BiasAway Documentation

Release v3.2.7

Aziz Khan and Anthony Mathelier

Oct 05, 2020

Table of contents

1 Introduction **3**

2 Installation **5**
- 2.1 Quick installation . . . 5
- 2.2 Prerequisites . . . 5
- 2.3 Install BiasAway using *conda* . . . 6
- 2.4 Install BiasAway using *pip* . . . 6
- 2.5 Install BiasAway from source . . . 6

3 How to use BiasAway **7**

4 BiasAway modules **9**
- 4.1 K-mer shuffling . . . 9
- 4.2 K-mer shuffling within a sliding window . . . 10
- 4.3 Genomic mononucleotide distribution matched . . . 11
- 4.4 Genomic mononucleotide distribution within a sliding window matched . . . 12

5 BiasAway web-server **15**
- 5.1 Introduction . . . 15
- 5.2 K-mer shuffling . . . 15
- 5.3 K-mer shuffling within a sliding window . . . 18
- 5.4 Genomic mononucleotide distribution matched . . . 18
- 5.5 Genomic mononucleotide distribution within a sliding window matched . . . 20
- 5.6 Example result page and QC plots . . . 20
- 5.7 Generation of background repositories . . . 22
- 5.8 Availability . . . 22

6 Support **23**

7 Citation **25**

Welcome to BiasAway - an open-source command-line tool and web-server that provide four approaches to generate nucleotide composition-matched DNA sequences.

CHAPTER 1

Introduction

The BiasAway software tool is introduced to generate nucleotide composition-matched DNA sequences. It is available as open source code from bitbucket.

The tool provides users with four approaches to generate synthetic or genomic background sequences matching mono- or k-mer composition of user-provided foreground sequences:

1) synthetic k-mer shuffled sequences
2) synthetic k-mer shuffled sequences in a sliding window
3) genomic mononucleotide distribution matched sequences
4) genomic mononucleotide distribution within a sliding window matched sequences

The 1st approach shuffles each user-provided sequences independently by preserving the k-mer composition of the input sequences. The 2nd approach applies the same method as the 1st approach but within a sliding window along the user-provided sequences. For the 3rd and 4th approaches, the background sequences are selected from a pool of provided genomic sequences to match the distribution of mononucleotide for each target sequence. The 4th approach considers the mean and standard deviation of %GC computed within the sliding window along the user-provided sequences to match as closely as possible the distribution for each user-provided sequence.

The approaches based on a sliding window were considered because due to evolutionary changes such as insertion of repetitive sequences, local rearrangements, or biochemical missteps, the target sequences may have sub-regions of distinct nucleotide composition.

CHAPTER 2

Installation

BiasAway is available on PyPi, through Bioconda, and the source code is available on bitbucket. BiasAway takes care of the installation of all the required python modules. If you already have a working installation of python, the easiest way to install the required python modules is by installing biasaway using `pip`.

If you are setting up Python for the first time, we recommend to install it using the Conda or Miniconda Python distribution. This comes with several helpful scientific and data processing libraries available for platforms including Windows, Mac OSX, and Linux.

You can use one of the following ways to install BiasAway.

2.1 Quick installation

2.2 Prerequisites

BiasAway requires the following Python modules:

- biopython: https://biopython.org
- numpy: https://numpy.org
- matplotlib: https://matplotlib.org/
- seaborn: https://seaborn.pydata.org/

2.2.1 Install biopython, numpy, matplotlib, and seaborn

BiasAway uses biopython, numpy, matplotlib, and seaborn you can install them using *pip* or *conda*.

Note: If you install using `pip` or `bioconda` prerequisites will be installed.

2.3 Install BiasAway using *conda*

BiasAway is available on Bioconda for installation via `conda`.

```
conda install -c bioconda biasaway
```

2.4 Install BiasAway using *pip*

BiasAway is available on PyPi for installation via `pip`.

```
pip install biasaway
```

2.5 Install BiasAway from source

You can install the development version by using `git` from our bitbucket repository at https://bitbucket.org/CBGR/biasaway.

2.5.1 Install development version from *Bitbucket*

If you have *git* installed, use this:

```
git clone https://bitbucket.org/CBGR/biasaway.git
cd biasaway
python setup.py sdist install
```

CHAPTER 3

How to use BiasAway

Once you have installed BiasAway, you can type:

```
biasaway --help
```

It will print the main help, which lists the six subcommands/modules: `k`, `w`, `g`, and `c`.

```
usage: biasaway <subcommand> [options]

positional arguments <subcommand>: {k,w,g,c}

        List of subcommands
        k       k-mer shuffling
        w       k-mer shuffling within a sliding window
        g       mononucleotide distribution matched
        c       mononucleotide distribution within a sliding window matched

optional arguments:
  -h, --help            show this help message and exit
  -v, --version         show program's version number and exit
```

To view the help for the individual subcommands, please type:

Note: Please check `BiasAway modules` to see a detailed summary of available **options**.

To view `k` module help, type

```
biasaway k --help
```

To view `w` module help, type

```
biasaway w --help
```

To view `g` module help, type

```
biasaway g --help
```

To view `c` module help, type

```
biasaway c --help
```

CHAPTER 4

BiasAway modules

The BiasAway software tool is introduced to generate nucleotide composition-matched DNA sequences. It is available as open source code from bitbucket.

