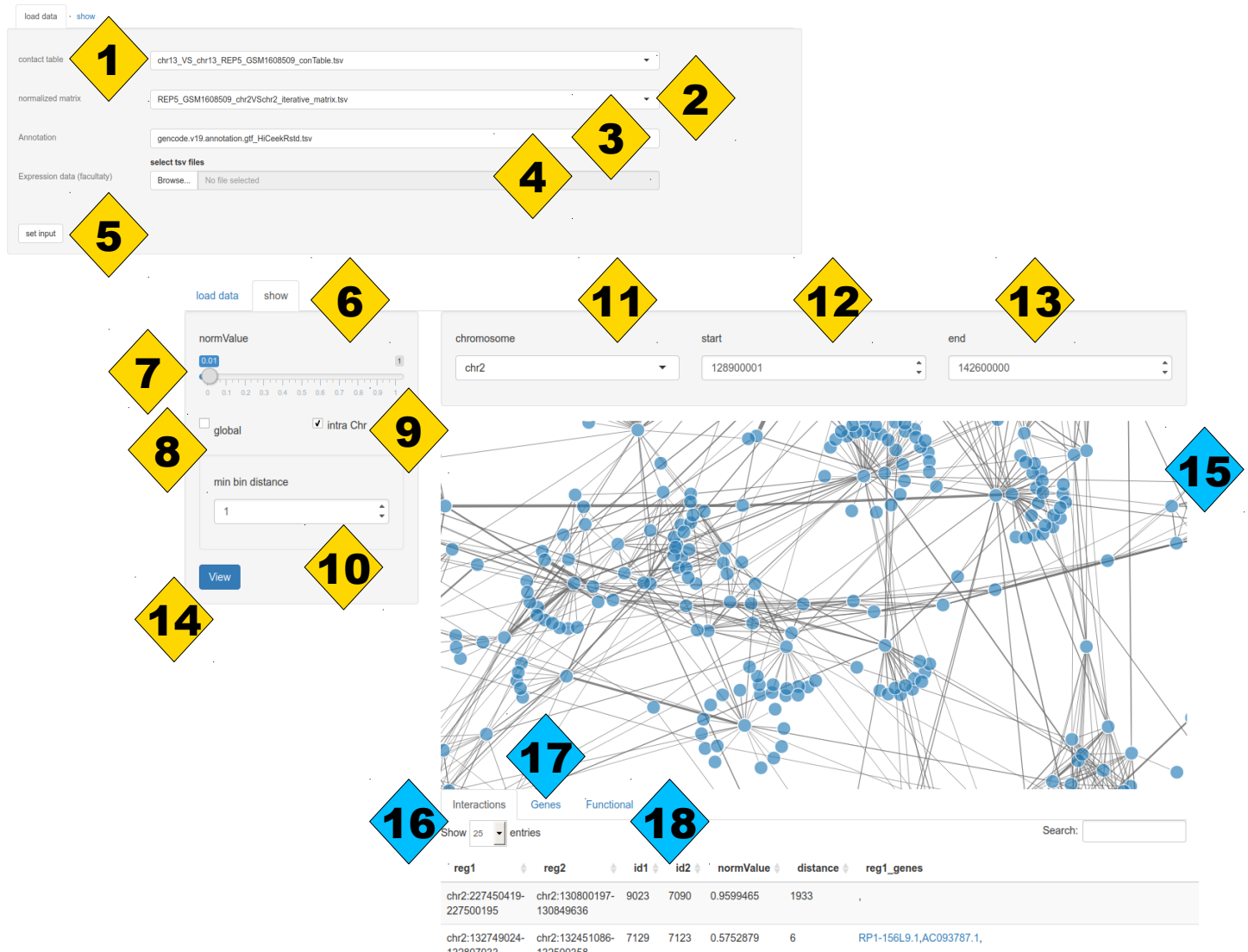
- 1) **Main Settings Panel:** select normalized contact matrix.
- 2) **Slot active:** active second visual track.
- 3) **file:** select file (PCA, TADs or epigenomic coverage).
- 4) **select column:** when multiple column file is chosen, you can select the column to show
- 5) **add:** un-check this box if you're going to replace latest column with new one.
- 6) **slot:** next slot (max 5 slot).
- 7) **coordinate selector:** it appears also if subset data check box are selected.
- 8) **region checkers:** check boxes can be used to select the coordinates to show, you can select axis to subset through a dedicated checkbox.
- 9) **region text box:** can be used to select specific coordinate to show.
- 10) **parameters selection panel:** can be used to select visualization parameters (press on Visualize button after settings configuration).
- 11) **plot names checkers:** can be used to select what to show on x-axis bin annotation (NAME and/or start bin coordinate and/or end bin coordinate).
- 12) **names limit:** how many names have to be showed on x-axis.
- 13) **bias:** heatmap color intensity.

