

BICseq2 manual page

» BICseq2

- [BICseq2-norm](#)
- [BICseq2-seg](#)

» [Mappability Files](#)

- Human hg19
 - [50mer File](#)
 - [75mer File](#)
 - [100mer File](#)
- Human hg18
 - [50mer File](#)

» [Reference Genome](#)

- [Human hg19](#)
- [Human hg18](#)

Introduction

BICseq2 is an algorithm developed for the normalization of high-throughput sequencing (HTS) data and detection of copy number variations (CNV) in the genome. BICseq2 can be used for detecting CNVs with or without a control genome. There are two main components of the algorithm:

1. BICseq2-norm is for normalizing potential biases in the sequencing data.
2. BICseq2-seg is for detecting CNVs based on the normalized data given by BICseq2-norm.

The general pipeline using BICseq2 for CNV detection is as follows.

- Only a case genome is sequenced and no control genome is available.
 1. Get the uniquely mapped reads from the bam file. (You may use our modified samtools provided [here](#)).
 2. Use BICseq2-norm to remove the biases in the data.
 3. Use BICseq2-seg to detect CNVs based on the normalized data.
- Both the case genome and control genome are available (In a cancer study, the case genome is the tumor genome, and the control genome can be the matched normal genome).
 1. Get the uniquely mapped reads from the case and control genome bam files, respectively.
 2. Normalize the case and control genomes individually using BICseq2-norm.
 3. Detect CNVs in the case genome based on the normalized data of the case genome and the control genome.

BICseq2-norm usage

Before using BICseq2-norm, you have to first compile the C code. To compile, you may simply type

```
make clean
make
```

After the compilation, you can use the perl program BICseq2-norm.pl for normalization.

Usage: BICseq2-norm.pl [options] <configFile><output>

Options:

```
--help
-l=<int>: read length
-s=<int>: fragment size
-p=<float>: a subsample percentage (Default 0.0002).
-b=<int>: bin the expected and observed as< int> bp bins (Default 100).
-gc_bin: if specified, report the GC-content in the bins
--NoMapBin: if specified, do NOT bin the reads according to the mappability
--bin_only: only bin the reads without normalization
--fig=<string>: plot the read count VS GC figure in the specified file (in PDF format)
--title=<string>: title of the figure
--tmp=<string>: the temp directory
```

< configFile> specifies the location of the configuration file that has the necessary information for normalization. See below for the format of the configuration file.

< output> is the file that stores the parameter estimates in the GAM model. This is not useful for general users.

The < configFile> has the following format:

chromName	faFile	MapFile	readPosFile	binFileNorm
chr1	chr1.fa	hg18.CRC.50mer.chr1.txt	chr1.seq	chr1.norm.bin
chr2	chr2.fa	hg18.CRC.50mer.chr2.txt	chr2.seq	chr2.norm.bin

In the < configFile>, the columns should be tab-delimited. The first row of this file is assumed to be the header of the file and will be omitted by BICseq2-norm.

The 1st column (chromName) is the chromosome name.

The 2nd column (faFile) is the reference sequence of this chromosome (Human hg18 and hg19 are available for download.).

The 3rd column (MapFile) is the mappability file of this chromosome (Human hg18 (50bp) and hg19 (50bp and 75bp) are available for download).

The 4th column (readPosFile) is the file that stores all the mapping positions of all reads that uniquely mapped to this chromosome.

The 5th column (binFile) is the file that stores the normalized data. The data will be binned with the bin size as specified by the option -b.

BICseq2-seg usage

Similar to BICseq2-norm, you can first compile BICseq2-seg with

```
make clean
make
```

After compilation, you can detect CNV with the perl program BICseq2-seg.pl.

Usage: BICseq2-seg.pl [options] <configFile><output>

Options:

```
--lambda=<float>: the (positive) penalty used for BICseq2
--tmp=<string>: the temp directory
--help: print this message
--fig=<string>: plot the CNV profile in a PNG file
--title=<string>: the title of the figure
--nrm: do not remove likely germline CNVs (with a matched normal) or segments with bad mappability (without a matched normal)
--bootstrap: perform bootstrap test to assign confidence (only for one sample case)
--noscale: do not automatically adjust the lambda parameter according to the noise level in the data
--strict: if specified, use a more stringent method to adjust the lambda parameter
```

- control: the data has a control genome
- detail: if specified, print the detailed segmentation result (for multiSample only)

As with the original BIC-seq algorithm, the --lambda parameter is the main parameter used for tuning the smoothness of the CNV profile. The larger the value, the fewer segments the file profile will have. The default value is 2.

< configFile> stores the necessary information for BICseq2-seg to detect CNVs.
 < output> stores the final CNV detection results.

< configFile> has the following format (tab-delimited; first row treated as header).

- If there is no control, the format is

chromName	binFileNorm
chr1	chr1.norm.bin
chr2	chr2.norm.bin

The 1st column (chromName) is just the chromosome name.

The 2nd column (binFileNorm) is the normalized bin file as obtained from BICseq2-norm.

- If there is a control, the format is

chromName	binFileNorm.Case	binFileNorm.Control
chr1	CaseChr1.norm.bin	ControlChr1.norm.bin
chr2	CaseChr1.norm.bin	ControlChr1.norm.bin

The 2nd column (binFileNorm.Case) is the normalized bin file of the case genome as obtained from BICseq2-norm.

The 3rd column (binFileNorm.Control) is the normalized bin file of the control genome as obtained from BICseq2-norm.

Note: If you have a control, you must specify to --control to let BICseq2 know that the data is a case/control study.

How to cite BIC-seq2:

Xi, R.*, Lee, S., Xia, Y., Kim, T. and Park, P.* (2016) Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants, Nucleic Acids Research, 44(13):6274-86.

Xi, R., Hadjipanayis, A.G., Luquette, L.J., Kim, T.M., Lee, E., Zhang, J.H., Johnson, M.D., Muzny, D.M., Wheeler, D.A., Kucherlapati, R., and Park, P.* (2011). Copy number alteration detection in sequencing data using the Bayesian information criterion, Proceedings of the National Academy of Sciences, USA, 108(46):E1128-36.

Frequently Asked Questions.

Please see this [document](#).