Please see https://github.com/theodorejperkins/RECAP (https://github.com/theodorejperkins/RECAP) for the latest stable build.
Please see https://github.com/jchitpin/RECAP (https://github.com/jchitpin/RECAP) for other iterations of the code. Last edited 16-January-2019

Introduction

ChIP-seq is used extensively to identify sites of transcription factor binding or regions of epigenetic modifications to the genome. The fundamental bioinformatics problem is to take ChIP-se read data and data representing some kind of control, and determine genomic regions that are enriched in the ChIP-se versus the control, also called "peak calling." While many programs have been designed to solve this task, nearly all fall into the statistical trap of using the data twice--once to determine candidate enriched regions, and a second time to assess enrichment by methods of classical statistical hypothesis testing. This double use of the data has the potential to invalidate the statistical significance assigned to enriched regions, or "peaks", and as a consequence, to invalidate false discovery rate estimates. Thus, the true significance or reliability of peak calls remains unknown. We propose a new wrapper algorithm, RECAP, that uses resampling of ChIP-seq and control data to estimate and correct for biases built into peak calling algorithms. RECAP is a powerful new tool for assessing the true statistical significance of ChIP-seq peak calls.

Installation

Download RECAP and extract to your desired directory. There should be five scripts:

1. RECAP_Re-Mix.sh
2. RECAP.pl
3. RECAP_MACS.sh
4. RECAP_SICER.sh
5. RECAP_diffReps.sh

The first two scripts should be runnable on any system with Bash and Perl installed. Several CPAN modules are required. To install them:

- cpan
- install List::BinarySearch
- install List::Util
- install List::MoreUtils
- install Math::Utils

Wrapper Script Usage

Argument Argument	Description Description
-i, --input	Input treatment/control BED file directory (absolute path)
-t, --treatment	Treatment BED file
-c, --control	Control BED file
-o, --output	Output file directory (absolute path and must exist)
-b, --bootstrap	Number of re-mixes
-e, --header	Header number of peak calling output files
-h, --help	Display this help and exit

From our simulated and ENCODE tests, we found RECAP-recalibrated MACS to yield the best peak calling results for ChIP-vs-Control data sets targeting sharp peaks from transcription factors. For broad peaks corresponding to histone modification, we would instead recommend using the SICER wrapper script or running MACS in broad mode. For differential ChIP-seq analysis between biological replicates, we would recommend using our diffeps wrapper script, although RECAP shows a minimal effect on p-value correction. If these three peak callers are unsuitable for your analysis, consider adapting the wrapper scripts based on the usage of `RECAP.pl` described further below.

MACS

Suppose we are interested in analyzing a treatment and control file with MACS.

1. Open RECAP_MACS.sh and modify any peak-calling preferences in line 101 and 109 ***except*** the p-value threshold. Possible options are listed on the MACS Github page https://github.com/taoliu/MACS (https://github.com/taoliu/MACS) . Currently, the recommended default MACS settings are used for regular peak calling on an hg18-sized genome.

2. Run the script with the arguments below. Absolute directory paths must be specified and the output directory must already exist. A bootstrap of 1 is recommended and the header should be set to 29 (28 if using MACS --nomodel parameter).
   ```
   bash RECAP_MACS.sh -i ~/ -t treatment_file.bed -c control_file.bed -o
   ~/output_directory -b 1 -e 29
   ```

3. Check the output directory to find the re-mixed bed files in `re-mix`, original peak calling output file in `MACS_original`, re-mixed peak calling output files in `MACS_re-mix`, and the final RECAP-recalibrated output file in `MACS_RECAP`.

SICER

Extra Arguments	Description
-w, --window	Window size
-g, --gap	Gap size

Suppose we are interested in analyzing a treatment and control file with SICER.

1. Open RECAP_MACS.sh and modify any peak-calling preferences in line 104 and 114 *except* the p-value threshold. Possible options are listed in the SICER README.pdf. Currently, the recommended default SICER settings are used for regular peak calling on an hg38-sized genome.

