Fantasio

Version 0.1

Isuru HAUPE & Marie MICHEL

2018-08-31

Contents

Introduction

Fantasio is composed of several functions. Its goals are:

- For population genetic studies: estimating and detecting inbreeding on individuals without known genealogy, estimating the population proportion of mating types and the individual probability to be offspring of different mating types
- For rare disease studies: performing homozygosity mapping with heterogeneity
- For multifactorial disease studies: HBD-GWAS strategy

Fantasio implements the creation of several random sparse submaps on genome-wide data (to remove linkage disequilibrium). It also provides graphical outputs to facilitate interpretations of homozygosity mapping results and plots.

In this vignette, we illustrate how to use the package, using the data set HGDP-CEPH, a ship which contains 1043 individuals and 660918 markers. In order to access the data, you will need to download the HGDP.CEPH package (see below how) and load it.

Not all options of the functions are described here, but rather their basic usage. The reader is advised to look at the manual page of the function for details.

Principal concepts

Fantasio implements a maximum likelihood method that uses a hidden Markov chain to model the dependencies along the genome between the (observed) marker genotypes of an individual, and its (unobserved) homozygous by descent (HBD) status. The emission probabilities of this hidden Markov model (HMM) depend on the allele frequencies. The transition probabilities depend on the genetic distance between two adjacent markers. This model allows estimating the inbreeding coefficient f of an individual, and a parameter a, where af is the instantaneous rate of change per unit map length (here cM) from no HBD to HBD. Both HBD and non-HBD segment lengths are assumed to be distributed exponentially with mean lengths $\frac{1}{a(1-f)}$ and $\frac{1}{af}$, respectively.

THe method requires the markers to be in minimal linkage disequilibrium (LD). Otherwise biased estimations of f are produced. A strategy consisting of generating multiple random sparse genome maps (submaps) has been proposed to avoid this bias (Leutenegger et al. 2011). When several submaps are considered, f is estimated as the median value on all the f estimation obtained on the different maps after removing submaps with $a > 1$ (having an average HBD segment length of 1 cM is unlikely to be detected with a SNP density of 1 per 0.5 cM). This strategy has the advantage of not requiring any LD computation on the sample and of minimizing loss of information, as compared with a strategy that based on a single map of markers in minimal LD.

Fantasio statistical framework allows fixing HMM parameters to compute the likelihood of a mating type. These likelihoods can be used for:

- Inferring an individual as inbred by comparing the maximized likelihood with the one to be outbred with a likelihood ratio test
- Estimating the population proportion of mating types
- Estimating the individual probability to be into different mating types

When multiple submaps are used, the median p-values/probabilities are considered. See Leutenegger et al. 2011 for more details on the calculations.

Homozygosity mapping (Lander and Botstein 1987) consists in focusing on inbred affected individuals and searching for a region of the genome of shared homozygosity. Original homozygosity mapping requires that the genealogy of patients be known so that inbred patients can be identified and their respective f estimated. Leutenegger et al. (Leutenegger et al. 2006) proposed using the f estimated on genome-wide genetic data to compute a FLOD score, similar to Morton's LOD score (Morton 1955).

Genin et al. (Genin et al. 2012) adapted the FLOD formula for multiple submaps. FLOD(i)(m,s) is computed for each individual i, each marker m on each submap s, using the equation (1):

$$FLOD^{(i)}(m,s) = log_{10}\frac{P\left(Y_{m,s}^{(i)}|H_1\right)}{P\left(Y_{m,s}^{(i)}|H_0\right)} = log_{10}\frac{P\left(X_{m,s}^{(i)}=1|Y_{m,s}^{(i)}\right)+q.P\left(X_{m,s}^{(i)}=0|Y_{m,s}^{(i)}\right)}{\hat{f}_s^{(i)}+q.\left(1-\hat{f}_s^{(i)}\right)}$$

With the following parameters :

- $Y_{m,s}^{(i)}$ the observed genotype of individual i at marker m on submap s
- H_1 the hypothesis where marker m is linked to the disease, and H_0 the one where it is not
- $X_{m,s}^{(i)}$ the HBD status of individual i at marker m on submap s that is estimated together with the inbreeding coefficient using the HMM of the package
- $\hat{f}_s^{(i)}$ the estimated inbreeding coefficient of individual i on submap s
- q the assumed frequency of the mutation involved in the disease for this individual.

