Frequently Asked Questions

1. What are .seq files and how to get the .seq files?
   I use .seq files to represent files containing mapping positions of short reads. Each .seq file contains mapping positions of one sample at ONE chromosome. A .seq file should have only one column and each row of the file is just the mapping position of one read.

   To get .seq file, you may just use the modified samtools or you can write a small script (e.g. perl or python scripts) to get these mapping positions. The modified samtools is a bit faster.

   Suppose that you have a bamfile named *mybam.bam* and you want to put the .seq files into the directory *myfolder*. Also assume that your data is mapped with BWA and your modified samtools in under the directory *modifiedSamtools*, you can use the following command to get the .seq files

   modifiedSamtools/samtools view –U BWA, *myfolder/,N,N mybam.bam*

   If you want to only consider the reads with minimum mapping quality 20, you may use

   modifiedSamtools/samtools view –U BWA, *myfolder/,N,N –q 20 mybam.bam*

   If your data is mapped by Bowtie, you can just replace BWA with Bowtie.

   If your data is of mixed read length, you should ONLY use reads of one read length; Otherwise, there might be false positives due to different read lengths (If your read is something like 100 and 101, it is fine to use these reads together). If you want to get the mapping positions in a certain range, say 100 to 101, you may simply use

   modifiedSamtools/samtools view –U BWA, *myfolder/,N,N,100,101 –q 20 mybam.bam*

2. How to get the mappability files?
   The mappability file contains mappable regions of short reads. The mappability file should have two columns separated by tabs. Each row of the mappability file corresponds to one mappable region. The first and the second column are the starting and ending positions of the mappable region, respectively.

   BIC-seq2 can work with uniquely mapped reads and/or multiply-mapped reads (reads that can be mapped to more than one position). If only uniquely mapped reads are used, the mappable regions are defined as uniquely mapped positions of certain read length. If multiply-mapped reads are also used, the mappable regions are defined as non-'N' regions in the reference

genome. Therefore, if both uniquely mapped reads and multiply-mapped reads are used, one can just take the non-'N' regions as the mappability file. Note that if multiply-mapped reads are used, you have to ask the aligner (such as BWA) to randomly report one of their mapped positions as their mapped position (in the bam file).

If only uniquely mapped reads are used and your data is human data, you may download mappability files for a few read lengths from the BIC-seq2 website. If your organism not Homo Sapiens, you have to generate these mappability files yourself. The UCSC genome browser provides mappability files for a few organisms (mostly in bigwig format), but you have to convert the bigwig format to the format as required by BIC-seq2 (e.g. first convert to wig format and then extract the uniquely mapped positions). If UCSC genome browser also does not have the mappability files you needed, you have to generate them yourself by aligning all possible k-mers (k is the read length of your data) to your reference genome and identify all positions which the k-mer starting from only have one mapping position allowing a few mismatches (e.g. 2 mismatches).

3. Can BIC-seq2 work with BWA mem mapping?
   Yes, but you have to first filter the hard-clipped reads. To do that, you could just filter your bam file by read length. BIC-seq2 actually works a little better with BWA aln.