
nucleo_miner Documentation

Release 2.3.42

Florent CHUFFART, Jean-Baptiste VEYRIERAS, Gael YVERT

May 27, 2014

CONTENTS

1	Readme / Documentation for <i>NucleoMiner2</i>	1
1.1	License	1
1.2	Installation Instructions	1
1.3	usage	2
2	Tutorial	3
2.1	Dataset and Configuration File	3
2.2	Preprocessing Illumina Fastq Reads for Each Sample	4
2.3	Inferring Nucleosome Position and Extracting Read Counts	6
2.4	Results	7
3	References	11
3.1	Python Reference	11
3.2	R Reference	11
4	Indices and tables	39

README / DOCUMENTATION FOR *NUCLEOMINER2*

NucleoMiner2 offers Python API and R package allowing to perform quantitative analysis of nucleosomal epigenome. It is especially well suited for scripting to extract natural Single-Nucleosome Epi- Polymorphisms (SNEP) from ChIP-Seq data.

1.1 License

Copyright CNRS 2012-2013

- Florent CHUFFART
- Jean-Baptiste VEYRIERAS
- Gael YVERT

This software is a computer program which purpose is to perform quantitative analysis of epigenetic marks at single nucleosome resolution.

This software is governed by the CeCILL license under French law and abiding by the rules of distribution of free software. You can use, modify and/ or redistribute the software under the terms of the CeCILL license as circulated by CEA, CNRS and INRIA at the following URL "<http://www.cecill.info>".

As a counterpart to the access to the source code and rights to copy, modify and redistribute granted by the license, users are provided only with a limited warranty and the software's author, the holder of the economic rights, and the successive licensors have only limited liability.

In this respect, the user's attention is drawn to the risks associated with loading, using, modifying and/or developing or reproducing the software by the user in light of its specific status of free software, that may mean that it is complicated to manipulate, and that also therefore means that it is reserved for developers and experienced professionals having in-depth computer knowledge. Users are therefore encouraged to load and test the software's suitability as regards their requirements in conditions enabling the security of their systems and/or data to be ensured and, more generally, to use and operate it in the same conditions as regards security.

The fact that you are presently reading this means that you have had knowledge of the CeCILL license and that you accept its terms.

1.2 Installation Instructions

1.2.1 Links

NucleoMiner2 home page and documentation: <https://forge.cbp.ens-lyon.fr/redmine/projects/nucleominer>

Gael Yvert lab page: <http://www.ens-lyon.fr/LBMC/gisv/>

1.2.2 Installation

- Download archive
- Compile bowtie2
- Compile samtools
- Compile bedtools
- Compile TemplateFilter

Required R packages:

- bot - fork
- rjson
- seqinr

```
cd src/r_packages/  
tar xfvz R-latest.tar.gz  
cd R-patched  
./configure --with-x=no PDFLATEX="ls"  
make  
cd ../../..  
R_BIN=src/r_packages/R-patched/bin/R  
$R_BIN CMD INSTALL src/r_packages/rjson_0.2.12.tar.gz  
$R_BIN CMD INSTALL src/r_packages/seqinr_3.0-7.tar.gz  
$R_BIN CMD INSTALL src/r_packages/plotrix_3.4-5.tar.gz  
$R_BIN CMD INSTALL src/r_packages/nm_2.0.tar.gz  
$R_BIN CMD INSTALL src/r_packages/fork_1.2.4.tar.gz  
$R_BIN CMD INSTALL src/r_packages/bot_0.9.tar.gz  
$R_BIN CMD INSTALL src/r_packages/DESeq_1.14.0.tar.gz  
  
...
```

1.3 usage

See html documentation for *NucleoMiner2*: <http://www.ens-lyon.fr/LBMC/gisv/>

TUTORIAL

This tutorial describes steps allowing to perform quantitative analysis of nucleosomal epigenome. We assume that files are organised around a given hierarchy and that all command lines are launched from project's root.

This tutorial is divided into two main parts. First one consists in the python script *wf.py* that aligns and convert Illumina reads. Second one is the R script *main.r* that extracts information (nucleosome position and indicators) from the dataset.

2.1 Dataset and Configuration File

We want to compare nucleosomes of 3 yeast strains:

- BY
- RM
- YJM

For each strain we perform Mnase-Seq and ChIP-Seq using the 5 following markers:

- H3K4me1
- H3K4me3
- H3K9ac
- H3K14ac
- H4K12ac

In order to simplify the design of experiment, we consider Mnase as a marker. For each couple (*strain*, *marker*) we perform 3 replicates. So, theoretically we should have $3 * (1 + 5) * 3 = 54$ samples. In practice we only obtain 2 replicates for (*YJM*, *H3K4me1*). Each one of the 53 samples is identified by a unique identifier. The file *CSV_SAMPLE_FILE* sums up this information.

We use a convention to link sample and Illumina fastq outputs. Illumina output files of the sample *ID* will be stored in the directory *ILLUMINA_OUTPUTFILE_PREFIX* + *ID*. For example, sample 41 outputs will be stored in the directory *data/2012-09-05/FASTQ/Sample_Yvert_Bq41/*.

For BY (resp. RM and YJM) we use following reference genome *saccharomyces_cerevisiae_BY_S288c_chromosomes.fasta* (resp. *saccharomyces_cerevisiae_rm11-1a_1_supercontigs.fasta* and *saccharomyces_cerevisiae_YJM_789_screencontig.fasta*). The index *FASTA_REFERENCE_GENOME_FILES* stores this information.

Each chromosome/contig is identified in the fasta file by an obscure identifier. For example, BY chromosome I is identified by *gil144228165|refNC_001133.7|* when TemplateFilter is waiting for an integer. So, we translate it. The index *FASTA_INDEXES* stores this translation.