Results are then averaged over the different submaps to obtain a single $FLOD^{(i)}(m)$ at each marker m.

Genin et al. (Genin et al. 2012) proposed to detect fully penetrant rare recessive variants by performing homozygosity mapping on inbred cases from Genome-Wide Association Study (GWAS) data. Linkage evidence is then evaluated over the entire set I of inbred cases by computing a FLOD score, HFLOD(m,α), at each marker m, in presence of genetic heterogeneity using a parameter α (2):

$$HFLOD(m,\alpha) = \sum log_{10}\left[\alpha.\frac{P\left(Y_{m,s}^{(i)}|H_1\right)}{P\left(Y_{m,s}^{(i)}|H_0\right)}+(1-\alpha)\right] = \sum log_{10}\left[\alpha.exp\left(FLOD^{(i)}(m)*log(10)\right)+(1-\alpha)\right]$$

This heterogeneity score is then maximized over α to evaluate the evidence of linkage at marker m where α is the estimate of the proportion of cases linked to this locus (3):

$$HFLOD(m) = max_{\alpha}(HFLOD(m, \alpha))$$

1. Getting started

The first thing you should know is that the package Fantasio depends on package gaston, please make sure to have it installed.

Please refer to the vignette of this package for more information.

Since we explained the concept behind the package let's make a usage example of it. For this we will use the data provided in the package HGDP-CEPH.

1.1 Installation

First and faremost install the package with the following command :

```
install.packages("Fantasio")
```

After doing that we will need to run the following commands :

```
require(Fantasio)
```

```
## Loading required package: Fantasio

## Loading required package: methods

## Loading required package: parallel

## Loading required package: gaston

## Loading required package: Rcpp

## Loading required package: RcppParallel

##
## Attaching package: 'RcppParallel'

## The following object is masked from 'package:Rcpp':
##
##     LdFlags

## Gaston set number of threads to 8. Use setThreadOptions() to modify this.

##
## Attaching package: 'gaston'

## The following object is masked from 'package:stats':
##
##     sigma

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

1.2 Input HGDP-CEPH data file

```
install.packages("HGDP.CEPH", repos="https://genostats.github.io/R/")
```

```
require(HGDP.CEPH)
```

```
## Loading required package: HGDP.CEPH
```

From now on, we can use the package.

```
filepath <-system.file("extdata", "hgdp_ceph.bed", package="HGDP.CEPH")
```

1.3 Creation of the bed matrix

Let us first create a bed.matrix object (see gaston package for details) for the data file we loaded from the package HGDP.CEPH with this command :

```
x <- read.bed.matrix(filepath)
```

```
## Reading /ext/home/haupe/R/x86_64-pc-linux-gnu-library/3.4/HGDP.CEPH/extdata/hgdp_ceph.rds
## Reading /ext/home/haupe/R/x86_64-pc-linux-gnu-library/3.4/HGDP.CEPH/extdata/hgdp_ceph.bed
```

This command returns an updated 'bed.matrix' object (refer to gaston vignette for more informations and function documentation) :

```
x <- set.stats(x)
```

```
## ped stats and snps stats have been set.
## 'p' has been set.
## 'mu' and 'sigma' have been set.
```

Here we only want to work on the Bedouin's population, so we selected this population with the following command :

```
x.me <- select.inds(x, population == "Bedouin")
```

Please refer to the manual function of read.bed.matrix, set.stats and select.inds if needed (package gaston).

By default, the package only computes HBD, FLOD scores and HFLOD scores for affected individuals (phenotype = 2). So make sure that affected individuals have status 2. In the case of the HGDP-CEPH data, there is no information on phenotypes so phenotype is 1. We show below how deal with this situation.

You can insure that your bed.matrix object is created and have the data needed with :