Results

- 14) **Plots:** plots generated after the "visualization" button is pressed.

4.5.2 Networks

Using the Networks visualization the users can view interactions from a set of bins of interest against all other bins in form of networks where the vertices represent the bins and edges represent the detected interactions. The module is divided in two tab: “load data”, “show”.



Parameters

- 1) **Contact table:** table of contacts produced during the binning step.
- 2) **normalization matrix:** the contact matrix.
- 3) **annotation file:** annotations GTF file.
- 4) **expression data:** file of RNA-seq/microarray gene expressions (two columns Ensembl id and expression value).
- 5) **set input:** press on it to start preliminary analysis
- 6) **show panel:** it appears after pressing the set button
- 7) **normValue:** normvalue threshold to remove weak connections
- 8) **global:** if checked, the normalize Threshold are fixed on a specific value.
- 9) **intra chr:** show only chromosomal interactions
- 10) **min bin distance:** minimal bin distance (calculated in bins) to remove close interactions.
- 11) **chromosome:** chromosome region of interest
- 12) **start:** region of interest start position
- 13) **end:** region of interest end position
- 14) **View:** press to start the analysis.

Results

- 15) **Network:** interactive plot produced after the “View” button click.
- 16) **Interactions table:** table of identified interactions, with distances (calculated in bin) and genes contained inside its (see next section).
- 17) **Genes table:** All genes with coordinates inside the bins of interest, where for each gene the user can associated a gene expression value (if expression data has been selected in the “load data” panel; see next section).
- 18) **Functional table:** Functional gene-enrichment results obtained using the list of genes contained in bins (see show next section).

4.5.2.1 Intersections Table

Interactions

Genes

Functional

Show 25 entries

Search:

reg1	reg2	id1	id2	normValue	distance	reg1_genes	reg2_genes
chr2:40349794-40403050	chr2:40201887-40252751	5363	5360	0.021700660	3	SLC8A1,SLC8A1-AS1,	SLC8A1-AS1,
chr2:40298044	chr2:40100854-40149447	5361	5358	0.019076970	3	SLC8A1-AS1,	SLC8A1-AS1,
chr2:40403047-40449689	chr2:40201887-40252751	5364	5360	0.018195570	4	SLC8A1,SLC8A1-AS1,	SLC8A1-AS1,
chr2:40546390-40598607	chr2:40298041-40349797	5367	5362	0.015454020	5	SLC8A1,	SLC8A1,SLC8A1-AS1,
chr2:40298041-40349797	chr2:40149444-40201890	5362	5359	0.015266930	3	SLC8A1,SLC8A1-AS1,	SLC8A1-AS1,AC007253.1,
chr2:40500579-40546393	chr2:40252748-40298044	5366	5361	0.014682850	5	SLC8A1,	SLC8A1-AS1,
chr2:40403047-40449689	chr2:40252748-40298044	5364	5361	0.014640760	3	SLC8A1,SLC8A1-AS1,	SLC8A1-AS1,
chr2:40201887-40252751	chr2:40045577-40100857	5360	5357	0.014462080	3	SLC8A1-AS1,	SLC8A1-AS1,
chr2:40749995-40799151	chr2:40201887-40252751	5371	5360	0.014243710	11	SLC8A1,	SLC8A1-AS1,
chr2:40449686	chr2:40201887-40252751	5365	5360	0.014005500	5	SLC8A1,SLC8A1-AS1,	SLC8A1-AS1,