From a pragmatismal point of view we discard some part of the genome (repeated sequence etc...). The list of the black listed area is explicitly detailed in *AREA_BLACK_LIST*.

For BY-RM (resp. BY-YJM and RM-YJM) genome sequence alignment we use previously compute .c2c file *data/2012-03_primarydata/BY_RM_gxcomp.c2c* (resp. *BY_YJM_GComp_All.c2c* and *RM_YJM_gxcomp.c2c*). For more information about .c2c files, please read section 5 of the manual of *NucleoMiner*, the old version of *NucleoMiner2* (http://www.ens-lyon.fr/LBMC/gisv/NucleoMiner_Manual/manual.pdf).

nucleominer uses specific directory to work in, these are described in *INDEX_DIR*, *ALIGN_DIR* and *LOG_DIR*.

Finally, *nucleominer* use external ressources, the path to these ressources are describe in *BOWTIE_BUILD_BIN*, *BOWTIE2_BIN*, *SAMTOOLS_BIN*, *BEDTOOLS_BIN* and *TF_BIN* and *TF_TEMPLATES_FILE*.

All paths, prefixes and indexes could be change in the *src/current/nucleominer_config.json* file.

2.2 Preprocessing Illumina Fastq Reads for Each Sample

This preprocessing step consists in the 4 main steps embed in the *wf.py* and described bellow. As a preamble, this script computes *samples* *samples_mnase* and *strains* that will be used along the 4 steps.

2.2.1 Creating Bowtie Index from each Reference Genome

For each strain, we need to create bowtie index. Bowtie index of a strain is a tree view of the genomoe reference for this strain. It will be used by bowtie to align reads. This step is performed by the following part of the *wf.py* script:

The following table sum up involved file sizes and process durations concerning this step.

strain	fasta genome file size	bowtie index file size	process duration
BY	12 Mo	25 Mo	11 s.
RM	12 Mo	24 Mo	9 s.
YJM	12 Mo	25 Mo	11 s.

2.2.2 Aligning Reads to Reference Genome

Next, we launch bowtie to align reads to the reference genome. It produces a *.sam* file that we convert into a *.bed* file. Binaries for *bowtie*, *samtools* and *bedtools* are wrapped using python *subprocess* class. This step is performed by the followinw part of the *wf.py* script:

2.2.3 Convert Aligned Reads for TemplateFilter

TemplateFilter use particular input format for reads, so we convert *.bed* file. TemplateFilter expect reads as following: *chr coord strand #read* where:

- chr is the number of the chromosome;
- coord is the coordinate of the reads;
- strand is *F* for forward and *R* for reverse;
- #reads the number of reads for this position.

Each entry is *tab*-separated.

WARNING for reverse strand bowtie returns the position of left first nucleotid when TemplateFilter is waiting for right one. So this step takes it into account and add lenght of reads (in our case 50) to reverse reads coordinate.

This step is performed by the following part of the *wf.py* script:

The following table sum up number of reads, involved file sizes and process durations concerning the two last steps. In our case, alignment process have been multithreaded over over 3 cores.

id	Illumina reads	aligned and filtered reads	ratio	.bed file size	TF input file size	process duration
1	16436138	10199695	62,06%	1064 Mo	60 Mo	383 s.
2	16911132	12512727	73,99%	1298 Mo	64 Mo	437 s.
3	15946902	12340426	77,38%	1280 Mo	65 Mo	423 s.
4	13765584	10381903	75,42%	931 Mo	59 Mo	352 s.
5	15168268	11502855	75,83%	1031 Mo	64 Mo	386 s.
6	18850820	14024905	74,40%	1254 Mo	69 Mo	482 s.
7	15591124	12126623	77,78%	1163 Mo	72 Mo	405 s.
8	15659905	12475664	79,67%	1194 Mo	71 Mo	416 s.
9	14668641	10960565	74,72%	1052 Mo	70 Mo	375 s.
10	14339179	10454451	72,91%	1049 Mo	51 Mo	363 s.
11	18019895	13688774	75,96%	1378 Mo	59 Mo	474 s.
12	13746796	10810022	78,64%	1084 Mo	54 Mo	360 s.
13	15205065	11766016	77,38%	990 Mo	54 Mo	381 s.
14	17803097	13838883	77,73%	1154 Mo	60 Mo	452 s.
15	15434564	12307878	79,74%	1032 Mo	57 Mo	394 s.
16	16802587	12725665	75,74%	1221 Mo	48 Mo	438 s.
17	16058417	12513734	77,93%	1192 Mo	63 Mo	422 s.
18	16154482	13204331	81,74%	1277 Mo	52 Mo	430 s.
19	21013924	17102120	81,38%	1646 Mo	59 Mo	555 s.
20	17213114	14433357	83,85%	1389 Mo	53 Mo	459 s.
21	17360907	14733001	84,86%	1203 Mo	55 Mo	450 s.
22	18136816	15389581	84,85%	1257 Mo	53 Mo	469 s.
23	14763678	12173025	82,45%	1140 Mo	56 Mo	393 s.
24	15541709	12890345	82,94%	1057 Mo	48 Mo	398 s.
25	16433215	13094314	79,68%	1241 Mo	57 Mo	433 s.
26	17370850	14264136	82,12%	1347 Mo	51 Mo	466 s.
27	14613512	8654495	59,22%	887 Mo	56 Mo	339 s.
28	15248545	11367589	74,55%	1166 Mo	67 Mo	405 s.
29	14316809	10767926	75,21%	1103 Mo	63 Mo	379 s.
30	15178058	12265794	80,81%	1030 Mo	66 Mo	390 s.
31	14968579	11876186	79,34%	1009 Mo	63 Mo	387 s.
32	16912705	13550508	80,12%	1143 Mo	70 Mo	442 s.
33	16782154	12755111	76,00%	1227 Mo	65 Mo	438 s.
34	16741443	13168071	78,66%	1260 Mo	71 Mo	442 s.
35	13096171	10367041	79,16%	992 Mo	62 Mo	350 s.
36	17715118	14092985	79,55%	1404 Mo	68 Mo	483 s.
37	17288466	7402082	42,82%	741 Mo	48 Mo	339 s.
38	16116394	13178457	81,77%	1101 Mo	63 Mo	420 s.
39	14241106	10537228	73,99%	880 Mo	57 Mo	348 s.
40	13784738	10598464	76,89%	1005 Mo	64 Mo	358 s.
41	12438007	9620975	77,35%	911 Mo	60 Mo	326 s.
42	13853959	11031238	79,63%	1045 Mo	64 Mo	365 s.
43	12036162	6654780	55,29%	684 Mo	46 Mo	268 s.
44	13873129	10251074	73,89%	1048 Mo	61 Mo	365 s.
45	19817751	14904502	75,21%	1520 Mo	72 Mo	528 s.
46	13368959	10818619	80,92%	912 Mo	63 Mo	350 s.
47	7566467	6139001	81,13%	520 Mo	44 Mo	201 s.
48	32586928	21191363	65,03%	1816 Mo	82 Mo	766 s.

