**Genetic Location Identification via Bionano**

**Version 1.6**

**User manual**

**Chunlong-Chen , Weitao Wang**

**2019-12-11**

**Contents**

1 Prerequisites

1. JDK version 1.8
2. R is required for peak calling
3. About 7GB of available RAM
4. Steps in this document assumes that your current working directory is ‘example’ and that the ‘replicon’ executable is in the parent directory.

2 Introduction

This is an extension software based on high throughput Bionano platform for the genomic position determination of any biological process correlated with fluorescently labeled DNA. (For example: DNA replication, DNA recombination, DNA repair) and provide the frequency occurred in population-based data for each found genomic position.

1. Clustering signal and get segmentation

The input data ask for Bionano [bnx](https://bionanogenomics.com/wp-content/uploads/2018/04/30038-BNX-File-Format-Specification-Sheet.pdf) and [xmap](https://bionanogenomics.com/wp-content/uploads/2017/03/30040-XMAP-File-Format-Specification-Sheet.pdf) dataformat files as most basic raw data. All analysis is based on these 2 files. The jar package should be run under linux terminal. The Patent\_GetSegment.jar can cluster neighbored signal based on the distance cutoff value set by user and merge the neighbored clustered primary segmentation. The default cluster cutoff value is 16384 and merged cutoff value is 32768. If you don’t want to merge the clustered primary segmentation to final segmentation, you can set the -MergeCutoff value as -1.

The must have parameter is -BNX, -XMP, -o, -FN and -Signal\_C

-BNX: The parent path to BNX file (The document where .bnx file exist)

-XMP: The parent path to XMP file (The document where .xmap file exist)

-O: The path of output file

-FN: The sample name list under the BNX and XMP path (Please pay attention the .bnx and .xmap must be with same file name beside suffix .bnx and .xmap, and must be end with such kind of suffix for dataformat recognition)

-Signal\_C: The bionano channel of signal in your experiment corresponding to mapping channel.

**The output is like below**

1. A .txt file record all detailed information about mapped fiber with red signal including red signal position, signal intensity and SNR.
2. A fiber coordinate bed file recorded all mapped fiber
3. A segmentation bed file recorded all final segmentation

**The other option**

-FL: You will get the mapped or unmapped DNA fiber length

-SignalPerSeg: pick up the segmentation which containing at least signal number you set here

-SignalPerFiber: pick up the fiber which containing at least signal number you set here

-FiberWithRed: filtering the fiber without red signal

-TXT: If you second try to modify any parameter, you can use the -TXT to set the .txt file generated by previous running.

**The example (Quick start)**

Java -jar Patent\_GetSegment.jar -BNX xxx/PathToBNX -XMP xxx/ PathToXMP -FN 1905async 1905\_FC1 1905\_FC2 -o xxx/test -Signal\_C 2 -FL -FiberWithRed -SignalPerSeg 3

Note: Under path to BNX must have 3 document named as 1905async.bnx , 1905\_FC1.bnx , 1905\_FC2.bnx. Same as path to XMP with 1905async.xmap , 1905\_FC1.xmap , 1905\_FC2.xmap

1. Count the signal number in each bin

The input data needs .txt file got by Patent\_GetSegment.jar. It used to generate bin files along the genome and count the signal number in each bin. The must have parameter is -WindowSize, -TXT, -Chr\_Length, -o, -Mode

-WindowSize: The bin width you set

-TXT: The path to file generated by Patent\_GetSegment.jar

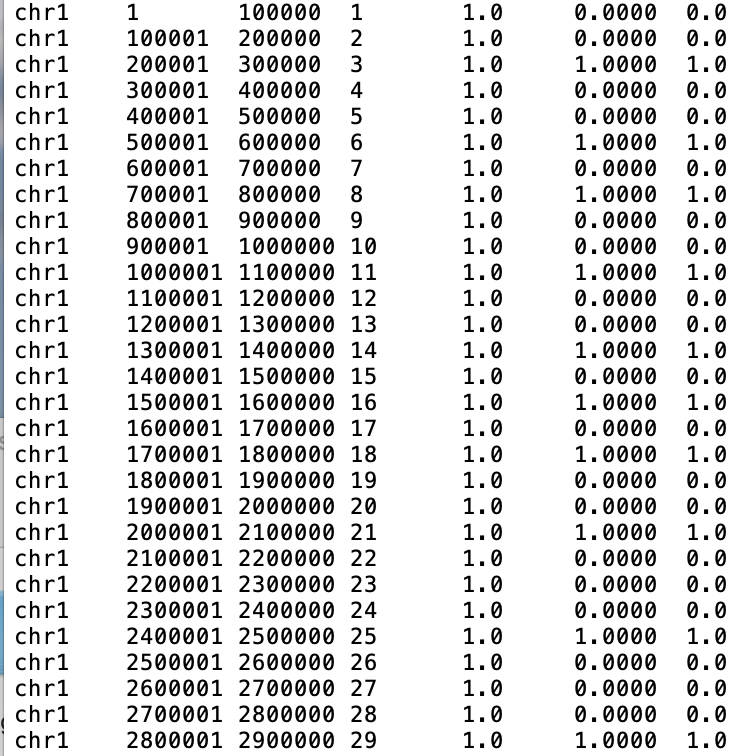
-Chr\_Length: The chromosome length (Refer to [UCSC](https://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes))