The tool provides users with four approaches to generate synthetic or genomic background sequences matching mono- and k-mer composition of user-provided foreground sequences:

Note: BiasAway can generate distribution plots for QC. Plots provide information about distribution of %GC, dinucleotides, and lengths for the input sequences and generated sequences. Moreover, BiasAway provides the following QC metrics for comparing these distributions whenever possible: mean absolute error and goodness of fit computed as Pearson's chi-squared statistic, log-likelihood ratio test (G-test), and the Cressie-Read power divergence.

Note: BiasAway also comes with a Web App available at http://biasaway.uio.no.

4.1 K-mer shuffling

Each user-provided sequence will be shuffled to keep its k-mer composition. This module can be used for any k, for instance use -k 1 for conserving the mononucleotide composition of the input sequences.

Usage:

```
biasaway k [options]
```

Note: Please scroll down to see a detailed summary of available **options**.

Help:

```
biasaway k --help
```

Example:

```
biasaway k -f path/to/FASTA/file/my_fasta_file.fa
```

It will output the generated sequences on stdout, keeping the dinucleotide composition of the input sequence by default (k-mer with k=2 is the default). If you wish to save the sequences in a specific file, you can type:

```
biasaway d -f path/to/FASTA/file/my_fasta_file.fa > path/to/output/FASTA/file/my_
↪fasta_output.fa
```

Summary of options

Option	Description
-h, –help	To show the help message and exit
-f, –foreground	Foreground file in fasta format.
-k, –kmer	K-mer to be used for shuffling (`default: 2` for dinucleotide shuffling)
-n, –nfold	How many background sequences per each foreground sequence will be generated (`default: 1`)
-e, –seed	Seed number to initialize the random number generator for reproducibility (`default: integer from the current time`)
-p, –plot-filename	Base filename for all the plots and related statistics looking at %GC, dinucleotide, and lengths distributions (“default: not activated so no plot and statistics produced)

4.2 K-mer shuffling within a sliding window

For each user-provided sequence, a window will slide along to shuffle the nucleotides within the window, keeping the local k-mer composition. As such, the generated sequences will preserve the local k-mer composition of the input sequences along them.

Usage:

```
biasaway w [options]
```

Note: Please scroll down to see a detailed summary of available **options**.

Help:

```
biasaway w --help
```

Example:

```
biasaway w -f path/to/FASTA/file/my_fasta_file.fa
```

It will output the generated sequences on stdout, keeping the local dinucleotide composition of the input sequences (k=2 for dinucleotide shuffling is used as default). If you wish to save the sequences in a specific file, you can type:

```
biasaway w -f path/to/FASTA/file/my_fasta_file.fa > path/to/output/FASTA/file/my_
↪fasta_output.fa
```

Summary of options

Option	Description
-h, –help	To show the help message and exit
-f, –foreground	Foreground file in fasta format.
-k, –kmer	K-mer to be used for shuffling (`default: 2` for dinucleotide shuffling)
-n, –nfold	How many background sequences per each foreground sequence will be generated (`default: 1`)
-w, –winlen	Window length (`default: 100`)
-s, –step	Sliding step (`default: 50`)
-e, –seed	Seed number to initialize the random number generator for reproducibility (`default: integer from the current time`)
-p, –plotfilename	Base filename for all the plots and related statistics looking at %GC, dinucleotide, and lengths distributions (“default: not activated so no plot and statistics produced)

4.3 Genomic mononucleotide distribution matched

Given a set of available background sequences (pre-computed or provided by the user), each user-provided foreground sequence will be matched to a background sequence having the same mononucleotide composition.

The first time you run this module, you need to provide a set of potential background sequences using the *–background* argument. The *–bgdirectory* argument is necessary and will contain the decomposition of the background sequences in dedicated files per %GC content.

If you already have such a pre-computed background directory, you can only use the *–bgdirectory* argument to speed-up the process.

Usage:

```
biasaway g [options]
```

Note: Please scroll down to see a detailed summary of available **options**.

Help:

```
biasaway g --help
```

Example:

```
biasaway g -f path/to/FASTA/file/my_fasta_file.fa -b path/to/background.fa -r path/to/
↪bgdirectory
```

It will output the generated sequences on stdout. If you wish to save the sequences in a specific file, you can type:

```
biasaway g -f path/to/FASTA/file/my_fasta_file.fa -b path/to/background.fa -r path/to/
↪bgdirectory > path/to/output/FASTA/file/my_fasta_output.fa
```

Summary of options

Option	Description
-h, –help	To show the help message and exit
-f, –foreground	Foreground file in fasta format.
-n, –nfold	How many background sequences per each foreground sequence will be generated (`default: 1`)
-r, –bgdirectory	Background directory (must be empty if –background is used). See documentation for details.
-b, –background	Background file in fasta format. Not necessary if a background directory has already been computed previously.
-l, –length	Try to match the length as closely as possible (`not set by default`)
-e, –seed	Seed number to initialize the random number generator for reproducibility (`default: integer from the current time`)
-p, –plotfilename	Base filename for all the plots and related statistics looking at %GC, dinucleotide, and lengths distributions (“default: not activated so no plot and statistics produced)

4.4 Genomic mononucleotide distribution within a sliding window matched

Given a set of available background sequences (pre-computed or provided by the user), each user-provided foreground sequence will be matched to a background sequence having a close mononucleotide local composition. Specifically, distribution of %GC composition in a sliding window are computed for foreground and background sequences; a foreground sequence with a mean m_f and standard deviation sdev_f of %GC in the sliding window is matched to a background sequence if its mean %GC m_b is such that: .. math:

```
m_f - N * sdev_f <= m_b <= m_f + N * sdev_f
```

with *N* equals to 2.6 by default.