```
str(x.me)
```

```
## Formal class 'bed.matrix' [package "gaston"] with 8 slots
##   ..@ ped                 :'data.frame': 48 obs. of  34 variables:
##   .. ..$ famid      : chr [1:48] "HGDP00607" "HGDP00608" "HGDP00609" "HGDP00610" ...
##   .. ..$ id         : chr [1:48] "HGDP00607" "HGDP00608" "HGDP00609" "HGDP00610" ...
##   .. ..$ father     : int [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ mother     : int [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ sex        : int [1:48] 2 1 1 1 1 2 2 2 2 1 ...
##   .. ..$ pheno      : int [1:48] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..$ population : Factor w/ 57 levels "Adygei","Balochi",..: 10 10 10 10 10 10 10 10 10 10 ...
##   .. ..$ region     : Factor w/ 27 levels "Algeria (Mzab)",..: 12 12 12 12 12 12 12 12 12 12 ...
##   .. ..$ region7    : Factor w/ 7 levels "Africa","America",..: 6 6 6 6 6 6 6 6 6 6 ...
##   .. ..$ H952       : logi [1:48] TRUE TRUE TRUE TRUE TRUE TRUE ...
##   .. ..$ N0         : int [1:48] 56890 58576 58375 60261 53403 61180 54121 55903 61857 62151 ...
##   .. ..$ N1         : int [1:48] 192705 189349 190265 186828 203585 183119 199168 196327 183012 18074
##   .. ..$ N2         : int [1:48] 394377 395838 394286 396726 386823 399508 390610 391298 398496 40115
##   .. ..$ NAs        : int [1:48] 301 510 1347 458 462 466 374 745 908 222 ...
##   .. ..$ N0.x       : int [1:48] 1514 3766 4313 4160 4434 1476 1491 1540 2261 4107 ...
##   .. ..$ N1.x       : int [1:48] 4936 0 0 0 0 4847 5354 5112 3243 0 ...
##   .. ..$ N2.x       : int [1:48] 10017 12688 12136 12289 12008 10140 9599 9810 10953 12361 ...
##   .. ..$ NAs.x      : int [1:48] 5 18 23 23 30 9 28 10 15 4 ...
##   .. ..$ N0.y       : int [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ N1.y       : int [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ N2.y       : int [1:48] 0 10 10 10 10 0 0 0 0 10 ...
##   .. ..$ NAs.y      : int [1:48] 10 0 0 0 0 10 10 10 10 0 ...
##   .. ..$ N0.mt      : int [1:48] 4 9 19 5 9 9 4 5 19 1 ...
##   .. ..$ N1.mt      : int [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ N2.mt      : int [1:48] 157 143 141 157 153 151 157 153 140 161 ...
##   .. ..$ NAs.mt     : int [1:48] 2 11 3 1 1 3 2 5 4 1 ...
##   .. ..$ callrate   : num [1:48] 1 0.999 0.998 0.999 0.999 ...
##   .. ..$ hz         : num [1:48] 0.299 0.294 0.296 0.29 0.316 ...
##   .. ..$ callrate.x : num [1:48] 1 0.999 0.999 0.999 0.998 ...
##   .. ..$ hz.x       : num [1:48] 0.3 0 0 0 0 ...
##   .. ..$ callrate.y : num [1:48] 0 1 1 1 1 0 0 0 0 1 ...
##   .. ..$ hz.y       : num [1:48] NaN 0 0 0 0 NaN NaN NaN NaN 0 ...
##   .. ..$ callrate.mt: num [1:48] 0.988 0.933 0.982 0.994 0.994 ...
##   .. ..$ hz.mt      : num [1:48] 0 0 0 0 0 0 0 0 0 0 ...
##   ..@ snps                :'data.frame': 660918 obs. of  17 variables:
##   .. ..$ chr     : int [1:660918] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..$ id      : chr [1:660918] "rs3094315" "rs12562034" "rs3934834" "rs9442372" ...
##   .. ..$ dist    : num [1:660918] 0.0916 0.0992 0.4963 0.5039 0.508 ...
##   .. ..$ pos     : int [1:660918] 742429 758311 995669 1008567 1011278 1011521 1011558 1020428 102140
##   .. ..$ A1      : chr [1:660918] "C" "A" "T" "A" ...
##   .. ..$ A2      : chr [1:660918] "T" "G" "C" "G" ...
##   .. ..