-o: The path of output file

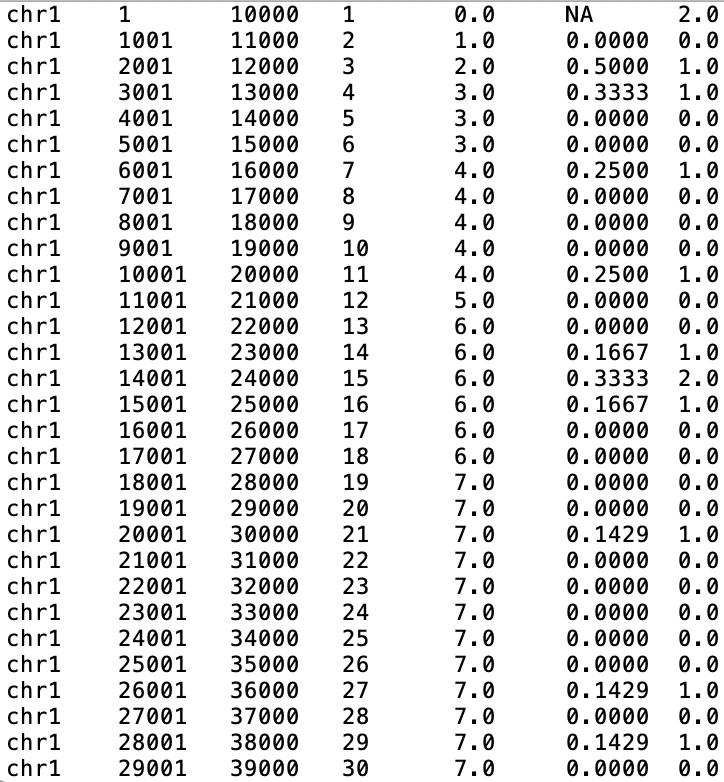
-Mode: (Adjacent/Sliding) the type of bin windows you want

**The output is like below**

Adjacent Bin (window\_size=100000, no overlapped adjacent window)



Sliding Bin (window\_size=10000, SlidingStep=1000)



**The other option**

-StepLength: When mode is Sliding you must set sliding step

-Normalize\_ByFiber: The path to xxx\_fiber.bed file generated by Patent\_GetSegment.jar used to normalize the signal by mapped fiber depth in local bins

**The example (Quick start)**

Java -jar Patent\_GetSegment.jar **-WindowSize 10000 -StepLength 1000**

**-TXT /Users/wwang/Desktop/test/1905async.txt**

**-o /Users/wwang/Desktop/test/**

**-Mode Sliding**

**-Chr\_Length /Users/wwang/Desktop/enrichment/ChrLength\_hg19.txt**

**-Normalize\_ByFiber /Users/wwang/Desktop/test/1905async\_fiber.bed**

1. Generate GTF

The input data needs .txt file got by AllRawDataRefining\_1R2G.jar or AllRawDataRefining\_1G2R.jar. It used to generate gtf like files to record genomic positions of all red signals and ends position of fiber. The must have parameter is -I, -O,-SampleName

-I: The input txt file’s parent path

-O: The output gtf file’s parent path

-SampleName: the name of all txt file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.

-Contain: The option could only be T when it show up, the default value is false when it is missing. It used to filter out the fiber without signal.

**The example (Quick start)**

java -jar GenerateGTF\_ByAllDataRefining\_Reformat.jar

-I /Volumes/WWT/Reformat/TXT/1802/

-O /Users/wwang/Desktop/ORM/

-SampleName 1802.0a.txt,1802.0b.txt

-Contain T

Notice: The red option could be removed based on clients’ requirement.

