GenomeCluster Tutorial

GenomeCluster requires information about the islands of Ns present in the genome sequence to analyze. To this end, you can use the python script N.py:

```
python N.py -i <input FASTA or multiFASTA file> -o <output BED file>
```

GenomeCluster also requires that the terminal commands *'awk'*, *'sort'* and *'cut'* are available. These commands are usually available on Linux systems. To run the script on *Windows* you must first check if these commands are in the path.

By running *GenomeCluster* without parameters you obtain an explanation of the required parameters:

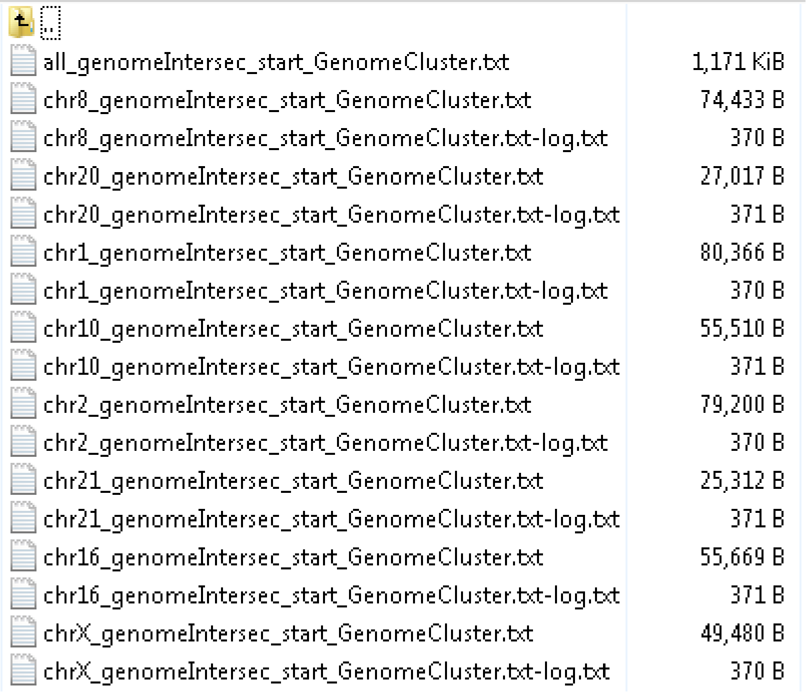
```
$/GenomeCluster$ perl GenomeCluster.pl

---------------------------------------------------------------------------
---------------------------------------------------------------------------
---------                                                         ---------
---------     Computational Genomics and Bioinformatics Group     ---------
---------        University of Granada, Dept. of Genetics         ---------
---------                                                         ---------
---------               Web: http://bioinfo2.ugr.es               ---------
---------        CGI: http://bioinfo2.ugr.es/GenomeCluster        ---------
---------                                                         ---------
---------              GenomeCluster (1.0) 11/30/13               ---------
---------                                                         ---------
---------------------------------------------------------------------------
---------------------------------------------------------------------------

Example of use:

perl GenomeCluster.pl <cMethod> <BED> <d> <P-value> [<assembly> [<N_BED>
[<maxN>]]]

cMethod:   Clustering Method. Type "element", "start", "middle" or "end"
             in order to select the method to find clusters: taking into
             account the whole elements or the start, middle or end
             coordinates of each element, respectively.

BED:       File input in BED format. Input elements do not need to be sorted
             nor merged. This program will sort, merge and select the input
             depending on the arguments.

d:         The threshold distance on basis of a given percentile.
             For example: d=25 calculates the percentile 25 of the genomic
             distance distribution and takes this value as the threshold
             distance.
             The recommended value is 50 (median distance).
             You can add multiple comma-separated percentile values, "ci"
             (chromosome intersection) or "gi" (genome intersection).
             Example: gi,25,60,ci,50

P-value:   The maximal P-value under which an element group is considered
             as a cluster. The recommended limit is 1E-5.

assembly:  Directory containing sequence files in FASTA format. If none is
             provided the program will estimate several parameters.

N_BED:     File containing coordinates of N blocks in BED format. If none
             is provided it will assume that sequences do not contain any N.

maxN:      Maximum number of Ns allowed. Default is 0.
```

##

A real example:

```
perl GenomeCluster.pl start snp137Common_hg19.bed gi 1e-5 /opt/seq/hg19/fasta
hg19_n.bed
```

The results are then obtained in the directory ‘snp137Common_hg19.bed_GCresult’:



```
..
all_genomeIntersec_start_GenomeCluster.txt                1,171 KiB
chr8_genomeIntersec_start_GenomeCluster.txt                74,433 B
chr8_genomeIntersec_start_GenomeCluster.txt-log.txt           370 B
chr20_genomeIntersec_start_GenomeCluster.txt               27,017 B
chr20_genomeIntersec_start_GenomeCluster.txt-log.txt          371 B
chr1_genomeIntersec_start_GenomeCluster.txt                80,366 B
chr1_genomeIntersec_start_GenomeCluster.txt-log.txt           370 B
chr10_genomeIntersec_start_GenomeCluster.txt               55,510 B
chr10_genomeIntersec_start_GenomeCluster.txt-log.txt          371 B
chr2_genomeIntersec_start_GenomeCluster.txt                79,200 B
chr2_genomeIntersec_start_GenomeCluster.txt-log.txt           370 B
chr21_genomeIntersec_start_GenomeCluster.txt               25,312 B
chr21_genomeIntersec_start_GenomeCluster.txt-log.txt          371 B
chr16_genomeIntersec_start_GenomeCluster.txt               55,669 B
chr16_genomeIntersec_start_GenomeCluster.txt-log.txt          371 B
chrX_genomeIntersec_start_GenomeCluster.txt                49,480 B
chrX_genomeIntersec_start_GenomeCluster.txt-log.txt           370 B
```