MAnorm Documentation

Release 1.1.4

ShaoLab

Jul 16, 2019

Contents

MAnorm is a robust model for quantitative comparison of ChIP-Seq data sets.

CHAPTER 1

Features

- Quantitatively compare ChIP-Seq samples
- Evaluate the overlap enrichment of protein binding sites compared to random
- Robust linear regression on common protein binding sites(peaks) for normalization
- The normalized *M-value* could serve as a quantitative measure of the differential binding
- Reflect authentic biological differences
- Support multiple format of sequencing reads

CHAPTER 2

Contents

2.1 Introduction

ChIP-Seq is widely used to characterize genome-wide binding patterns of transcription factors (TFs) and other chromatin-associated proteins. Although comparison of ChIP-Seq data sets is critical for understanding the role of their cell type/state-specific binding on modulating gene regulation programs, few quantitative approaches have been developed.

Here, we present a simple and effective method, **MAnorm**, for quantitative comparison of ChIP-Seq data sets describing transcription factor binding sites and epigenetic modifications. The quantitative binding differences inferred by MAnorm showed a strong correlation with both the changes in expression of target genes and the binding of cell type-specific regulators.

MAnorm uses common peaks of two samples as a reference to build the rescaling model for normalization, which is based on the empirical assumption that if a chromatin-associated protein has a substantial number of peaks shared in two conditions, the binding at these common regions will tend to be determined by similar mechanisms, and thus should exhibit similar global binding intensities across samples.

The normalized *M* value given by MAnorm was used as a **quantitative** measure of **differential binding** in each peak region between two samples, with peak regions associated with larger absolute *M* values exhibiting greater binding differences between two samples.

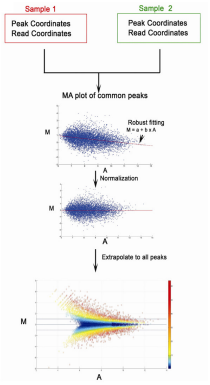
MAnorm exhibited excellent performance in quantitative comparison of ChIP-Seq data sets for both epigenetic modifications and transcription factors (TFs). The quantitative binding differences inferred by MAnorm were highly correlated with both the changes in expression of target genes and also the binding of cell type-specific regulators. With the accumulation of ChIP-seq data sets, MAnorm should serve as a powerful tool for obtaining a more comprehensive understanding of cell type-specific and cell state-specific regulation during organism development and disease onset.

2.1.1 Model Description

Assumptions

- First, we assume the true intensities of most common peaks are the same between two ChIP-Seq samples. This assumption is valid when the binding regions represented by the common peaks show a much higher level of co-localization between samples than that expected at random, and thus binding at the common peaks should be determined by similar mechanisms and exhibit similar global binding intensity between samples.
- Second, the observed differences in sequence read density in common peaks are presumed to reflect the scaling relationship of ChIP-Seq signals between two samples, which can thus be applied to all peaks.

Workflow



2.2 Tutorial

2.2.1 Installation

Like many other Python packages and bioinformatics softwares, MAnorm can be obtained easily from PyPI or Bioconda. The command below shows how to install the latest release of MAnorm in a convenient way, but you can also install it from source code alternatively.

Prerequisites

Tip: MAnorm is implemented under **Python 2.7** and will support **Python 3.X** in the following updates.

- **Python 2.7**
- setuptools
- numpy
- matplotlib
- statsmodels
- scipy

Install with pip

The latest release of MAnorm is available at PyPI, you can install via `pip`:

```
$ pip install manorm
```

Install with conda

You can also install MAnorm with conda through Bioconda channel:

```
$ conda install -c bioconda manorm
```

Install from source code

It's highly recommended to install MAnorm with `pip` or `conda`. If you prefer to install it from source code, please read the following steps:

The source code of MAnorm is hosted on GitHub, and setuptools is required for installation.

First, clone the repository of MAnorm:

```
$ git clone https://github.com/shao-lab/MAnorm.git
```

Then, install MAnorm in the source directory:

```
$ cd MAnorm
$ python setup.py install
```

Note:

- You may need to install all dependencies listed in `requirements.txt`.

- You may need to modify `$PATH` and `$PYTHONPATH` manually to make it work.

Galaxy Installation

MAnorm is available on Galaxy, you can incorporate MAnorm into your own Galaxy instance.

Please search and install MAnorm via the Galaxy Tool Shed.

2.2.2 Usage of MAnorm

To check whether MAnorm is properly installed, you can inspect the version of MAnorm by `-v/--version` option:

```
$ manorm -v
$ manorm --version
```

Command-Line Usage

MAnorm provide a console script `manorm` for running the program, the basic usage should as follows:

$ manorm –p1 peaks_file1.xls –p2 peaks_file2.xls –r1 reads_file1.bed –r2 reads_file2.bed -o output_name

Tip: Please use `-h/--help` for the details of all options.