1. GetFiberCoordinate\_ByTXT

The input data needs .txt file got by AllRawDataRefining\_1R2G.jar or AllRawDataRefining\_1G2R.jar. It used to generate fiber’s bed files. The must have parameter is -I, -O,-SampleName

-I: The input txt file’s parent path

-O: The output file’s parent path

-SampleName: the name of all txt file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.

**The example (Quick start)**

java -jar GetFiberCoordinate\_ByTXT.jar

-I /Volumes/WWT/Final-Version/1807\_1802\_1905\_1708\_Merge/TXT/AlldataRefining/0min/

-O /Users/wwang/Desktop/ORM/GetFiberCoordinate\_ByTXT/Test

-SampleName 1905.FC0.txt,1905.FC1.txt

1. GetRedflagNumberInSlidingWindow\_ByAllDataRefining

The input data needs .txt file got by AllRawDataRefining\_1R2G.jar or AllRawDataRefining\_1G2R.jar. It used to generate fiber’s bed files. The must have parameter is -I, -O, -SampleName

-I: The input txt file’s parent path

-O: The output file’s parent path

-SampleName: the name of all txt file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.

-REF: The input REF file Unit kb, the example like below

chr1 249250621

chr2 243199373

chr3 198022430

…

Please pay attention, here the chromosome name must be same as the chromosome names show up in .txt files even including uppercase and lowercase letters Chr and chr is different, sometrimes people use chr23 and chr24 to represent chrX and chrY in .txt file. So corresponding chromosome name in this REF file should be chr23 and chr 24 instead of chrX and chrY

-WS: window size, Unit kb

-SS: SlidingStep, Unit kb

**The example (Quick start)**

java -jar GetRedflagNumberInSlidingWindow\_ByAllDataRefining.jar

-REF /Users/wwang/Desktop/enrichment/ChrLength\_hg19.txt

-WS 10

-SS 1

-I /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-SampleName 1905.FC1.txt,1905.FC2.txt

-O /Users/wwang/Desktop/ORM/GetRedflagNumberInSlidingWindow\_ByAllDataRefining/Test

**Output format**

Chr Bin\_start Bin\_end Bin\_ID RedsignalNumber FiberNumber FireEfficiency FiberwithoutSignal  
  
FireEfficiency = RedsignalNumber / FiberNumber , it is a value after fiber depth normalization to estimate signal showing up probability

1. GetRedflagNumberInAdjacentWindow\_ByAllDataRefining

The input data needs .txt file got by AllRawDataRefining\_1R2G.jar or AllRawDataRefining\_1G2R.jar. It used to generate fiber’s bed files. The must have parameter is -I, -O,-SampleName

-I: The input txt file’s parent path

-O: The output file’s parent path

-SampleName: the name of all txt file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.

-REF: The input REF file Unit kb, the example like below

chr1 249250621

chr2 243199373

chr3 198022430

…

Please pay attention, here the chromosome name must be same as the chromosome names show up in .txt files even including uppercase and lowercase letters Chr and chr is different, sometrimes people use chr23 and chr24 to represent chrX and chrY in .txt file. So corresponding chromosome name in this REF file should be chr23 and chr 24 instead of chrX and chrY

-WS: window size, Unit kb

**The example (Quick start)**

java -jar GetRedflagNumberInSlidingWindow\_ByAllDataRefining.jar

-REF /Users/wwang/Desktop/enrichment/ChrLength\_hg19.txt

-WS 10

-I /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-SampleName 1905.FC1.txt,1905.FC2.txt

-O /Users/wwang/Desktop/ORM/GetRedflagNumberInSlidingWindow\_ByAllDataRefining/Test

**Output format**

Chr Bin\_start Bin\_end Bin\_ID RedsignalNumber FiberNumber FireEfficiency FiberwithoutSignal  
  
FireEfficiency = RedsignalNumber / FiberNumber , it is a value after fiber depth normalization to estimate signal showing up probability

1. GetNewSegmentation\_AddSoloSignal

-I: The input txt file’s parent path

-O: The output file’s parent path

-SampleName: the name of all txt file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.

-GetSignalDistance: if it is “T” the script will generate a document record all the signal internal distance, the default value is false.

-FC: The first cutoff values to cluster the closet signal. Multi cutoff values for different files could be separate by “,” User point a cutoff value for internal signal distance and all signals within this value will cluster into segemtation

-GetSegmentDistance: if it is “T” the script will generate a document record all the signal internal distance, the default value is false. And you should put the -FC options in front of -GetSegmentDistance. Because -FC is a must have options for GetSegmentDistance.