Results

Interactions table: table of detected interactions containing the name of the first region/bin, the name of the second region/bin, the first region bin index, the second region bin index, the normalized value of the interaction, the distance in bin between the two regions, the list of genes annotated in first region, the list of genes contained in second interacting bin
(Note all gene symbols are linked to genecard web page).

4.5.2.2 Genes Table

Interactions

Genes

Functional

select bin of interest

chr2:40349794-40403050

start

Show 25 entries

Search:

bin	start	end	geneName	ensembl	Copy_GSM2400247
chr2:37451599-37499234	37458774	37480546	NDUFAF7	ENSG00000003509	23.52
chr2:75852578-75899632	75879126	75938115	GCFC2	ENSG00000005436	22.89
chr2:242358164-242398810	242295658	242434256	FARP2	ENSG00000006607	46.09
chr2:37200800-37248383	37195526	37311485	HEATR5B	ENSG00000008869	43.39
chr2:37298066-37347382	37195526	37311485	HEATR5B	ENSG00000008869	43.39
chr2:37248380-37298069	37195526	37311485	HEATR5B	ENSG00000008869	43.39

Parameters

- 1)select bin of interest: choose the bin that you want to explore.
- 2)start button: HiceekR seeks for genes (and eventually gene expression level) inside bin of interest

Results

3)Genes Table: table containing bin name and bin coordinates (start and end), geneName, ensembl identifier, gene expression value in the RNA-seq/microarray uploaded file. Note all gene symbols are linked to genecard web page, all Ensembl identifiers are linked to Ensembl browser

Parameters

- 1)organism: select reference organism.
- 2)Enrich button: star Enrichment analysis

Results

3)Enrichment Table: gProfilerR package result that include, term unique ID; domani of annotation, terms, intersection of gene, p-value, term-size and overlap^{10,11}