$ N0      : int [1:660918] 3 4 6 16 3 0 3 5 8 3 ...
##   .. ..$ N1      : int [1:660918] 23 15 15 21 16 1 21 13 18 10 ...
##   .. ..$ N2      : int [1:660918] 22 29 27 11 29 47 24 30 22 35 ...
##   .. ..$ NAs     : int [1:660918] 0 0 0 0 0 0 0 0 0 0 ...
##   .. ..$ N0.f    : int [1:660918] NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ N1.f    : int [1:660918] NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ N2.f    : int [1:660918] NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ NAs.f   : int [1:660918] NA NA NA NA NA NA NA NA NA NA ...
##   .. ..$ callrate: num [1:660918] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..$ maf     : num [1:660918] 0.302 0.24 0.281 0.448 0.229 ...
##   .. ..$ hz      : num [1:660918] 0.479 0.312 0.312 0.438 0.333 ...
##   ..@ bed                  :<externalptr>
##   ..@ p                    : num [1:660918] 0.698 0.76 0.719 0.448 0.771 ...
##   ..@ mu                   : num [1:660918] 1.396 1.521 1.438 0.896 1.542 ...
##   ..@ sigma                : num [1:660918] 0.61 0.652 0.712 0.751 0.617 ...
##   ..@ standardize_p        : logi FALSE
##   ..@ standardize_mu_sigma: logi FALSE
```

This object contains two slots :

- ped : which gives you information about all the individuals in the data
- snps : which gives you information about the snps themselves

More information in the vignette of the gaston package.

2. Running Fantasio

We created a wrapper to make the usage of the package more simple.

We implemented in the package two differents methods in order to create n submaps :

- By "Hotspots" : with this method we use a file of recombination hotspots (downloaded from the HapMap website in hg17 (*)) to segment the genome. Segments should contain at least number_of_marker markers. Markers are then randomly selected within each segment along the genome. By doing this process we obtain a submap (a list of marker). The recombination hotspots have been converted to other buid (hg18, hg19) using hgLiftOver. The default recombination hotspots file used is in hg19. (*) http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/2006-10_rel21_phaseI+II/hotspots
- By "Distance" : with this method we use a fix step based on genetic or physiscal distance (0.5 cM by default) to pick a marker randomly along the genome. More technically, segments are created whenever there is a gap larger than the step (0.5 cM) between adjacent markers. Each segment is then subdivided in several mini-segments. By default we create 20 mini-segments, each containing at least 50 markers. If this is not possible (not enough markers), we do not create mini-segments. After this process is done, we loop over the mini-segments, pick a random marker and walk through the mini-segments by picking the nearest marker after taking a step (default 0.5 cM) downstream and upstream the mini-segments.

The wrapper calls two different functions : `createSegmentsListBySnps` and `createSegmentsListBySnps`. The first function `createSegmentsListBySnps` is used to create a list of segments though the genome. The second function `makeAllSubmapsBySnps` or `makeAllSubsmapsbyHotspots` is used to create the submaps.

2.1 Hotspots (Default)

By default, the submaps are created using the file of recombination hotspots and summarizing the results for each snp that appears in a submap.