-SC: The second cutoff values to cluster the closet segmentation. Multi cutoff values for different files could be separate by “,” User point a cutoff value for internal segmentation distance and all segmentations within this value will cluster into one segmentation and the script will generate the final bed files

"-AddSolo" : If it is “T”, the final result will contain all single signal coordinate without clustering. The default is false;

**The example (Quick start)**

java -jar GetNewSegmentation\_AddSoloSignal.jar

-GetSignalDistance

-I /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-SampleName 1905.FC1.txt,1905.FC2.txt

-O /Users/wwang/Desktop/ORM/GetNewSegmentation\_AddSoloSignal/Test

java -jar GetNewSegmentation\_AddSoloSignal.jar

-FC 16384,16384

-GetSegmentDistance

-I /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-SampleName 1905.FC1.txt,1905.FC2.txt

-O /Users/wwang/Desktop/ORM/GetNewSegmentation\_AddSoloSignal/Test

java -jar GetNewSegmentation\_AddSoloSignal.jar

-FC 16384,16384

-SC 32768,32768

-I /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-SampleName 1905.FC1.txt,1905.FC2.txt

-O /Users/wwang/Desktop/ORM/GetNewSegmentation\_AddSoloSignal/Test

(-AddSolo T)

1. GMM

This is an R package used for drawing gaussian mixture model distributions. For better setting the cutoff value to cluster signal and primary segment, client need to do the gaussian mixture model classification. Most of time we classify into 3 distribution and choose the tail of first distribution as the cutoff value. The distance between signals and distance between primary segments comes from jar package GetNewSegmentation\_AddSoloSignal.

The user need to set the Input\_ParentPath, Output\_ParentPath, Names of all samples and the x range for the final plot.

1. Add\_FDI\_ToSegment

-FDI\_N : The cutoff value for filtering the segmentation with negative FDI bigger than this cutoff

-FDI\_P : The cutoff value for filtering the segmentation with postive FDI smaller than this cutoff

-Signal\_N：The cutoff value for filtering the segmentation with negative FDI containing signals smaller than the cutoff

-Signal\_P：The cutoff value for filtering the segmentation with postive FDI containing signals smaller than the cutoff

-I-T: The input txt file’s parent path

-I-B: The input bed file’s parent path

-SampleName: the name of all input file, multi files will be separated by “,” and the name of txt file can’t contain character “\_” since it has been used as delimiter in script.Please pay attension the .txt and .bed should show up in pairs. The .bed file is generated by script GetNewSegmentation\_AddSoloSignal. And the name of the pairs should be same besides suffix, for example Sample1.txt and Sample1.bed

**The example (Quick start)**

Java -jar Add\_FDI\_ToSegment.jar

-FDI\_N -0.6

-FDI\_P 0.8

-Signal\_N 3

-Signal\_P 3

-I-T /Volumes/WWT/Final-Version/1905/TXT/AllDataRefining/

-I-B /Volumes/WWT/Final-Version/1905/Bed/

-SampleName 1905.FC1,1905.FC2

-O /Users/wwang/Desktop/ORM/Add\_FDI\_ToSegment/Test/

Function for DNA replication field

1. . GetS50Timing

S50, defined as the fraction of the S phase (0<S50<1) at which 50% of DNA is replicated in a defined region, was computed by linear interpolation of enrichment values in the six S phase compartments. It can reflect the temporal order of replication in some specific genomic position. And if you want to run this package you need to provide S50 timing along the genome.([REF](https://www.nature.com/articles/ncomms10208#Sec14))  
  
-I the parent path of the bed file you want to annotate S50 value, please pay attention the input bed file should be sort by chromosome order, ChrX and ChrY will be changed into Chr23,Chr24

-O the parent path of the output file

-S50 the parent path of the document containing all S50

-SampleName: the name of all input file, multi files will be separated by “,” and the name of bed file can’t contain character “\_” since it has been used as delimiter in script.