Options

-h, --help	Show help message and exit.
-v, --version	Show version number and exit.
--p1	**[Required]** Peaks file of sample1.
--p2	**[Required]** Peaks file of sample2.
--r1	**[Required]** Reads file of sample1.
--r2	**[Required]** Reads file of sample2.
--s1	Reads shiftsize of sample1. Default: 100
--s2	Reads shiftsize of sample2. Default: 100
-w	Width of window to calculate read density. Default: 1000
-d	Summit-to-summit distance cutoff for common peaks. Default: `-w`/2
-n	Number of simulations to test the enrichment of peaks overlap between two samples.
-m	*M-value* cutoff to distinguish biased (sample-specific) peaks from unbiased peaks.
-p	*P-value* cutoff to define biased peaks.
-s	Output additional files which contains the results of original peaks.
--name1	Name of sample1. (experiment condition, cell-type etc.)

--name2	Name of sample2.
-o	**[Required]** Output directory.

Further explanation:

- `--s1/--s2`: These values are used to shift reads towards 3' direction to determine the precise binding site. Set as half of the fragment length.
- `-w`: Half of the window size when counting reads of the peak regions. MAnorm uses windows with unified length of 2 * `-w` centered at peak summits/midpoints to calculate the read density. This value should match the typical length of peaks, a value of 1000 is recommended for sharp histone marks like H3K4me3 and H3K9/27ac, and 500 for transcription factors or DNase-Seq.
- `-d`: Summit-to-summit distance cutoff for common peaks. Default= `-w` / 2. Only overlapped peaks with summit-to-summit distance less than than this value are considered as real common peaks of two samples when fitting M-A normalization model.
- `-m`: *M-value* (log2 fold change) cutoff to distinguish biased peaks from unbiased peaks. Peaks with M-value >= `-m` and P-value <= `-p` are defined as sample1-biased(specific) peaks, while peaks with M-value <= -1 * `-m` and P-value <= `-p` are defined as sample2-biased peaks.
- `-s`: By default, MAnorm will write the comparison results of unique and merged common peaks in a single output file. With this option on, MAnorm will output two extra files which contains the results of the original(unmerged) peaks.
- `--name1/--name2`: If specified, it will be used to replace the peaks/reads input file name as the sample name in output files.
- `-o`: Output directory. When `--name1` and `--name2` are not specified, MAnorm will use it as the prefix of comparison output file.

2.2.3 Input Format

Format of Peaks file

Standard **BED** format and **MACS xls** format are supported, other supported format are listed below:

```
* 3-columns tab split format

  # chr   start end
    chr1  2345  4345
    chr1  3456  5456
    chr2  6543  8543

* 4-columns tab split format

  # chr   start end   summit
    chr1  2345  4345  254
    chr1  3456  5456  127
    chr2  6543  8543  302
```

Note: The fourth column **summit** is the relative position to **start**.

Format of Reads file

Only **BED** format are supported for now. More format will be embedded in the following updates.

2.2.4 MAnorm Output

1. output_name_all_MAvalues.xls

This is the main output result of MAnorm which contains the M-A values and normalized read density of each peak, common peaks from two samples are merged together.

- chr: chromosome name
- start: start position of the peak
- end: end position of the peak
- summit: summit position of the peak (relative to start)
- m_value: M value (log2 Fold change) of normalized read densities under comparison
- a_value: A value (average signal strength) of normalized read densities under comparison
- p_value
- peak_group: indicates where the peak is come from
- normalized_read_density_in _sample1
- normalized_read_density_in_sample2

Note: Coordinates in .xls file is under **1-based** coordinate-system.

---

2. output_filters/

- sample1_biased_peaks.bed
- sample2_biased_peaks.bed
- output_name_unbiased_peaks.bed

3. output_tracks/

- output_name_M_values.wig
- output_name_A_values.wig
- output_name_P_values.wig

4. output_figures/

- output_name_MA_plot_before_normalization.png
- output_name_MA_plot_after_normalization.png
- output_name_MA_plot_with_P-value.png
- output_name_read_density_on_common_peaks.png

2.3 ChangeLog

2.3.1 v1.1.4 (2018-08-17)

- Fix an issue in setting matplotlib backend

2.3.2 v1.1.3 (2018-01-19)

- Fix a bug in the file name of filtered biased peaks
- Fix a typo

2.3.3 v1.1.2 (2018-01-18)

- Keep five digits for floats in the output files
- Fix a typo

2.3.4 v1.1.1 (2018-01-11)

- Add test module

Bugs fixed:

- Rename the file name of filtered biased peaks

2.3.5 v1.1 (2017-11-07)

Improvements:

- Refactor the package for better performance and compatibility

Bugs fixed:

- Fix the coordinates of peaks to be consistent with the corresponding coordinate system
- Fix the approximate equation in p-value calculation
- Fix the summit calculation of merged common peaks

2.4 FAQ

TODO

2.5 License

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

2.6 Contact

2.6.1 GitHub Issue

Welcome to ask questions or report bugs on GitHub:

https://github.com/shao-lab/MAnorm/issues

2.6.2 Email

Please contact:

- Hongduo Sun (sunhongduo@picb.ac.cn)
- Zhen Shao (shaozhen@picb.ac.cn)

CHAPTER 3

Citation

If you use MAnorm or any derived code, please cite this paper in your publication:

Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol. Mar 16;13(3):R16.

The Python version of MAnorm is developed by ShaoLab at CAS-MPG Partner Institute for Computational Biology, SIBS, CAS.

See also:

GitHub repository of MAnorm: https://github.com/shao-lab/MAnorm