```
F1 <- Fantasio(bedmatrix=x.me, segments="Hotspots", n=5, list.id = "all")
```

We require that at least n.consecutive.marker markers are HBD before calling a HBD segment. Default value for n.consecutive.marker=5.

Here we need to use argument list.id = "all" because we do not have phenotype information for the HGDP-CEPH individuals. By default, Fantasio focuses on affected individuals only (pheno = 2).

2.2 Hotspots by Segments

For the "Hotspots" method, the results can also be summarized globally for each segment using option recap.by.segments=TRUE. In that case, n.consecutive.marker should be set to 1.

```
F2 <- Fantasio(bedmatrix=x.me, segments="Hotspots", recap.by.segments=TRUE, n.consecutive.marker=1, n=5
```

2.3 Distance

```
F3 <- Fantasio(bedmatrix=x.me, segments="Distance", n=5, list.id = "all")
```

2.4 How to use the segment.option argument

In order to use the `segment.option` argument you need to pass a list of arguments, each variable name in the list must be an argument name in the function. The function that will be called is either `createsSegmentsListBySnps` if `segments` argument is equal to "Distance" or `createSegmentsListByHotspots` if `segments` argument is equal to "Hotspots" and the arguments list will be passed to it. So refer to these functions for possible arguments.

```
l <- list(number_of_marker=50) #default is 0
F1.l <- Fantasio(bedmatrix=x, segments="Hotspots", segment.options=l, list.id = "all")
```

In the case of "Hotspots", by default, we do not require to have a minimum number of markers in each segment (number_of_marker = 0). With the above command line, we impose to have at least 50 markers in a segment. Otherwise it is merged with the previous segment.

3. Step by step usage of the package Fantasio

3.1 Hotspots

3.1.1 Creation of the segments list

We will now create segments, which will be use to create the submaps later, further explication below, for now use this command :

```
s1 <- createSegmentsListByHotspots(x.me)
```

```
## You are currently using version hg19 of hotspot
## Gathering all hotspots for the genome : ......................
## Gathering all the genome's markers : ......................
## Finding which markers are between two hotspots : ......................
```

This function creates a list of chromosomes, in each, you have a list of several segments created thanks to the hotspots file given in argument (files are given with the package), in each segments you have SNPs index.

You can watch a summary of what was done with :

```
segmentsListSummary(s1)
```

```
##    chromosome number_of_segments number_of_markers
## 1           1                904             48532
## 2           2                895             52722
## 3           3                750             43507
## 4           4                702             39040
## 5           5                721             39933
## 6           6                719             42184
## 7           7                564             34754
## 8           8                589             36417
## 9           9                563             30276
## 10         10                669             33427
## 11         11                537             31255
## 12         12                601             31039
## 13         13                475             24595
## 14         14                401             20926
## 15         15                383             19098
## 16         16                429             19028
## 17         17                374             16052
## 18         18                442             19496
## 19         19                217             10397
## 20         20                384             16228
## 21         21                225              9301
## 22         22                207              9379
```

This function creates a dataframe with three colums :

- chromosome
- number_of_segments
- number_of_marker

3.1.2 Creation of the submaps and computation

We will now head toward the creation of submaps using the following commands :

```
F1 <- makeAllSubmapsByHotspots(x.me, 5, s1, verbose=FALSE, list.id = "all")
```

For the sake of clarity we have only created 5 submaps, but generally we do 100.

This function will creates 5 submaps, all the parameters can be modified (use args(makeAllSubmapsByHotspots) for more informations).

The variable submaps becomes an list.submaps object, you can watch the different elements of it with :

```
str(F1) #careful it can become huge depending on your data sizes
```

3.1.3 Descrition of the object F1

This object contains all the results of the different computations executed during the process of creating n submaps. Here is a complete description of each structure in this object :

- segments_list : the object Segments created previously (s1)

```r
str(F1@segments_list) #careful it can become huge depending on your data sizes
```

- atlas : the list of all the submaps created during the process. Each element of the list is a S4 object. Depending on the method you used the object can be either a `snsp.matrix` or a `hotspots.matrix` (here we use the hotspots method). Each submaps contains 15 slots :
  - submap : the index of each marker picked (index from the bed.matrix object)
  - ncol : the total number of marker picked
  - nrow : the number of individuals
  - ped : a dataframe with all the individuals' genotypes
  - map : a dataframe with all the SNP information
  - epsilon : genotyping error rate
  - delta.dist : distance between each marker in cM/bp
  - log.emiss : log of all the emission probabilities of the hidden Markov model
  - a : a matrix with all the a's estimation
  - f : a matrix with all the f's estimation
  - likelihood0 : a matrix with all the likehood under the null hypothesis ($f = 0$)
  - likelihood1 : a matrix with all the likehood under the inbred hypothesis ($f = 1$)
  - p.lrt : p value of the likelihood ratio test
  - HBD.prob : a matrix with all the HBD probabilities computed for each individual
  - FLOD : a matrix with all the FLOD score computed

```r
str(F1@atlas)
```

- likelihood_summary : a dataframe with all the likelihood0 and likelihood1 computed over the submaps.

```r
str(F1@likelihood_summary)
```

- estimation_summary : a dataframe with all the a and f computed over the submaps

```r
str(F1@estimation_summary)
```

- marker_summary : a dataframe, which gives the number of markers and the number of times it has been picked,
  - number_of_time_picked
  - number_of_marker