**The example (Quick start)**

Java -jar Add\_FDI\_ToSegment.jar   
 -I /Volumes/WWT/Final-Version/1905/Bed/sort   
 -O /Users/wwang/Desktop/ORM/GetS50Timing/Test   
 -S50 /Users/wwang/Desktop/cluster/S50\_1kbSlide\_100kbWindow/   
 -SampleName 1905.FC0.bed,1905.FC1.bed

1. Add\_DeltaRFD\_ToSegment

In DNA replication field RFD curve is used to call initial zone by OKseq method, it is calculated by number of Okazaki fragments to left(L) and right(R), RFD=(L-R)/(L+R). We calculate RFD in 1kb bin and in typical replication origin region, RFD curves will be in a sharp increasing trend , so the difference corresponding RFD value in end and start of origin region will be a positive value, we call it Delta\_RFD the bigger the Delta\_RFD is the region will be more likely to be an real replication origin ([REF](https://www.nature.com/articles/ncomms10208#Sec14))  
  
-I the parent path of the bed file you want to annotate Delta\_RFD value, please pay attention the input bed file should be sort by chromosome order, ChrX and ChrY will be changed into Chr23,Chr24

-O the parent path of the output file

-RFD the parent path of the document containing all RFD in 1kb bin along the genome

-SampleName: the name of all input file, multi files will be separated by “,” and the name of bed file can’t contain character “\_” since it has been used as delimiter in script.

**The example (Quick start)**

Java -jar Add\_DeltaRFD\_ToSegment.jar

-SampleName 1905.FC0.bed,1905.FC1.bed

-O /Users/wwang/Desktop/ORM/Add\_DeltaRFD\_ToSegment/Test

-I /Volumes/WWT/Final-Version/1905/Bed/sort

-RFD /Users/wwang/Desktop/Final-Version/ForkDirection/AddFDItoBed/RFD\_InWindow/output/1kb/NewRFD\_1kb\_sort.bedgraph

1. Calculate\_FDI\_RFD  
   Similar to RFD, the FDI\_RFD is also calculated by formula (L-R)/(L+R), but L and R is the segments with positive and negative FDI within 1kb bin. Firstly, the segments bed file should be treated by jar package Add\_FDI\_ToSegment, And then we will map these regions to 1kb adjacent bins along the genome. Only if the bin is overlapped with the region I will add 1 to the number of positive or negative FDI (L or R) based on the value of the segmentation. Finally, all bins will have their own L and R values. I will calculate the FDI\_RFD for all bins.

-REF : The input REF file Unit kb, the example like below

chr1 249250621

chr2 243199373

chr3 198022430

…

Please pay attention, here the chromosome name must be same as the chromosome names show up in .txt files even including uppercase and lowercase letters Chr and chr is different, sometrimes people use chr23 and chr24 to represent chrX and chrY in .txt file. So corresponding chromosome name in this REF file should be chr23 and chr 24 instead of chrX and chrY

-WS : window size,length of bin along the genome, unit kb

-I : The parent path for multi bed files with FDI value or the direct path for single bed file with FDI

-O : The parent path for multi results with FDI\_RFD or the direct path for single bed file with FDI\_RFD

-Colum\_FDI : the Colum number of input file’s FDI

-SignalColumn : the Colum number of input file’s signal number, if the bed file is generated by jar package GetNewSegmentation\_AddSoloSignal there should be one column is signal number. If the bed file has this column, user can set the cutoff for filtering segmentation with few signals to reduce the false positive value the suggested cutoff is 3. The default value is 2, if not set.

-FDI\_N\_Cutoff: The cutoff value for filtering the segmentation with negative FDI bigger than this cutoff

-FDI\_P\_Cutoff : The cutoff value for filtering the segmentation with postive FDI smaller

than this cutoff, the default will be zero if not set

-Signal\_N：The cutoff value for filtering the segmentation with negative FDI containing

signals smaller than the cutoff, the default will be zero if not set

**The example (Quick start)**

Single bed input with FDI cutoff setting

Java -jar -REF  /Users/wwang/Desktop/enrichment/ChrLength\_hg19.txt

-WS 1

-I /Users/wwang/Desktop/ORM/FDI\_RFD/Test/All0min.bed

-O /Users/wwang/Desktop/ORM/FDI\_RFD/Test/Output.bed

-Colum\_FDI 7

-SignalColumn 5

-SignalCutoff 2

-FDI\_P\_Cutoff 1

-FDI\_N\_Cutoff -1

Multi bed input with default FDI cutoff

Java -jar -REF  /Users/wwang/Desktop/enrichment/ChrLength\_hg19.txt

-WS 1

-I /Users/wwang/Desktop/ORM/FDI\_RFD/Test/

-O /Users/wwang/Desktop/ORM/FDI\_RFD/Test/Output.bed

-Colum\_FDI 7

-SignalColumn 5

-SignalCutoff 2

-SampleName 1905.FC0.bed,1905.FC1.bed