2. Run the script with the arguments below. Absolute directory paths must be specified and the output directory must already exist. A bootstrap of 1 is recommended and the header should be set to 0. The window and gap size must be specified here because they are used to name the resulting `*-islands-summary` files. Window = 100 and gap = 200 are the recommended default SICER settings for regular peak calling.
   ```
   bash RECAP_SICER.sh -i ~/ -t treatment_file.bed -c control_file.bed -o ~/output_directory -b 1 -e 0 -w 100 -g 200
   ```

3. Check the output directory to find the re-mixed bed files in `re-mix`, original peak calling output file in `SICER_original`, re-mixed peak calling output files in `SICER_re-mix`, and the final RECAP-recalibrated output file in `SICER_RECAP`.

diffReps

Suppose we are interested in analyzing a treatment and control file with diffReps.

1. Open RECAP_diffReps.sh and modify any peak-calling preferences in line 104 and 113 *except* the p-value threshold. Possible options are listed using `diffReps.pl --help`. Currently, the recommended default diffReps settings are used for regular peak calling with an hg19-sized genome.

2. Run the script with the arguments below. Absolute directory paths must be specified and the output directory must already exist. A bootstrap of 1 is recommended and the header should be set to 33.
   ```
   bash RECAP_diffReps.sh -i ~/ -t treatment_file.bed -c control_file.bed -o ~/output_directory -b 1 -e 33
   ```

3. Check the output directory to find the re-mixed bed files in `re-mix`, original peak calling output file in `diffReps_original`, re-mixed peak calling output files in `diffReps_re-mix`, and the final RECAP-recalibrated output file in `diffReps_RECAP`.

A different peak caller

Any peak caller can work with `RECAP.pl` so long as it uses a p-value cut-off to identify significant peaks. One of the wrapper scripts above would have to be altered by replacing the

peak caller and parameters with your desired one. The number of header lines in the output file would have to be changed too. The RECAP.pl arguments `software` and `delim` would be set to `0` for other and either `t` for tab or `c` for comma. Be sure to only keep the re-mixed summary files containing the list of picked peaks. `RECAP.pl` won't be able to pick up the right re-mixed summary file otherwise, especially if bootstrap is greater than `1`.

RECAP_Re-Mix.sh Usage

```
bash RECAP_Re-Mix.sh [-i] [-t] [-c] [-o] [-m] [-b] [-h]
```

Argument	Description
-i, --input	Input treatment/control BED file directory (absolute path)
-t, --treatment	Treatment BED file
-c, --control	Control BED file
-o, --output	Output file directory (absolute path and must exist)
-m, --method	Method of re-mixing*
-b, --bootstrap	Number of re-mixes*
-h, --help	Display this help and exit

Options(*)

-m, --method
Choose either ***equal*** or ***unequal***. ***Equal*** distributes the treatment and control file reads into two equally sized re-mixed BED files. ***Unequal*** distributes the treatment and control file reads into two re-mixed BED files with the same read counts as the input files. ***Unequal*** is the recommended parameter.

-b, --bootstrap
Bootstrap is the number of times the treatment and control files are re-mixed to generate re-mixed BED files.

RECAP.pl Usage

```
perl RECAP.pl [--dirOrig]   [--nameOrig]   [--dirRemix] [--nameRemix]
              [--dirOutput] [--nameOutput] [bootstrap]  [--header]
              [--pvalCol]   [--delim]      [--software] [--help]
```

Argument	Description

Argument	Description
--dirOrig	Input original peak calling summary file directory (absolute path)
--nameOrig	Original peak calling summary file
--dirRemix	Input re-mixed peak calling summary file directory (absolute path)
--nameRemix	Re-mixed peak calling summary file ending in '.bootstrap_#.bed
--dirOutput	Output directory (absolute path and must exist)
--nameOutput	Original peak calling summary file with RECAP
--bootstrap	Number of re-mixing procedures*
--header	Number of header lines in peak calling summary file
--pvalCol	Column number containing *p*-values in summary file
--delim	Delimiter type*
--software	Type of peak caller used*
-- help	Display this help and exit

Options(*)

--bootstrap Ensure that the re-mixed peak calling file ends in '.boostrap_#.bed. Replace '#' with the bootstrap number.