```r
str(F1@marker_summary)
```

- submaps_summary : a dataframe which gives several informations about the a and f computed over the submaps. The dataframe contains 13 columns:
  - FID: family identifier
  - IID: individual identifier
  - STATUS: status (1 non-affected, 2 affected, 0 unknown)
  - SUBMAPS: number of submaps used
  - QUALITY: percentage of valid submaps (i.e. submaps with $a < 1$)
  - F_MIN: minimum f on valid submaps
  - F_MAX: maximum f on valid submaps
  - F_MEAN: mean f on valid submaps

  - F_MEDIAN: median f on valid submaps (recommended to estimate f)
  - A_MEDIAN: median a on valid submaps (recommended to estimate a)
  - pLRT_MEDIAN: median p-value of LRT tests on valid submaps
  - INBRED: a flag indicating if the individual is inbred (pLRT_MEDIAN < 0.05) or not
  - pLRT_<0.05: number of valid submaps with a LRT (likelihood ratio test) having a p-value below 0.05

```
head(F1@submap_summary)
```

```
##         FID       IID STATUS SUBMAPS QUALITY       F_MIN       F_MAX
## 1 HGDP00607 HGDP00607      1   5 / 5     100 0.022681387 0.028548500
## 2 HGDP00608 HGDP00608      1   5 / 5     100 0.035159598 0.041289176
## 3 HGDP00609 HGDP00609      1   5 / 5     100 0.037669140 0.042819999
## 4 HGDP00610 HGDP00610      1   5 / 5     100 0.047400500 0.055147242
## 5 HGDP00611 HGDP00611      1   1 / 5      20 0.008446423 0.008446423
## 6 HGDP00612 HGDP00612      1   5 / 5     100 0.055547043 0.068661120
##        F_MEAN    F_MEDIAN   A_MEDIAN  pLRT_MEDIAN INBRED pLRT_inf_0.05
## 1 0.025676286 0.025423941 0.11517577 9.692524e-27   TRUE             5
## 2 0.037822117 0.036643094 0.06197893 1.934656e-58   TRUE             5
## 3 0.040338855 0.040665408 0.11392824 9.857859e-48   TRUE             5
## 4 0.052075003 0.053193213 0.16090315 4.148441e-60   TRUE             5
## 5 0.008446423 0.008446423 0.95653179 2.253991e-02   TRUE             1
## 6 0.061891615 0.061273186 0.30671225 6.920514e-44   TRUE             5
```

- bySegments : a boolean indicating whether the creation of summary statistics for HBD and FLOD has to be made by segments or not. By default, it is FALSE. The description below of HBD_recap and FLOD_recap depends on this choice.
- HBD_recap : a dataframe with individuals in row and marker positions in column. It contains the mean of the HBD probability at a marker over the submaps for each individual. If the marker only appears in one submap (no mean needed), then it is just the HBD proba for that marker in that submap.

```
F1@HBD_recap[1:10, 1:10] # an individual * marker matrix
```

- FLOD_recap : a dataframe like above. It contains the mean of the FLOD score at a marker over the submaps for each individual. Again if a marker only appears in one submap, no mean is needed.

```
F1@FLOD_recap[1:10, 1:10] # an individual * marker matrix
```

- HBD_segments : a list of dataframe, each datafram is for an individual and it contains a list of HBD segments (start and end positions). By default, a region is HBD if it contains at least 5 markers (n.consecutive.marker=5) with a HBD probability larger than 0.5 (threshold=0.5).

```
str(F1@HBD_segments[[1]])
```

- HFLOD : a dataframe with the value of HFLOD scores for every markers that appeared at least once in a submap. It uses the formula (2) with FLOD_recap to obtain HFLOD.

```
str(F1@HFLOD)
```

- bedmatrix : the bedmatrix object

```
str(F1@bedmatrix)
```

- unit : the unit of the marker (cM or Bp).

3.2 Hotspots by segments

We implemented a second inner method for the "Hotspots" method. The only paramater that changes is `recap.by.segments`, it is put to TRUE.

In the default "Hotspots" method the HBD probabilities and FLOD scores are computed for each marker (see above).

With the "Hotspots by segment" method, HBD probabilities and FLOD scores correspond to the mean of HBD probabilities and FLOD score of all the markers of a segment that have been selected in one of the sumaps. There is hence only one value for a segment delimited by hotspots. So we also recommend to set n.consecutive.marker to 1. Since results over multiple markers have already been averaged.

3.2.1 Creation of the segments list

We use the same segment list that is used before (s1).

3.2.2 Creation of the submaps and computation

As said before the only argument that changes is "recap.by.segments", it is put to TRUE. And we recommend to adjust "n.consecutive.marker".