4.5.2.3 Enrichment Table

1

Functional

organism

hsapiens

Enrich

Show 25 entries

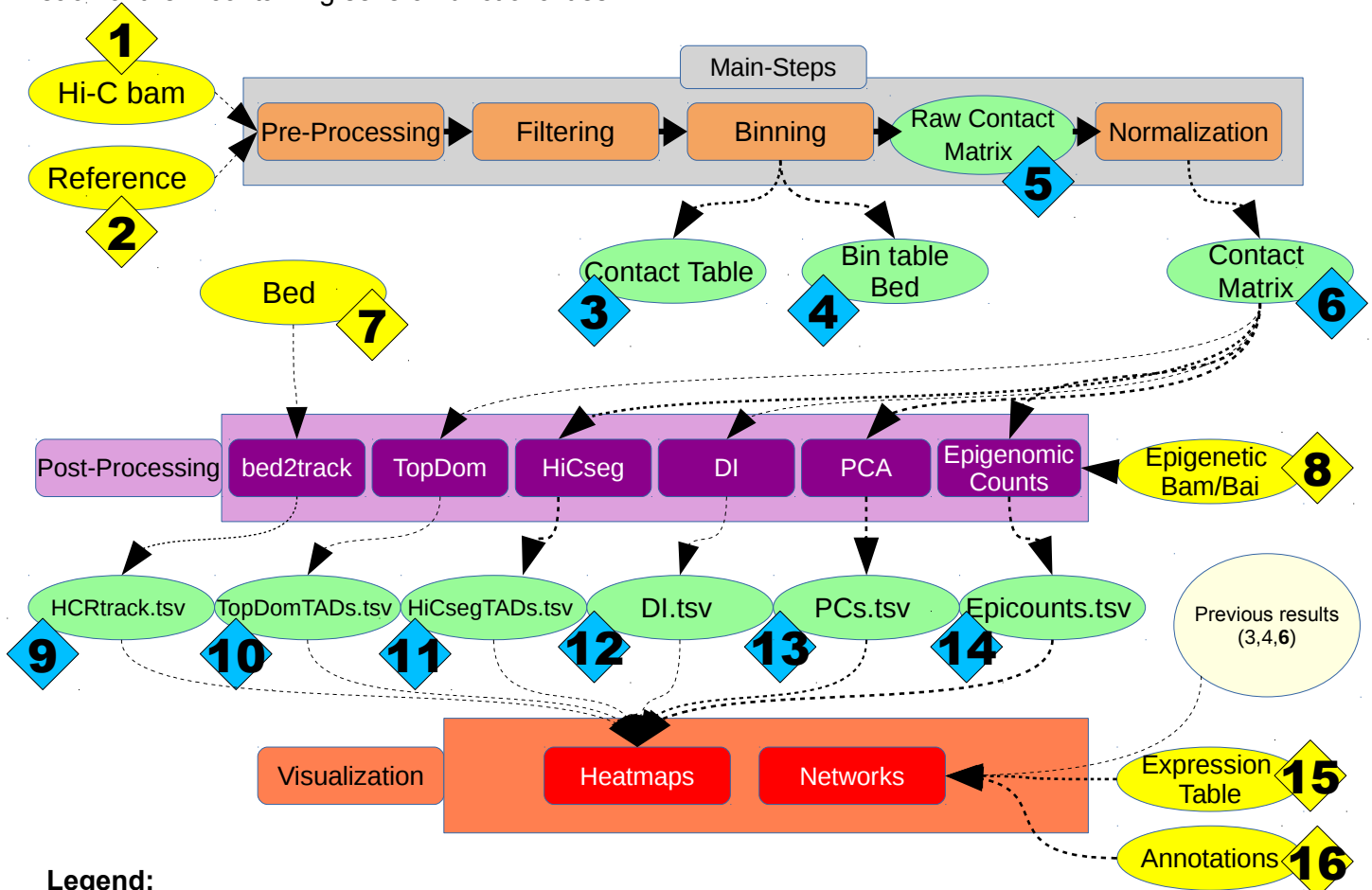
Search:

3

term.id	domain	term.name	intersection	P-value	term.size	overlap.size
GO:0035625	BP	obsolete epidermal growth factor-activated receptor transactivation by G-protein coupled receptor signaling pathway	ADRA2B	1	4	1
GO:0008150	BP	biological_process	NDUFAF7,GCFC2,MAP4K3,TTCT27,SPAST,VRK2,LTBP1,EIF2AK2,PUM2,MTA3,PMS1,ASB1,GTf2A1L,SRBD1,PSME4,NCK2,MAP4K4,TRIB2,ABCB11,ARHGAP15,DNAJC10,MAP2,RIF1,IGKV5-2,LBH,FOXJ3,NOTO,ARHGFEF33,RAD51AP2,PPP3R1,FTCDNL1,UGT1A5,RDH14,L	1	17568	288
GO:0051179	BP	localization	SPAST,LTBP1,CTNNA2,TTCT7A,STON1-GTF2A1L,NCK2,MAP4K4,ABCB11,MAP2,LRP2,APC	1	6415	110
GO:0051674	BP	localization of cell	CTNNA2,NCK2,APOB,ITGA4,GRB14,RTN4,TACR1,FN1,DNAH6,IL1R1,SOS1,EPHA4,TNP1,	1	1609	29
GO:0051234	BP	establishment of localization	SPAST,STON1-GTF2A1L,MAP4K4,ABCB11,MAP2,LRP2,APOB,NCOA1,RAB10,GCKR,CAD,	1	5083	86
GO:0006810	BP	transport	SPAST,STON1-GTF2A1L,MAP4K4,ABCB11,MAP2,LRP2,APOB,NCOA1,RAB10,GCKR,CAD,	1	4965	85
GO:0071705	BP	nitrogen compound transport	MAP4K4,LRP2,APOB,RAB10,GCKR,STAM2,ITGB6,SPTBN1,FN1,EHBP1,IL1R2,AFTPH,CC	1	2231	35
GO:0042886	BP	amide transport	MAP4K4,LRP2,APOB,RAB10,GCKR,STAM2,ITGB6,SPTBN1,FN1,EHBP1,IL1R2,AFTPH,CC	1	1938	31