The first time you run this module, you need to provide a set of potential background sequences using the *–background* argument. The *–bgdirectory* argument is necessary and will contain the decomposition of the background sequences in dedicated files per %GC content.

If you already have such a pre-computed background directory, you can only use the *–bgdirectory* argument to speed-up the process.

Usage:

```
biasaway c [options]
```

Note: Please scroll down to see a detailed summary of available **options**.

Help:

```
biasaway c --help
```

Example:

```
biasaway c -f path/to/FASTA/file/my_fasta_file.fa -b path/to/background.fa -r path/to/
↪bgdirectory
```

It will output the generated sequences on stdout. If you wish to save the sequences in a specific file, you can type:

```
biasaway c -f path/to/FASTA/file/my_fasta_file.fa -b path/to/background.fa -r path/to/
↪bgdirectory > path/to/output/FASTA/file/my_fasta_output.fa
```

Summary of options

Option	Description
-h, –help	To show the help message and exit
-f, –foreground	Foreground file in fasta format.
-n, –nfold	How many background sequences per each foreground sequence will be generated (`default: 1`)
-r, –bgdirectory	Background directory (must be empty if –background is used). See documentation for details.
-b, –background	Background file in fasta format. Not necessary if a background directory has already been computed previously.
-l, –length	Try to match the length as closely as possible (`not set by default`)
-w, –winlen	Window length (`default: 100`)
-s, –step	Sliding step (`default: 50`)
-d, –deviation	Deviation from the mean (`default: 2.6 for a threshold of mean + 2.6 * stdev`)
-e, –seed	Seed number to initialize the random number generator for reproducibility (`default: integer from the current time`)
-p, –plotfilename	Base filename for all the plots and related statistics looking at %GC, dinucleotide, and lengths distributions (“default: not activated so no plot and statistics produced)

CHAPTER 5

BiasAway web-server

5.1 Introduction

The BiasAway web-server provides an interactive and easy to use interface for users to upload FASTA files and to generate background sequences. It comes with precomputed genomic partitions of 100, 250, 500, 750, and 1000 bp bins for the genome of nine species (*Arabidopsis thaliana*; *Caenorhabditis elegans*; *Danio rerio*; *Drosophila melanogaster*; *Homo sapiens*; *Mus musculus*; *Rattus norvegicus*; *Saccharomyces cerevisiae*; and *Schizosaccharomyces pombe*). These background sequences are provided through Zenodo at 10.5281/zenodo.3923866. These background sequences were generated using the script at https://bitbucket.org/CBGR/biasaway_background_construction, which can be used by users to generate their own background sequences. The result page provides information about mononucleotide, dinucleotide, and length distributions for the provided and generated sequences for comparison.

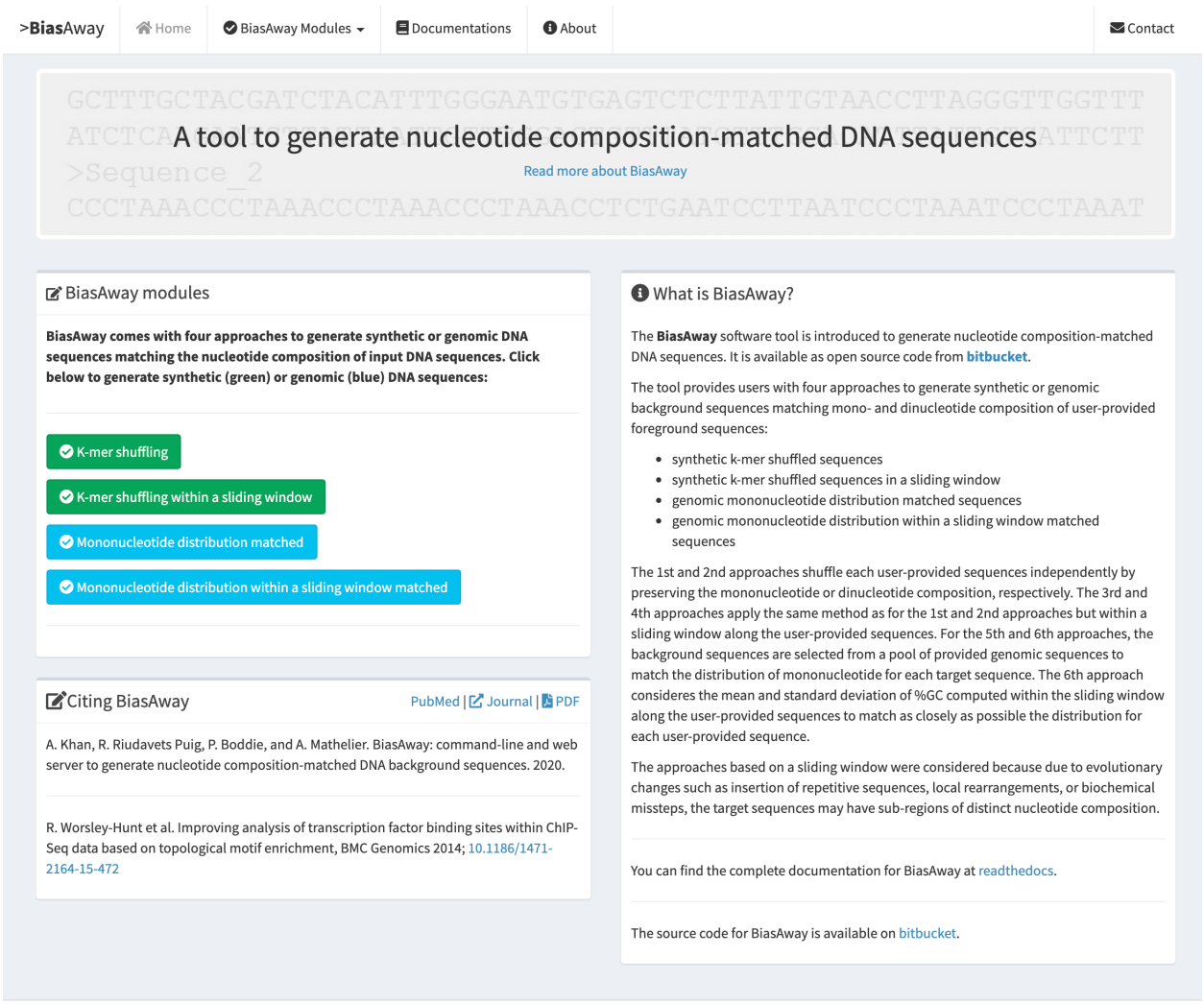
BiasAway has four modules:

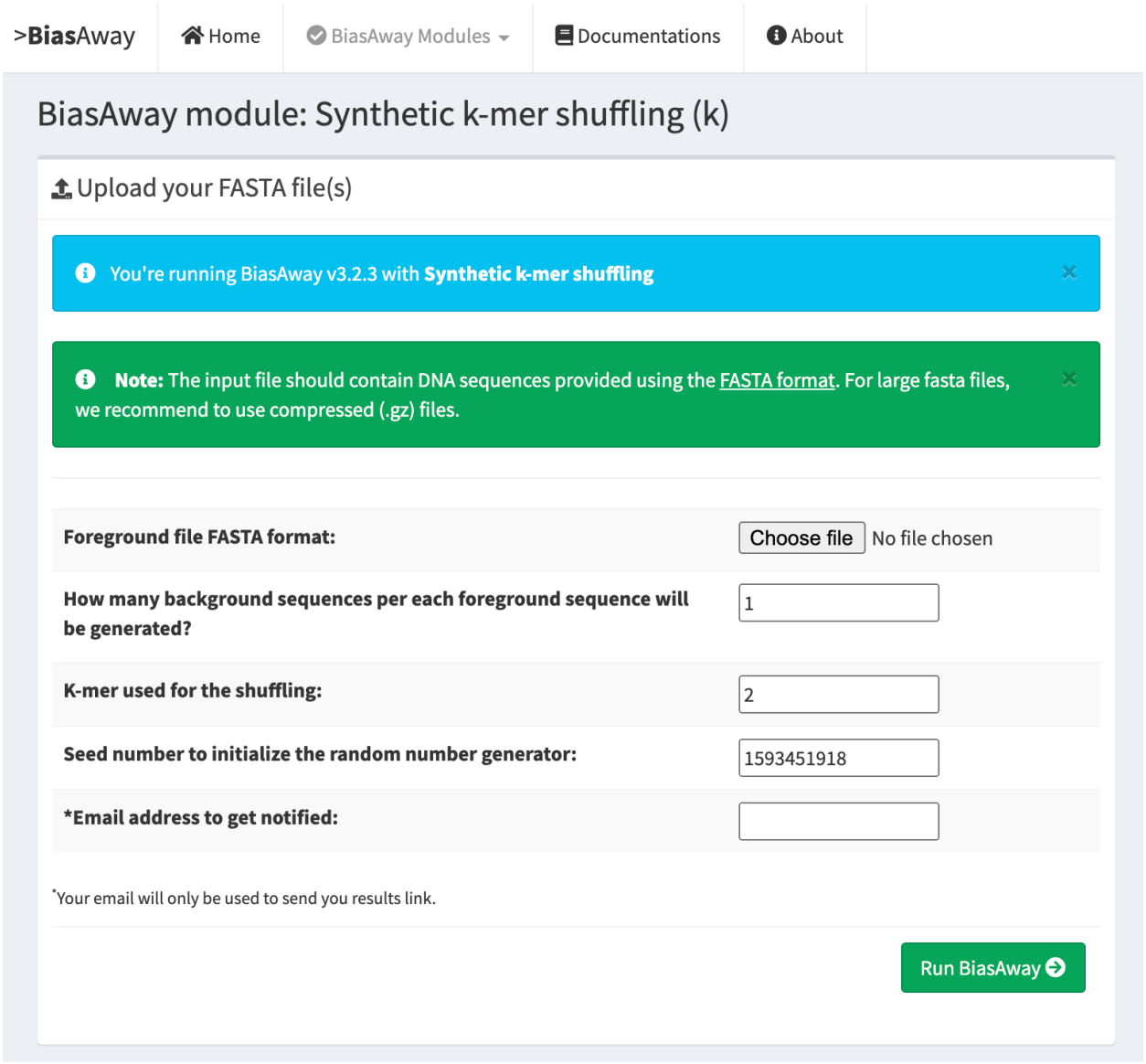
Note: The BiasAway web-application automatically generate distribution plots for QC. Plots provide information about distribution of %GC, dinucleotides, and lengths for the input sequences and generated sequences. Moreover, BiasAway provides the following QC metrics for comparing these distributions whenever possible: mean absolute error and goodness of fit computed as Pearson's chi-squared statistic, log-likelihood ratio test (G-test), and the Cressie-Read power divergence.