Continued on next page

Table 2.1 – continued from previous page

id	Illumina reads	aligned and filtered reads	ratio	.bed file size	TF input file size	process duration
49	30733184	18791373	61,14%	1801 Mo	89 Mo	721 s.
50	41287616	30383875	73,59%	2911 Mo	112 Mo	1065 s.
51	40439965	31177914	77,10%	2981 Mo	117 Mo	1070 s.
53	40876476	33780065	82,64%	3316 Mo	103 Mo	1165 s.
55	52424414	47117107	89,88%	3811 Mo	119 Mo	1477 s.

For some reasons (manipulation efficiency, e.g. PCR...), we remove samples 33, 45, 48 and 55.

2.2.4 Run TemplateFilter on Mnase Samples

Finally, for each sample we performe TemplateFilter analysis.

WARNING TemplateFilter returns a list of nucleosomes. Each nucleosome is define by its center and its width. An odd width leads us to considere non interger lower and upper bound.

WARNING TemplateFilter is not design to deal with replicate. So we choose to keep a maximum of nucleosome and filter in a second time using the benefit of replicate. To do that we set a low correlation threshold parameter (0.5) and a particularly high value of overlapping (300%).

This step is performed by the followinw part of the *wf.py* script:

id	strain	found nucs	nuc file size	process duration
1	BY	96214	68 Mo	1022 s.
2	BY	91694	65 Mo	1038 s.
3	BY	91205	65 Mo	1036 s.
4	RM	88076	62 Mo	984 s.
5	RM	90141	64 Mo	967 s.
6	RM	87517	62 Mo	980 s.
7	YJM	88945	64 Mo	566 s.
8	YJM	88689	64 Mo	570 s.
9	YJM	88128	63 Mo	565 s.

2.3 Inferring Nucleosome Position and Extracting Read Counts

This preprocessing step consists in the 4 main steps embed in the *wf.py* and described bellow. As a preamble, this script computes *samples* *samples_mnase* and *strains* that will be used along the 4 steps.

The second part of the tutorial use *R* (<http://http://www.r-project.org>). It consists in the following main steps:

- compute_rois.R
- extract_maps.R
- compare_common_wp.R
- split_samples.R
- count_reads.R
- get_size_factors
- launch_deseq.R

2.3.1 Computing Common Genome Region Between Strains

```
R CMD BATCH src/current/compute_rois.R
```

2.3.2 Extracting Maps for Well Positionned and Fuzzy Nucleosomes

```
R CMD BATCH src/current/extract_maps.R
```

2.3.3 Compute Distance Between Well Positionned Nucleosomes

```
R CMD BATCH src/current/compare_common_wp.R
```

2.3.4 Split and Compress Samples According CURs

```
R CMD BATCH src/current/split_samples.R
```

2.3.5 Count Reads for Each Nucleosome

```
R CMD BATCH src/current/count_reads.R
```

2.3.6 Get Size Factors Using DESeq

```
R CMD BATCH src/current/get_size_factors.R
```

2.3.7 Performing DESeq Analysis

```
R CMD BATCH src/current/launch_deseq.R
```

2.4 Results

2.4.1 Output Files Organisation

Previous steps produce following 45 files. Each filename is under the form

```
results/current/[combi]_[marker]_[form]_snep.tab
```

Where combi is in {BY_RM, BY_YJM, RM_YJM} for each strain combination, marker is in {H3K4me1, H3K4me3, H3K9ac, H3K14ac, H4K12ac} for each post translational histone modification and form is in {wp, fuzzy, wpfuzzy} considering well positionned nucleosomes, fuzzy nucleosomes or both for SNEP computation.

```
chr_BY lower_bound_BY upper_bound_BY index_nuc_BY chr_RM lower_bound_RM upper_bound_RM in-
dex_nuc_RM roi_index form BY_Mnase_Seq_1 BY_Mnase_Seq_2 BY_Mnase_Seq_3 RM_Mnase_Seq_4
RM_Mnase_Seq_5 RM_Mnase_Seq_6 BY_H3K14ac_36 BY_H3K14ac_37 BY_H3K14ac_53 RM_H3K14ac_38
RM_H3K14ac_39 pvalsGLM
```

For each file, there is 1 line per nucleosome and each line is composed of many columns divided into 3 main topics:

- nuc information
- number opf reads for each sample
- DESeq analysis results.