--delim Choose either *(c)omma* or *(t)ab* delimiters depending on the output of your peak caller.

--software Choose either *(M)ACS* for MACS2, or *(D)iffReps* for diffReps, or *(O)ther* for another type of peak caller. Choosing *M* negative antilogs the *p*-values, a necessary step during *p*-value recalibration with RECAP. Choosing 'D' filters off any downregulated p-values, a special feature of diffReps that must be removed during *p*-value recalibration with RECAP.

Notes

The re-mixing process takes minutes to perform. Recalibrating the *p*-values with the Perl script should take seconds to minutes.

Example (Manual) Workflow

The code below can be adapted for any peak caller that produces *p*-values and is similar to the codes in the MACS/SICER/diffReps wrapper scripts. If you'd like to create your own wrapper

script for another peak caller, please adapt the following:

1. Re-mix treatment and control BED files:

   ```
   bash RECAP_Re-Mix.sh -i ~/ChIP-Seq/files -t Treatment.bed -c Control.bed -o
   ~/ChIP-Seq/files -m unequal -b 1
   ```

   This will create a new directory `~/ChIP-Seq/files/re-mix` with files `Treatment.bootstrap_1.bed` and `Control.bootstrap_1.bed`

2. Analyze "original" files with MACS (please use *p*=0.1 for MACS and *p*=1 for SICER/diffReps):

   ```
   cd ~/ChIP-Seq/files macs2 callpeak -t Treatment.bed -c Control.bed -p 0.1 -n
   Analysis_Original --outdir ~/ChIP-Seq/analysis/
   ```

   This will create several files including `Analysis_Original_peaks.xls`.

3. Analyze "re-mixed" files with MACS (please use *p*=0.1 for MACS and *p*=1 for SICER/diffReps):

   ```
   cd ~/ChIP-Seq/files/re-mix macs2 callpeak -t Treatment.bootstrap_1.bed -c
   Control -p 0.1 -n Analysis_Remixed_bootstrap_1 --outdir ~/ChIP-Seq/analysis/
   ```

   This will create several files including `Analysis_Remixed_bootstrap_1_peaks.xls`. Please retain only the `_peaks.xls` file which is to be recalibrated.

   ```
   cd ~/ChIP-Seq/analysis
   find . -type f -name 'Analysis_Remixed*_peaks.narrowPeak' -delete
   find . -type f -name 'Analysis_Remixed*_model.r' -delete
   find . -type f -name 'Analysis_Remixed*_summits.bed' -delete
   ```

4. Recalibrate the *p*-values:

   ```
   perl RECAP.pl --dirOrig ~/ChIP-Seq/analysis/ --nameOrig
   Analysis_Original_peaks.xls --dirRemix ~/ChIP-Seq/analysis --nameRemix
   Analysis_Remixed --dirOutput ~/ChIP-Seq/analysis/ --nameOutput
   Treatment.RECAP.bootstrap_1.txt --bootstrap 1 --header 29 --pvalCol 7 --delim t
   --software M
   ```

   This will create a file in `~/ChIP-Seq/analysis/` called `Treatment.RECAP.bootstrap_1.txt`. The output will look the same as `Analysis_Original_peaks.xls` but with two extra columns of RECAP recalibrated and Benjamini-Hochberg adjusted RECAP p-values.

   NOTE: There are generally 29 header lines in the MACS summary file (28 if using `--nomodel`). The 7th column contains the *p*-values. The output file `Treatment.RECAP.bootstrap_1.txt` will retain the same header as the original summary file but contain a new column of recalibrated *p*-values and Benjamini-Hochberg FDR-adjusted recalibrated *p*-values. If you're not using MACS or diffReps, change the `software` value to O for other.