Below are screenshots for individual modules.

5.2 K-mer shuffling

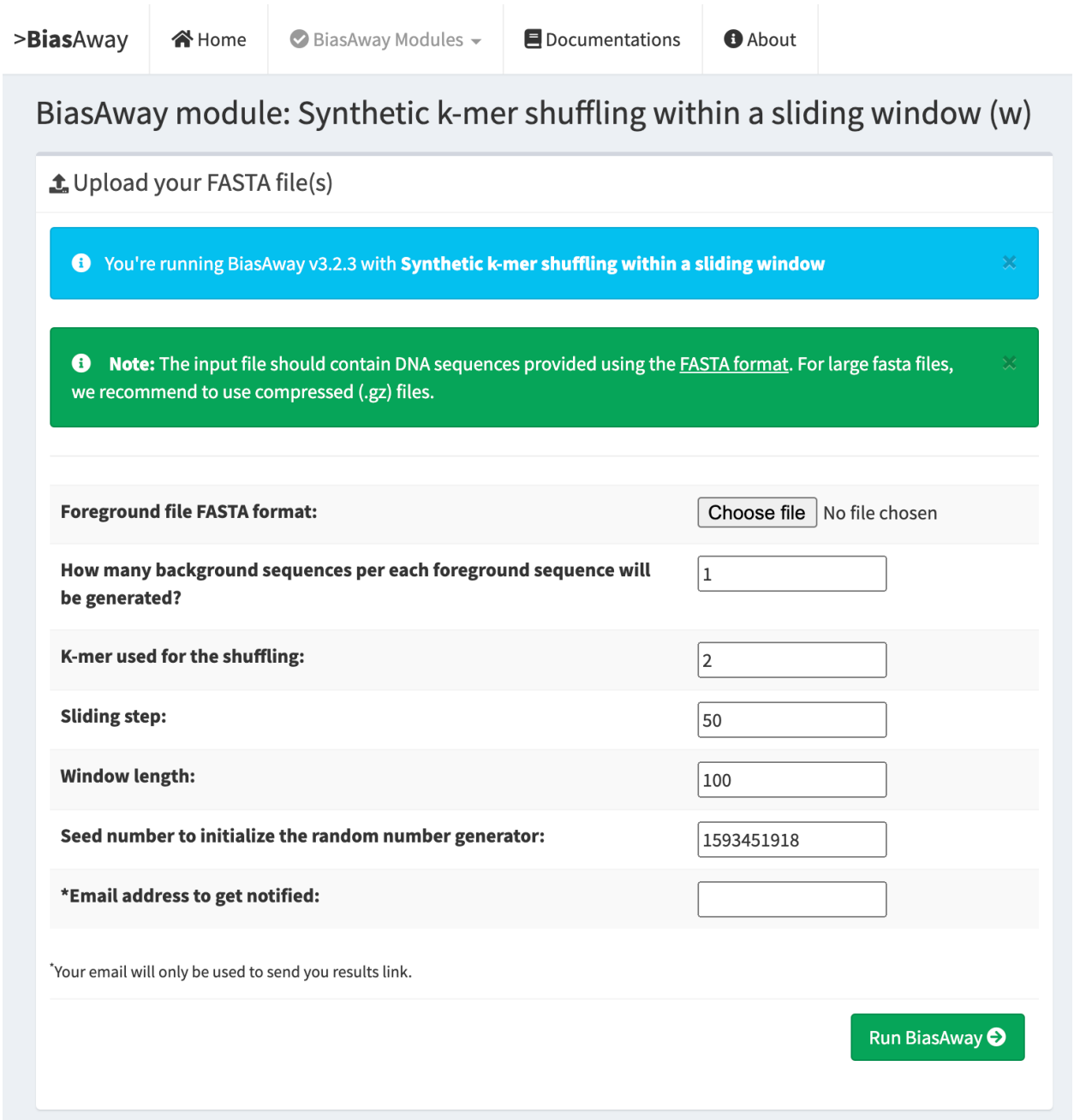
This module should be run when the user aims at preserving the global k-mer nucleotide frequencies of input sequences.


>BiasAway
Home
BiasAway Modules
Documentations
About
Contact
A tool to generate nucleotide composition-matched DNA sequences
Read more about BiasAway
BiasAway modules
BiasAway comes with four approaches to generate synthetic or genomic DNA sequences matching the nucleotide composition of input DNA sequences. Click below to generate synthetic (green) or genomic (blue) DNA sequences:
K-mer shuffling
K-mer shuffling within a sliding window
Mononucleotide distribution matched
Mononucleotide distribution within a sliding window matched
Citing BiasAway
PubMed | Journal | PDF
A. Khan, R. Riudavets Puig, P. Boddie, and A. Mathelier. BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences. 2020.
R. Worsley-Hunt et al. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment, BMC Genomics 2014; 10.1186/1471-2164-15-472
What is BiasAway?
The BiasAway software tool is introduced to generate nucleotide composition-matched DNA sequences. It is available as open source code from bitbucket.
The tool provides users with four approaches to generate synthetic or genomic background sequences matching mono- and dinucleotide composition of user-provided foreground sequences:
synthetic k-mer shuffled sequences
synthetic k-mer shuffled sequences in a sliding window
genomic mononucleotide distribution matched sequences
genomic mononucleotide distribution within a sliding window matched sequences
The 1st and 2nd approaches shuffle each user-provided sequences independently by preserving the mononucleotide or dinucleotide composition, respectively. The 3rd and 4th approaches apply the same method as for the 1st and 2nd approaches but within a sliding window along the user-provided sequences. For the 5th and 6th approaches, the background sequences are selected from a pool of provided genomic sequences to match the distribution of mononucleotide for each target sequence. The 6th approach consideres the mean and standard deviation of %GC computed within the sliding window along the user-provided sequences to match as closely as possible the distribution for each user-provided sequence.
The approaches based on a sliding window were considered because due to evolutionary changes such as insertion of repetitive sequences, local rearrangements, or biochemical missteps, the target sequences may have sub-regions of distinct nucleotide composition.
You can find the complete documentation for BiasAway at readthedocs.
The source code for BiasAway is available on bitbucket.
BiasAway Copyright © 2020. The content is licensed under Creative Commons Attribution 4.0 International License.
NCMM
Centre for Molecular Medicine Norway
UiO : Faculty of Medicine
University of Oslo


