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 two scripts:

1. RECAP_Re-Mix.sh
2. RECAP.pl

Both 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 Math::Utils

Usage

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

Argument	Description
-i, --input	Input treatment/control BED file directory
-t, --treatment	Treatment BED file
-c, --control	Control BED file

Argument	Description
-o, --output	Output file directory
-m, --method	Method of re-mixing*
-b, --bootstrap	Number of re-mixes*

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.

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

Argument	Description
--dirOrig	Input original peak calling summary file directory
--nameOrig	Original peak calling summary file
--dirRemix	Input re-mixed peak calling summary file directory
--nameRemix	Re-mixed peak calling summary file name ending in '.bootstrap_#.bed
--dirOutput	Output directory
--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*

Argument	Description
-- 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 (Automated) Workflow

Suppose we are interested in analyzing a treatment and control file with MACS and recalibrating the resulting *p*-values.

1. Open RECAP_MACS.sh.
2. Fill out the following 6 parameters:
   a) **INPUT_DIR**: The ChIP/Control directory
   b) **CHIP_NAME**: Name of the ChIP bed file
   c) **CONTROL_NAME**: Name of the control bed file
   d) **OUTPUT_DIR**: Output directory for subsequent peak calling and RECAP analyses
   e) **BOOTSTRAP**: Number of RECAP re-mixes. (Default=1)
   f) **HEADER**: Number of header lines in the peak calling summary file (Default=29 for MACS)
3. Run RECAP_MACS.sh. `bash RECAP_MACS.sh`

Example (Manual) Workflow

Suppose we are interested in analyzing a treatment and control file with MACS and recalibrating the resulting *p*-values.

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:

```
macs2 callpeak -t Treatment.bed -c Control.bed --nomodel -p 0.1 -n Analysis --outdir ~/ChIP-Seq/analysis/
```

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

3. Analyze "re-mixed" files with MACS (please use p=0.1 for MACS and p=0.99 for SICER/diffReps):

```
macs2 callpeak -t Treatment.bootstrap_1.bed -c Control.bootstrap_1.bed --nomodel -p 0.1 -n Treatment.bootstrap_1 --outdir ~/ChIP-Seq/analysis/
```

This will create several files including `Treatment.bootstrap_1_peaks.xls`. Please retain only the `_peaks.xls` file which is to be recalibrated.

4. Recalibrate the p-values:

```
perl RECAP.pl --dirOrig ~/ChIP-Seq/analysis/ --nameOrig Treatment_peaks.xls --dirRemix ~/ChIP-Seq/analysis --nameRemix Treatment --dirOutput ~/ChIP-Seq/analysis/ --nameOutput Treatment.RECAP.bootstrap_1.txt --bootstrap 1 --header 28 --pvalCol 7 --delim t --software M
```

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 `Analysis.RECAP.bootstrap_1.txt` will retain the same header as the original summary file but contain a new column of recalibrated p-values and FDR-adjusted recalibrated p-values.