For example for the file *BY_RM_H3K14ac_wp_snep.tab* informations are:

- chr_BY, the BY chr involved
- lower_bound_BY, the lower bound of the BY nuc
- upper_bound_BY, the upper_bound of the BY nuc
- index_nuc_BY, the index of the nuc in the entire list of BY nucs
- **chr_RM, lower_bound_RM, upper_bound_RM, index_nuc_RM** are the same information for the RM strain
- roi_index, the index of the region of interest involved.

Next cols concern indicators for each sample. They are labeled [strain]_[marker]_[sample_id] and each value represents the number of reads for the current nuc for the sample *sample_id*.

The 5 final columns concern DESeq analysis:

- manip[a_manip] strain[a_strain] manip[a_strain]:strain[a_strain], the manip (marker) effect, the strain effect and the snep effect.
- pvalsGLM, the pvalue resulting of the comparison of the GLM model considering or the interaction term *marker:strain*
- snep_index, a boolean set to TRUE if the *pvalueGLM* value is under the threshold computed with FDR function with a rate set to 0.01%.

It also produces the file that explicts size factor for each involved sample in differents strain combination and nucleosomal region type:

TODO: include this file... /home/filleton/analyses/snepcatalog/data/2013-10-09/current/README.txt

results/current/size_factors.tab

2.4.2 Number of SNEPs

Here are the number of computed for each forms.

```
[1] "wp"
      #nucs H3K4me1 H3K4me3 H3K9ac H3K14ac H4K12ac
BY-RM  30234     520     798     83    3566     26
BY-YJM 31298     303     619    102     103    128
RM-YJM 29863     129     340     46    3177     18
[1] "fuzzy"
      #nucs H3K4me1 H3K4me3 H3K9ac H3K14ac H4K12ac
BY-RM  10748     294     308    101    1681     42
BY-YJM 10669     122     176    124     93     87
RM-YJM 11478      54     112     41    1389     20
[1] "wpfuzzy"
      #nucs H3K4me1 H3K4me3 H3K9ac H3K14ac H4K12ac
BY-RM  40982     770    1136    183    5404     73
```

BY-YJM	41967	439	804	214	198	199
RM-YJM	41341	184	468	87	4687	37

TODO:

- Print/study intra/inter strain LODs.
- Check the normality of sample using Shapiro–Wilk (Hypothesis for computing LODs)

REFERENCES

3.1 Python Reference

3.2 R Reference

3.2.1 Arabic to Roman pair list.

Description

Util to convert Arabicto Roman

Usage

```
ARAB2ROM ()
```

Author(s)

Florent Chuffart

R: False Discovery Rate

3.2.2 False Discovery Rate

Description

From a vector x of independent p-values, extract the cutoff corresponding to the specified FDR. See Benjamini & Hochberg 1995 paper

Usage

```
FDR (x, FDR)
```

Arguments

x

A vector x of independent p-values.

FDR

The specified FDR.

Value

Return the the corresponding cutoff.

Author(s)

Gael Yvert, Florent Chuffart

Examples

```
print("example")
```

R: Roman to Arabic pair list.

3.2.3 Roman to Arabic pair list.

Description

Util to convert Roman to Arabic

Usage

```
ROM2ARAB()
```

Author(s)

Florent Chuffart

R: Aggregate replicated sample's nucleosomes.

3.2.4 Aggregate replicated sample's nucleosomes.

Description

This function aggregates nucleosome for replicated samples. It uses TemplateFilter output of each sample as replicate. Each sample owns a set of nucleosomes computed using TemplateFilter and ordered by the position of their center. Adjacent nucleosomes are compared two by two. Comparison is based on a log likelihood ratio score. The issue of comparison is adjacent nucleosomes merge or separation. Finally the function returns a list of clusters and all computed *llr_scores*. Each cluster owns an attribute *wp* for “well positionned”. This attribute is set as *TRUE* if the cluster is composed of exactly one nucleosomes of each sample.

Usage

```
aggregate_intra_strain_nucs(samples, llr_thres = 20, coord_max = 2e+07)
```

Arguments

`samples`

A list of samples. Each sample is a list like `sample = list(id=..., marker=..., strain=..., roi=..., inputs=..., outputs=...)` with `roi = list(name=..., begin=..., end=..., chr=..., genome=...)`.

`llr_thres`

Log likelihood ration threshold.

`coord_max`

A too big value to be a coord for a nucleosome lower bound.

Value

Returns a list of clusterized nucleosomes, and all computed llr scores.

Author(s)

Florent Chuffart

Examples

```
# Dealing with a region of interest
roi = list(name="example", begin=1000, end=1300, chr="1", genome=rep("A",301))
samples = list()
for (i in 1:3) {
  # Create TF output
  tf_nuc = list("chr"=paste("chr", roi$chr, sep=""), "center"=(roi$end + roi$begin)/2, "width"= 150)
  outputs = dfadd(NULL,tf_nuc)
  outputs = filter_tf_outputs(outputs, roi$chr, roi$begin, roi$end)
  # Generate corresponding reads
  nb_reads = round(runif(1,170,230))
  reads = round(rnorm(nb_reads, tf_nuc$center,20))
  u_reads = sort(unique(reads))
  strands = sample(c(rep("R",ceiling(length(u_reads)/2)),rep("F",floor(length(u_reads)/2))))
  counts = apply(t(u_reads), 2, function(r) { sum(reads == r)})
  shifts = apply(t(strands), 2, function(s) { if (s == "F") return(-tf_nuc$width/2) else return(tf_nuc$width/2)})
  u_reads = u_reads + shifts
  inputs = data.frame(list("V1" = rep(roi$chr, length(u_reads)),
                           "V2" = u_reads,
                           "V3" = strands,
                           "V4" = counts), stringsAsFactors=FALSE)
  samples[[length(samples) + 1]] = list(id=1, marker="Mnase_Seq", strain="strain_ex", total_reads = nrow(inputs))
}
print(aggregate_intra_strain_nucs(samples))
```

R: Aligns nucleosomes between 2 strains.