>BiasAway
Home
BiasAway Modules
Documentations
About
BiasAway module: Synthetic k-mer shuffling (k)
Upload your FASTA file(s)
You're running BiasAway v3.2.3 with Synthetic k-mer shuffling
Note: The input file should contain DNA sequences provided using the FASTA format. For large fasta files, we recommend to use compressed (.gz) files.
Foreground file FASTA format:
Choose file
No file chosen
How many background sequences per each foreground sequence will be generated?
1
K-mer used for the shuffling:
2
Seed number to initialize the random number generator:
1593451918
*Email address to get notified:
*Your email will only be used to send you results link.
Run BiasAway

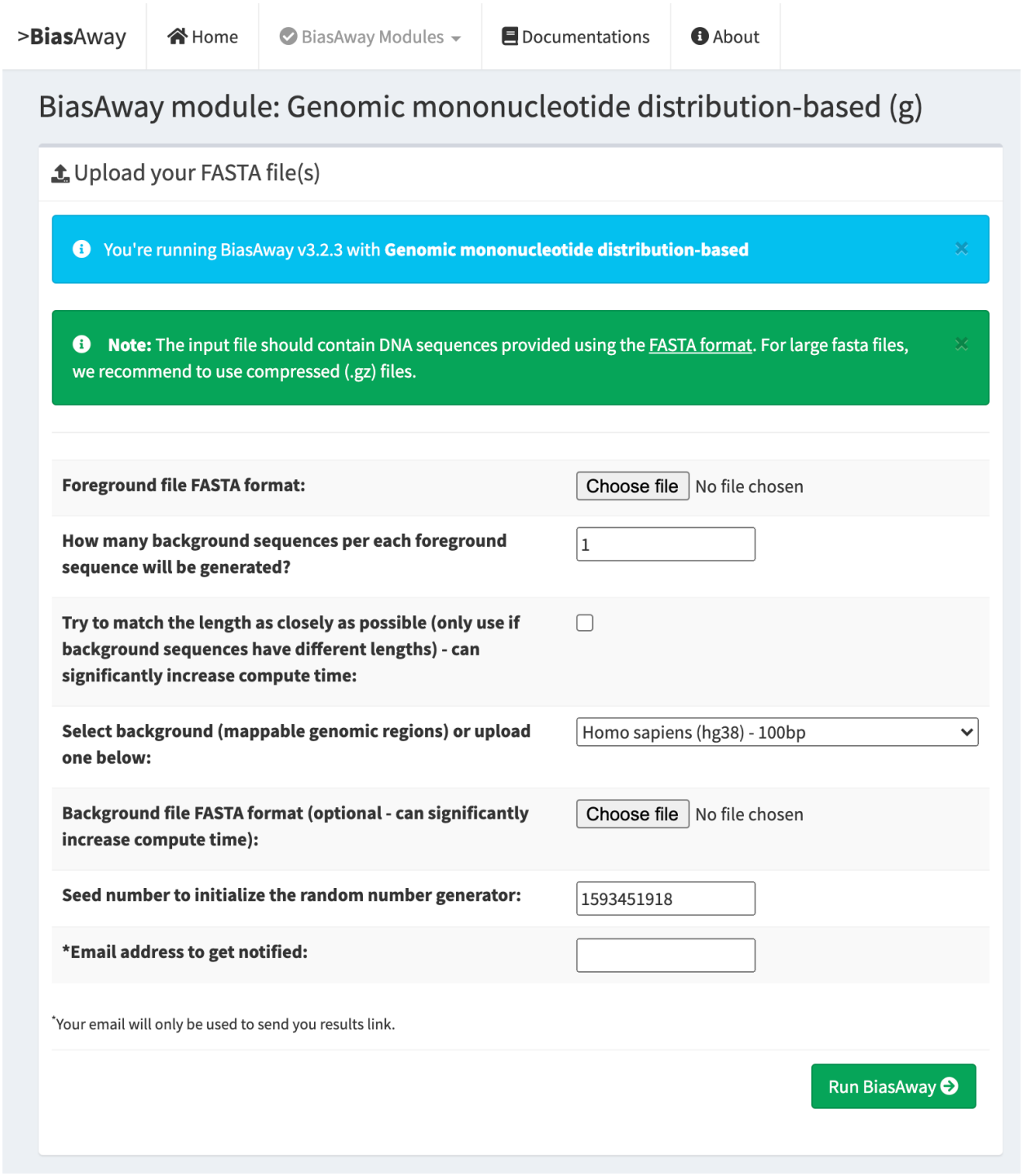
5.3 K-mer shuffling within a sliding window

This module should be run when the user aims at preserving the local k-mer nucleotide frequencies of input sequences.