5 Input/Output & Folder organization

The diagram below shows all steps executed inside HiCeekR and the input/output files of results. The pipeline can be divided in three macro-sections: **Main Steps**, **Post processing** and **Visualization**, each of them containing several functionalities.



Legend:

Yellow ellipses denote the input files that can be either experimental data or reference genome or gene annotations. **Green ellipses** denote intermediate or final results, in particular files labeled with 6,8,9,10 are those of major interest for the user and are used as input file for the visualization

- 1) **Hi-C Bam**: input Bam file produced from Hi-C data library sequencing.
- 2) **Reference**: input reference genome in fasta format.
- 3) **Contact table**: table of all contacts found during the Binning step. (tab delimited table. Each row is an interaction and has 8 columns respectively: chr, start, end and width of the first interacting regions versus chr, start, end and width of the second interacting region).
- 4) **bin table bed**: the bed file of all bins.
- 5) **Raw Contact Matrix**: Raw counts Contact matrix (Symmetric bin vs bin tab delimited matrix)
- 6) **Contact Matrix**: Normalized Contact Matrix (Symmetric bin vs bin tab delimited matrix)
- 7) **bed**: bed file with minimal required column (chromosome, start and end coordinates)
- 8) **Epigenetic Bam/Bai files**: Bam files (with bai indexes) obtained from external epigenetic Data analysis workflow.
- 9) **HCRtrack**: overlap between bin table bed [4] and bed file [7] (0 not overlap, 1 overlap with one or more bed file coordinates)
- 10) **TopDomTADs**: tab delimited TopDom boundaries detected (0 indicate not boundary bins; 1 indicate putative boundaries bins)
- 11) **HiCsegTADs**: tab delimited HiCseg boundaries detected (0 undetected; 1 indicate putative boundaries bins)
- 12) **DI**: tab delimited directional index results. Contains as rows the bins and as Columns Directionality indices.
- 13) **PCs**: tab delimited Eigenvector table that has as rows the bins and as columns the PCs loadings.
- 14) **epicounts**: tab delimited coverage table of epigenetic Bam/Bai files inside Hi-C bins (defined inside Bin Table Bed)
- 15) **Expression Table**: tab delimited table derived from expression experiments that has as rows the genes and as column the Expression Values (like FPKM or CPM etc.).
- 16) **Annotation file**: GTF file.