3.2.5 Aligns nucleosomes between 2 strains.

Description

This function aligns nucs between two strains for a given genome region.

Usage

```
align_inter_strain_nucs(replicates, wp_nucs_strain_ref1 = NULL,  
                        wp_nucs_strain_ref2 = NULL, corr_thres = 0.5, llr_thres = 100,  
                        config = NULL, ...)
```

Arguments

`replicates`

Set of replicates, ideally 3 per strain.

`wp_nucs_strain_ref1`

List of aggregates nucleosome for strain 1. If it's null this list will be computed.

`wp_nucs_strain_ref2`

List of aggregates nucleosome for strain 2. If it's null this list will be computed.

`corr_thres`

Correlation threshold.

`llr_thres`

LOD cut off.

`config`

GLOBAL config variable

`...`

A list of parameters that will be passed to *aggregate_intra_strain_nucs* if needed.

Value

Returns a list of clusterized nucleosomes, and all computed llr scores.

Author(s)

Florent Chuffart

Examples

```

# Define new translate_cur function...
translate_cur = function(roi, strain2, big_cur=NULL, config=NULL) {
  return(roi)
}
# Binding it by uncomment following lines.
unlockBinding("translate_cur", as.environment("package:nucleominer"))
unlockBinding("translate_cur", getNamespace("nucleominer"))
assign("translate_cur", translate_cur, "package:nucleominer")
assign("translate_cur", translate_cur, getNamespace("nucleominer"))
lockBinding("translate_cur", getNamespace("nucleominer"))
lockBinding("translate_cur", as.environment("package:nucleominer"))

# Dealing with a region of interest
roi = list(name="example", begin=1000, end=1300, chr="1", genome=rep("A",301), strain_ref1 = "STRAIN1")
roi2 = translate_cur(roi, roi$strain_ref1)
replicates = list()
for (j in 1:2) {
  samples = list()
  for (i in 1:3) {
    # Create TF output
    tf_nuc = list("chr"=paste("chr", roi$chr, sep=""), "center"=(roi$end + roi$begin)/2, "width"=roi$end-roi$begin)
    outputs = dfadd(NULL,tf_nuc)
    outputs = filter_tf_outputs(outputs, roi$chr, roi$begin, roi$end)
    # Generate corresponding reads
    nb_reads = round(runif(1,170,230))
    reads = round(rnorm(nb_reads, tf_nuc$center,20))
    u_reads = sort(unique(reads))
    strands = sample(c(rep("R",ceiling(length(u_reads)/2)),rep("F",floor(length(u_reads)/2))))
    counts = apply(t(u_reads), 2, function(r) { sum(reads == r)})
    shifts = apply(t(strands), 2, function(s) { if (s == "F") return(-tf_nuc$width/2) else return(0)})
    u_reads = u_reads + shifts
    inputs = data.frame(list("V1" = rep(roi$chr, length(u_reads)),
                           "V2" = u_reads,
                           "V3" = strands,
                           "V4" = counts), stringsAsFactors=FALSE)
    samples[[length(samples) + 1]] = list(id=1, marker="Mnase_Seq", strain=paste("strain_ex",j,sep="_"))
  }
  replicates[[length(replicates) + 1]] = samples
}
print(align_inter_strain_nucs(replicates))

```

R: Launch deseq methods.

3.2.6 Launch deseq methods.

Description

This function is based on deseq example. It normalizes data, fit data to GLM model with and without interaction term and compare the two l;=models.

Usage

```
analyse_design(snep_design, reads)
```

Arguments

sneq_design

The design to considere.

reads

The data to considere.

Author(s)

Florent Chuffart

R: Stage replicates data

3.2.7 Stage replicates data

Description

This function loads in memory data corresponding to the given experiments.

Usage

```
build_replicates(expe, roi, only_fetch = FALSE, get_genome = FALSE,  
                 all_samples, config = NULL)
```

Arguments

expe

a list of vector corresponding to vector of replicates.

roi

the region that we are interested in.

only_fetch

filter or not inputs.

get_genome

Load or not corresponding genome.

all_samples

Global list of samples.

config

GLOBAL config variable.

Author(s)

Florent Chuffart

Examples

```
# library(rjson)
# library(nucleominer)
#
# # Read config file
# json_conf_file = "nucleo_miner_config.json"
# config = fromJSON(paste(readLines(json_conf_file), collapse=""))
# # Read sample file
# all_samples = get_content(config$CSV_SAMPLE_FILE, "cvs", sep=";", head=TRUE, stringsAsFactors=FALSE)
# # here are the sample ids in a list
# expes = list(c(1))
# # here is the region that we want to see the coverage
# cur = list(chr="8", begin=472000, end=474000, strain_ref="BY")
# # it displays the coverage
# replicates = build_replicates(expes, cur, all_samples=all_samples, config=config)
# out = watch_samples(replicates, config$READ_LENGTH,
#   plot_coverage = TRUE,
#   plot_squared_reads = FALSE,
#   plot_ref_genome = FALSE,
#   plot_arrow_raw_reads = FALSE,
#   plot_arrow_nuc_reads = FALSE,
#   plot_gaussian_reads = FALSE,
#   plot_gaussian_unified_reads = FALSE,
#   plot_ellipse_nucs = FALSE,
#   plot_wp_nucs = FALSE,
#   plot_wp_nuc_model = FALSE,
#   plot_common_nucs = FALSE,
#   height = 50)
```

R: Extract a sub part of the corresponding c2c file

3.2.8 Extract a sub part of the corresponding c2c file

Description

This function allows access to a specific part of the c2c file.