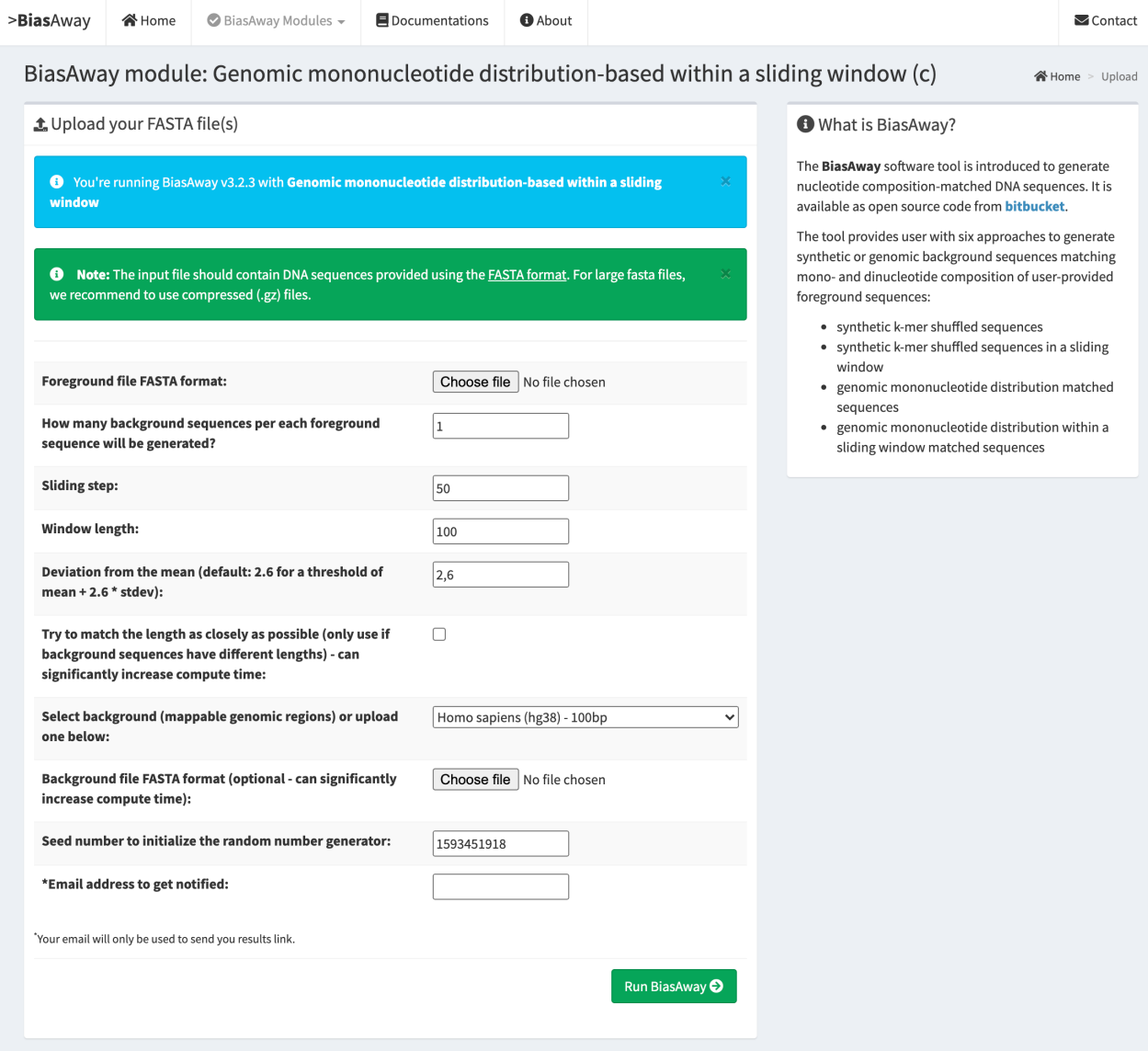
5.4 Genomic mononucleotide distribution matched

This module should be run when the user aims at selecting genuine genomic background sequences from a pool of provided genomic sequences to match the distribution of mononucleotide for each target sequence.


>BiasAway
Home
BiasAway Modules
Documentations
About
BiasAway module: Genomic mononucleotide distribution-based (g)
Upload your FASTA file(s)
You're running BiasAway v3.2.3 with Genomic mononucleotide distribution-based
Note: The input file should contain DNA sequences provided using the FASTA format. For large fasta files, we recommend to use compressed (.gz) files.
Foreground file FASTA format:
Choose file
No file chosen
How many background sequences per each foreground sequence will be generated?
1
Try to match the length as closely as possible (only use if background sequences have different lengths) - can significantly increase compute time:
Select background (mappable genomic regions) or upload one below:
Homo sapiens (hg38) - 100bp
Background file FASTA format (optional - can significantly increase compute time):
Choose file
No file chosen
Seed number to initialize the random number generator:
1593451918
*Email address to get notified:
*Your email will only be used to send you results link.
Run BiasAway

5.5 Genomic mononucleotide distribution within a sliding window matched

This module should be run when the user aims at selecting genuine genomic background sequences from a pool of provided genomic sequences to match the local distribution of mononucleotide for each target sequence.