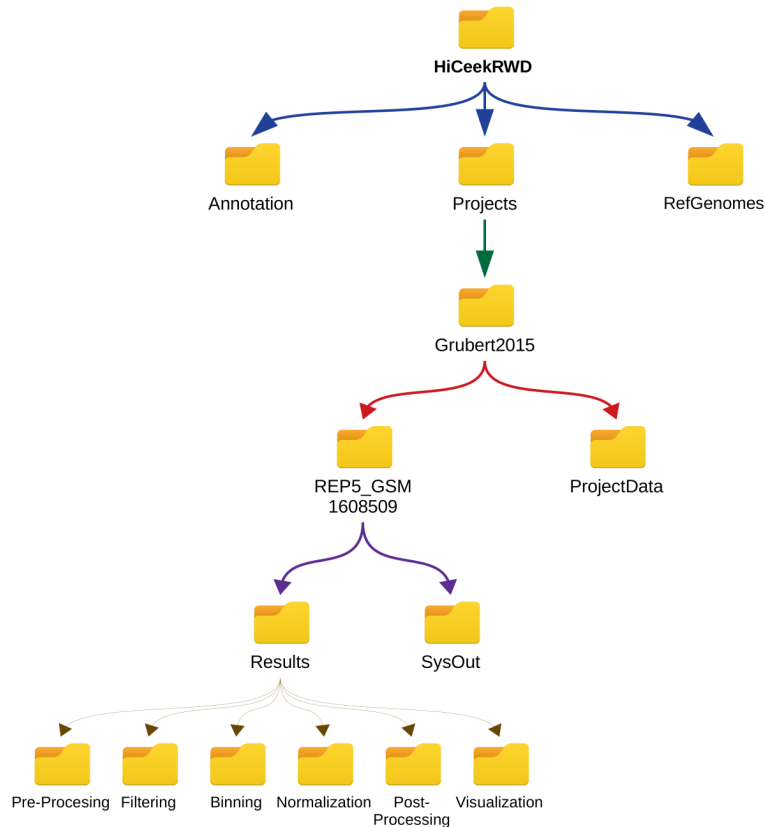
Folder organization:

HiCeekR allows to handle both user experimental data and other information such as the reference genome and annotations. All data and results are available under the **HiCeekRWD** folder that is created during HiCeekR configuration.

In particular, reference genomes have to be stored in the **RefGenomes** folder, gene annotations in the **Annotation** folder.

User defined project will be saved inside the **Projects** folder, where HiCeekR saves a sub-folder for each user project. Each used specific project contains a **ProjectData** folder containing both user experimental data Hi-C sequencing data (aligned bam file) and others omic related data (i.e., sequence aligned bam files from ChIP-seq experiments, gene expression files from RNA-seq or microarray experiments, that were obtained using other workflows) and an **Analysis** folder. The analysis folder contains the **Results** folder where results are organized according to the pipeline step in dedicated sub-folders.

Figures and tables as well as intermediated file that are generated in each step can be found in the corresponding sub-folders



References

1. <https://bioconductor.org/packages/release/bioc/html/diffHic.html>
2. Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi:10.1016/j.ymeth.2012.05.001
3. Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* 72, 65–75. doi:10.1016/j.ymeth.2014.10.031
4. Ay, F. and Noble, W. S. (2015). Analysis methods for studying the 3D architecture of the genome. *Genome Biology* 16, 183. doi:10.1186/s13059-015-0745-7
5. Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Iterative Correction of Hi-C Data Reveals Hallmarks of Chromosome Organization. *Nature Methods* 9, 999–1003. doi:10.1038/nmeth.2148
6. Shavit, Y. and Lio, P. (2014). Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. BioSyst.* 10, 1576–1585. doi:10.1039/C4MB00142G
7. <https://cran.r-project.org/web/packages/chromR/index.html>
8. <https://bioconductor.org/packages/release/bioc/html/HiTC.html>
9. Xun Lan, Heather Witt, Koichi Katsumura, Zhenqing Ye, Qianben Wang, Emery H. Bresnick, Peggy J. Farnham and Victor X. Jin (2012). Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. *Nucleic Acids Res* 40(16), 7690–7704. doi:10.1093/nar/gks501
10. <https://cran.r-project.org/web/packages/gProfileR/gProfileR.pdf>
11. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 1–8. doi:10.1093/nar/gkz369
12. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012 Apr 11;485(7398):376–80. doi: 10.1038/nature11082.
13. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*. 2014;30(17):i386–i392. doi:10.1093/bioinformatics/btu443
14. Shin H, Shi Y, Dai C, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res*. 2016;44(7):e70. doi:10.1093/nar/gkv1505