Usage

```
c2c_extraction(strain1, strain2, chr = NULL, lower_bound = NULL,
  upper_bound = NULL, config = NULL)
```

Arguments

strain1

the key strain

strain2

the target strain

chr

if defined, the c2c will be filtered according to the chromosome value

`lower_bound`

if defined, the c2c will filtered for part of the genome upper than `lower_bound`

`upper_bound`

if defined, the c2c will filtered for part of the genome lower than `upper_bound`

`config`

GLOBAL config variable

Author(s)

Florent Chuffart

R: reformat an “apply manipulated” list of regions

3.2.9 reformat an “apply manipulated” list of regions

Description

Utils to reformat an “apply manipulated” list of regions

Usage

```
collapse_regions(regions)
```

Arguments

regions	
---------	--

Author(s)

Florent Chuffart

R: Compute Common Uninterrupted Regions (CUR)

3.2.10 Compute Common Uninterrupted Regions (CUR)

Description

CURs are regions that can be aligned between the genomes

Usage

```
compute_inter_all_strain_curs(diff_allowed = 30, min_cur_width = 4000,  
                              config = NULL)
```

Arguments

`diff_allowed`

the maximum indel width allowed in a CUR

`min_cur_width`

The minimum width of a CUR

`config`

GLOBAL config variable

Author(s)

Florent Chuffart

R: Crop bound of regions according to region of interest bound

3.2.11 Crop bound of regions according to region of interest bound

Description

The function is no more necessary since we remove “big_cur” bug in `translate_cur` function.

Usage

```
crop_fuzzy(tmp_fuzzy_nucs, roi, strain, config = NULL)
```

Arguments

`tmp_fuzzy_nucs`

the regions to be cropped.

`roi`

The region of interest.

`strain`

The strain to consider.

`config`

GLOBAL config variable

Author(s)

Florent Chuffart

R: Adding list to a dataframe.

3.2.12 Adding list to a dataframe.

Description

Add a list *l* to a dataframe *df*. Create it if *df* is *NULL*. Return the dataframe *df*.

Usage

```
dfadd(df, l)
```

Arguments

df

A dataframe

l

A list

Value

Return the dataframe *df*.

Author(s)

Florent Chuffart

Examples

```
## Here dataframe is NULL
print(df)
df = NULL

# Initialize df
df = dfadd(df, list(key1 = "value1", key2 = "value2"))
print(df)

# Adding elements to df
df = dfadd(df, list(key1 = "value1'", key2 = "value2'"))
print(df)
```

R: Prefetch data

3.2.13 Prefetch data

Description

Fetch and filter inputs and outputs per region of interest. Organize it per replicates.

Usage

```
fetch_mnase_replicates(strain, roi, all_samples, config = NULL,  
    only_fetch = FALSE, get_genome = FALSE, get_ouputs = TRUE)
```

Arguments

strain

The strain we want mnase replicatesList of replicates. Each replicates is a vector of sample ids.

roi

Region of interest.

all_samples

Global list of samples.

config

GLOBAL config variable

only_fetch

If TRUE, only fetch and not filtering. It is used tio load sample files into memory before forking.

get_genome

If TRUE, load corresponding genome sequence.

get_ouputs

If TRUE, get also ouput corresponding TF output files.

Author(s)

Florent Chuffart

R: Filter TemplateFilter inputs

3.2.14 Filter TemplateFilter inputs

Description

This function filters TemplateFilter inputs according genome area observed properties. It takes into account reads that are at the frontier of this area and the strand of these reads.

Usage

```
filter_tf_inputs(inputs, chr, x_min, x_max, nuc_width = 160,  
    only_f = FALSE, only_r = FALSE, filter_for_coverage = FALSE)
```

Arguments

`inputs`

TF inputs to be filtered.

`chr`

Chromosome observed, here chr is an integer.

`x_min`

Coordinate of the first bp observed.

`x_max`

Coordinate of the last bp observed.

`nuc_width`

Nucleosome width.

`only_f`

Filter only F reads.

`only_r`

Filter only R reads.

`filter_for_coverage`

Does it filter for plot coverage?

Value

Returns filtered inputs.

Author(s)

Florent Chuffart

R: Filter TemplateFilter outputs

3.2.15 Filter TemplateFilter outputs

Description

This function filters TemplateFilter outputs according, not only genome area observed properties, but also correlation and overlap threshold.

Usage

```
filter_tf_outputs(tf_outputs, chr, x_min, x_max, nuc_width = 160,  
                  ol_bp = 59, corr_thres = 0.5)
```

Arguments

`tf_outputs`

TemplateFilter outputs.

`chr`

Chromosome observed, here chr is an integer.

`x_min`

Coordinate of the first bp observed.

`x_max`

Coordinate of the last bp observed.

`nuc_width`

Nucleosome width.

`ol_bp`

Overlap Threshold.

`corr_thres`

Correlation threshold.

Value

Returns filtered TemplateFilter Outputs

Author(s)

Florent Chuffart

R: to flat aggregate_intra_strain_nucs function output

3.2.16 to flat aggregate_intra_strain_nucs function output

Description

This function builds a dataframe of all clusters obtain from aggregate_intra_strain_nucs function.

Usage

```
flat_aggregated_intra_strain_nucs(partial_strain_maps, cur_index)
```

Arguments

`partial_strain_maps`

the output of aggregate_intra_strain_nucs function

`cur_index`

the index of the roi involved

Value

Returns a dataframe of all clusters obtain from `aggregate_intra_strain_nucs` function.

Author(s)

Florent Chuffart

R: flat reads

3.2.17 flat reads

Description

Extract reads coordinates from `TemplateFilter` input sequence

Usage

```
flat_reads(reads, nuc_width)
```

Arguments

`reads`

`TemplateFilter` input reads

`nuc_width`

Width used to shift F and R reads.