5.6 Example result page and QC plots

BiasAway provides quality control (QC) plots and metrics to assess the similarity of the mono- and di-nucleotide, and length distributions for the foreground and background sequences. Specifically, four plots are provided to visualize how similar the foreground and background sequences are when considering (2) their distributions of %GC content using density plots, (2) their dinucleotide contents considering all IUPAC nucleotides using a heatmap, (3) their dinucleotide contents considering adenine, cytosine, guanine, and thymine nucleotides using a heatmap, and (4) their distributions of lengths.

>BiasAway | Home | BiasAway Modules | Documentations | About | Contact

BiasAway Results

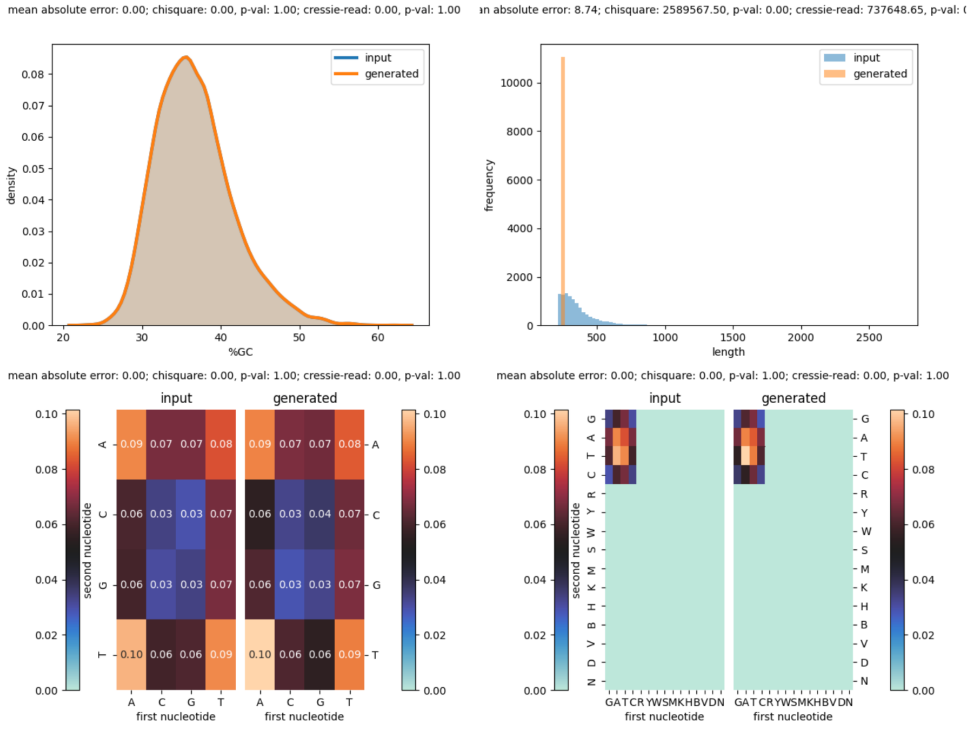
Home > Results

BiasAway module details and download sequences

Here is the URL to the file containing the background sequences generated. This file will be stored for 10 days.

BiasAway module:	g: Genomic mononucleotide distribution-based
BiasAway command with parameters:	`biasaway g --foreground /var/www/apps/biasaway-app/media/remap2020_LFY_nr_macs2_TAIR10_v1_0.fasta.gz --nfold 1 --bgdirectory /var/www/apps/biasaway-app/background_fasta/R64-1-1/250bp/ --seed 1593424198 --plot_filename /var/www/apps/biasaway-app/temp/BiasAway_g_vla0t891_20200629`
Job status:	Completed
Date and time:	June 29, 2020, 11:12 a.m.
Results URL:	https://testbiasaway.uio.no/result/68/

Background sequences are ready for download!



Download background sequences

5.7 Generation of background repositories

Modules *g* and *c* of BiasAway require the generation of a background repository for the genome of interest. This can be created with the script located at our BitBucket repository.

Our BiasAway Web-Server contains precomputed background repositories for 9 species. The genome fasta files used to create these can be found below:

- Homo sapiens: GRCh38/hg38
- Mus musculus: mm10
- Rattus norvegicus: Rnor 6.0
- Arabidopsis thaliana: TAIR10
- Danio rerio: GRCz11
- Drosophila melanogaster: dm6
- Caenorhabditis elegans: WBcel235
- Saccharomyces cerevisiae
- Schizosaccharomyces pombe: ASM294v2

Please note that some genome fasta files are separated by chromosomes in their original repositories. In that case, please make sure to concatenate all chromosome fasta files in one single genome fasta file.

We also provide a collection of precomputed background repositories for the nine organisms mentioned above using k-mers of size 100, 250, 500, 750 and 1000 base pairs. They can be found as individual compressed files in our Zenodo repository

5.8 Availability

The BiasAway web-server is freely available at:

> http://biasaway.uio.no

CHAPTER 6

Support

If you have questions, or found any bug in the program, please write to us at anthony.mathelier[at]ncmm.uio.no and azizk[at]stanford.edu.

You can also report the issues to our bitbucket repo

CHAPTER 7

Citation

If you used BiasAway, please cite:

- A. Khan, R. Riudavets Puig, P. Boddie, and A. Mathelier. BiasAway: command-line and web server to generate nucleotide composition-matched DNA background sequences, 2020.
- R. Worsley-Hunt *et al.* Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment, *BMC Genomics* 2014; 10.1186/1471-2164-15-472