Value

Returns a list of F reads, R reads and joint/shifted F and R reads.

Author(s)

Florent Chuffart

R: Retrieve Reads

3.2.18 Retrieve Reads

Description

Retrieve reads for a given marker, combi, form.

Usage

```
get_all_reads(marker, combi, form = "wp", config = NULL)
```

Arguments

marker

The marker to considere.

combi

The starin combination to considere.

form

The nuc form to considere.

config

GLOBAL config variable

Author(s)

Florent Chuffart

R: get comp strand

3.2.19 get comp strand

Description

Compute the complementatry strand.

Usage

```
get_comp_strand(strand)
```

Arguments

strand

The original strand.

Value

Returns the complementatry strand.

Author(s)

Florent Chuffart

R: Build the design for deseq

3.2.20 Build the design for deseq

Description

This function build the design according sample properties.

Usage

```
get_design(marker, combi, all_samples)
```

Arguments

`marker`

The marker to considere.

`combi`

The starin combination to considere.

`all_samples`

Global list of samples.

Author(s)

Florent Chuffart

R: Compute the fuzzy list for a given strain.

3.2.21 Compute the fuzzy list for a given strain.

Description

This function grabs the nucleosomes detxted by `template_filter` that have been rejected bt `aggregate_intra_strain_nucs` as well positions.

Usage

```
get_intra_strain_fuzzy(wp_map, roi, strain, config = NULL)
```

Arguments

`wp_map`

Well positionned nucleosomes map.

`roi`

The region of interest.

`strain`

The strain we want to extracvt the fuzzy map.

config

GLOBAL config variable.

Author(s)

Florent Chuffart

R: Compute the list of SNEPs for a given set of marker, strain...

3.2.22 Compute the list of SNEPs for a given set of marker, strain combination and nuc form.

Description

This function uses

Usage

```
get_sneps(marker, combi, form, all_samples, config = NULL)
```

Arguments

marker

The marker involved.

combi

The strain combination involved.

form

the nuc form involved.

all_samples

Global list of samples.

config

GLOBAL config variable

Author(s)

Florent Chuffart

Examples

```
marker = "H3K4me1"
combi = c("BY", "YJM")
form = "wpunr" # "wp" | "unr" | "wpunr"
# foo = get_sneps(marker, combi, form)
# foo = get_sneps("H4K12ac", c("BY", "RM"), "wp")
```

R: Compute the unaligned nucleosomal regions (UNRs).

3.2.23 Compute the unaligned nucleosomal regions (UNRs).

Description

This function aggregate non common wp nucs for each strain and subtract common wp nucs. It does not take care about the size of the resulting UNR. It will be take into account in the count read part of the pipeline.

Usage

```
get_unrs(combi, roi, cur_index, wp_maps, fuzzy_maps, common_nuc_results,  
         config = NULL)
```

Arguments

`combi`

The strain combination to consider.

`roi`

The region of interest.

`cur_index`

The region of interest index.

`wp_maps`

Well positionned nucleosomes maps.

`fuzzy_maps`

Fuzzy nucleosomes maps.

`common_nuc_results`

Common wp nuc maps

`config`

GLOBAL config variable

Author(s)

Florent Chuffart

R: Returns the intersection of 2 list on regions.

3.2.24 Returns the intersection of 2 list on regions.

Description

This function...

Usage

```
intersect_region(region1, region2)
```

Arguments

region1

Original regions.

region2

Regions to intersect.

Author(s)

Florent Chuffart

R: Likelihood ratio

3.2.25 Likelihood ratio

Description

Compute the log likelihood ratio of two or more set of value.

Usage

```
llr_score_nvecs(xs)
```

Arguments

xs

list of vectors.

Value

Returns the log likelihood ratio.

Author(s)

Florent Chuffart

Examples

```
# LOD score for 2 set of values
mean1=5; sd1=2; card2 = 250
mean2=6; sd2=3; card1 = 200
x1 = rnorm(card1, mean1, sd1)
x2 = rnorm(card2, mean2, sd2)
min = floor(min(c(x1,x2)))
max = ceiling(max(c(x1,x2)))
hist(c(x1,x2), xlim=c(min, max), breaks=min:max)
lines(min:max,dnorm(min:max,mean1,sd1)*card1,col=2)
lines(min:max,dnorm(min:max,mean2,sd2)*card2,col=3)
lines(min:max,dnorm(min:max,mean(c(x1,x2)),sd(c(x1,x2)))*card2,col=4)
llr_score_nvecs(list(x1,x2))
```

R: nm

3.2.26 nm

Description

It provides a set of useful functions allowing to perform quantitative analysis of nucleosomal epigenome.

Details

Package:	nucleominer
Maintainer:	Florent Chuffart < florent.chuffart@ens-lyon.fr >
Author:	Florent Chuffart
Version:	2.3.42
License:	CeCILL
Title:	nm
Depends:	seqinr, plotrix, DESeq, cachecache

Author(s)

Florent Chuffart

R: Plot the distribution of reads.

3.2.27 Plot the distribution of reads.

Description

This fuxntion use the deseq nomalization feature to compare qualitatively the distribution.

Usage

```
plot_dist_samples(strain, marker, res, all_samples, NEWPLOT = TRUE)
```

Arguments

`strain`

The strain to considere.

`marker`

The marker to considere.

`res`

Data

`all_samples`

Global list of samples.

`NEWPLOT`

If FALSE the curve will be add to the current plot.

Author(s)

Florent Chuffart

R: sign from strand

3.2.28 sign from strand

Description

Get the sign of strand

Usage

```
sign_from_strand(strands)
```

Arguments

<code>strands</code>	
----------------------	--

Value

If strand in forward then returns 1 else returns -1

Author(s)

Florent Chuffart

R: Substract to a list of regions an other list of regions that...

3.2.29 Subtract to a list of regions an other list of regions that intersect it.

Description

This function embed a recursive part. It occurs when a subtracted region split an original region on two.

Usage

```
subtract_region(region1, region2)
```

Arguments

region1

Original regions.

region2

Regions to subtract.

Author(s)

Florent Chuffart

R: Switch a pairlist

3.2.30 Switch a pairlist

Description

Take a pairlist key:value and return the switched pairlist value:key.

Usage

```
switch_pairlist(l)
```

Arguments

l

The pairlist to switch.

Value

The switched pairlist.

Author(s)

Florent Chuffart

Examples

```
l = list(key1 = "value1", key2 = "value2")
print(switch_pairlist(l))
```

R: Translate coords of a genome region.

3.2.31 Translate coords of a genome region.

Description

This function is used in the examples, usually you have to define your own translation function and overwrite this one using *unlockBinding* features. Please, refer to the example.

Usage

```
translate_cur(roi, strain2, config = NULL, big_cur = NULL)
```

Arguments

roi

Original genome region of interest.

strain2

The strain in wich you want the genome region of interest.

config

GLOBAL config variable

big_cur

A largest region than roi use to filter c2c if it is needed.

Author(s)

Florent Chuffart

Examples

```
# Define new translate_cur function...
translate_cur = function(roi, strain2, config) {
  strain1 = roi$strain_ref
  if (strain1 == strain2) {
    return(roi)
  } else {
    stop("Here is my new translate_cur function...")
  }
}

# Binding it by uncomment follwing lines.
# unlockBinding("translate_cur", as.environment("package:nm"))
# unlockBinding("translate_cur", getNamespace("nm"))
```

```
# assign("translate_cur", translate_cur, "package:nm")
# assign("translate_cur", translate_cur, getNamespace("nm"))
# lockBinding("translate_cur", getNamespace("nm"))
# lockBinding("translate_cur", as.environment("package:nm"))
```

R: Translate a list of regions from a strain ref to another.

3.2.32 Translate a list of regions from a strain ref to another.

Description

This function is an elaborated call to `translate_cur`.

Usage

```
translate_regions(regions, combi, cur_index, config = NULL, roi)
```

Arguments

`regions`

Regions to be translated.

`combi`

Combination of strains.

`cur_index`

The region of interest index.

`config`

GLOBAL config variable

`roi`

The region of interest.

Author(s)

Florent Chuffart

R: Aggregate regions that intersect themselves.

3.2.33 Aggregate regions that intersect themselves.

Description

This function is based on sort of lower bounds to detect regions that intersect. We compare lower bound and upper bound of the porevious item. This function embed a while loop and break break regions list become stable.

Usage

```
union_regions(regions)
```

Arguments

regions

The Regions to be aggregated

Author(s)

Florent Chuffart

R: Watching analysis of samples

3.2.34 Watching analysis of samples

Description

This function allows to view analysis for a particular region of the genome.

Usage

```
watch_samples(replicates, read_length, plot_ref_genome = TRUE,  
  plot_arrow_raw_reads = TRUE, plot_arrow_nuc_reads = TRUE,  
  plot_squared_reads = TRUE, plot_coverage = FALSE, plot_gaussian_reads = TRUE,  
  plot_gaussian_unified_reads = TRUE, plot_ellipse_nucs = TRUE,  
  change_col = TRUE, plot_wp_nucs = TRUE, plot_fuzzy_nucs = TRUE,  
  plot_wp_nuc_model = TRUE, plot_common_nucs = FALSE, plot_common_unrs = FALSE,  
  plot_wp_nucs_4_nonmnase = FALSE, plot_chain = FALSE, plot_sample_id = FALSE,  
  aggregated_intra_strain_nucs = NULL, aligned_inter_strain_nucs = NULL,  
  height = 10, main = NULL, xlab = NULL, ylab = "#reads (per million reads)",  
  config = NULL)
```

Arguments

replicates

replicates under the form...

read_length

length of the reads

plot_ref_genome

Plot (or not) reference genome.

plot_arrow_raw_reads

Plot (or not) arrows for raw reads.

plot_arrow_nuc_reads

Plot (or not) arrows for reads associated to a nucleosome.

plot_squared_reads

Plot (or not) reads in the square fashion.

plot_coverage

Plot (or not) reads in the coverage fashion. fashion.

plot_gaussian_reads

Plot (or not) gaussian model of a F and R reads.

plot_gaussian_unified_reads

Plot (or not) gaussian model of a nuc.

plot_ellipse_nucs

Plot (or not) ellipse for a nuc.

change_col

Change the color of each nucleosome.

plot_wp_nucs

Plot (or not) cluster of nucs

plot_fuzzy_nucs

Plot (or not) cluster of fuzzy

plot_wp_nuc_model

Plot (or not) gaussian model for a cluster of nucs

plot_common_nucs

Plot (or not) aligned reads.

plot_common_unrs

Plot (or not) unaligned nucleosomal regions (UNRs).

plot_wp_nucs_4_nonmnase

Plot (or not) clusters for non inputs samples.

plot_chain

Plot (or not) clusterised nucleosomes between mnase samples.

plot_sample_id

Plot (or not) the sample id for each sample.

aggregated_intra_strain_nucs

list of aggregated intra strain nucs. If NULL, it will be computed.

aligned_inter_strain_nucs

list of aligned inter strain nucs. If NULL, it will be computed.

height

Number of reads in per million read for each sample, graphical parametre for the y axis.

main

main title of the produced plot

xlab

xlab of the produced plot

ylab

ylab of the produced plot

config

GLOBAL config variable

Author(s)

Florent Chuffart

INDICES AND TABLES

- *genindex*
- *search*