```
F2 <- makeAllSubmapsByHotspots(x.me, 5, s1, verbose=FALSE, recap.by.segments = TRUE, n.consecutive.mark
```

3.3 Distance

3.3.1 Creation of the segments list

We will now create segments, which will be used to create the submaps later :

```
s3 <- createSegmentsListBySnps(x.me)
```

```
## Finding segments for the genome : .......................
## Finding which markers are between two segments: .......................
## Finding mini segments .......................
```

This function creates segments based on the markers available in the data and the gaps between them. The value of the gap is imposed by the minimun step required between markers. If 2 adjacent markers have an inter-distance larger than the step, then there is a gap between them. The first marker of the pair is the end of a segment and the second marker is the start of a new segment. The function creates an object which will contain three slots :

- gap : the value of the minimal fix step required between sampled markers of a submap
- unit : the unit of the markers ("cM" or "Bp")
- snps.segments : for each chromosome, the list of segments; each segment being a list of SNP indexes

You can watch a summary of what was done with :

```r
segmentsListSummary(s3)
```

This function creates a dataframe with three colums :

- chromosome
- number_of_segments
- number_of_marker

3.3.2 Creation of the submaps and computation

We will now head toward the creation of submaps using the following commands :

```r
F3 <- makeAllSubmapsBySnps(x.me, 5, s3, verbose=FALSE, list.id = "all")
```

The variable submaps becomes an list.submaps object, you can watch the different elements of it with :

```r
str(F3) #careful it can become huge depending on your data sizes
```

4. Parallelism with the package

We implemented a paralellism method to make the creation of the submaps more efficient. We paralellized the creation of the submaps, that is to say, the selection of markers. Make sure to have an environment that can support the usage of multiple CPU.

Use the `n.cores` argument, i.e the number of CPU that will be used to make the differents submaps in the following functions :

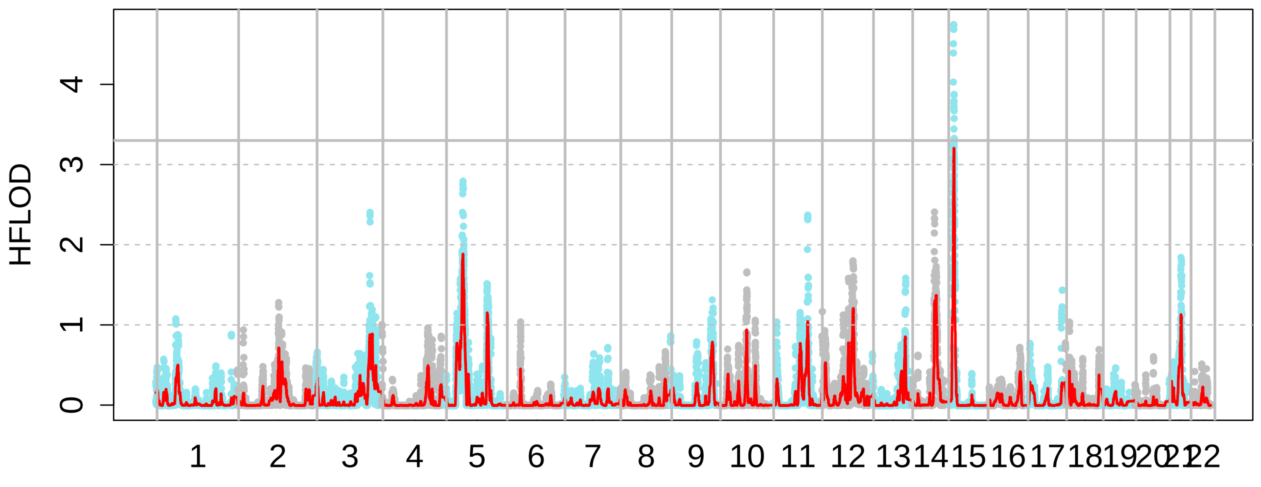
- Fantasio
- makeAllSubmapsByHotspots
- makeAllSubmapsBySnps

```r
F6 <- Fantasio(bedmatrix=x.me, segments="Hotspots", n=5, verbose=FALSE, n.cores=10, list.id = "all")
```

5. Plotting

5.1 HFLOD manhattan plot

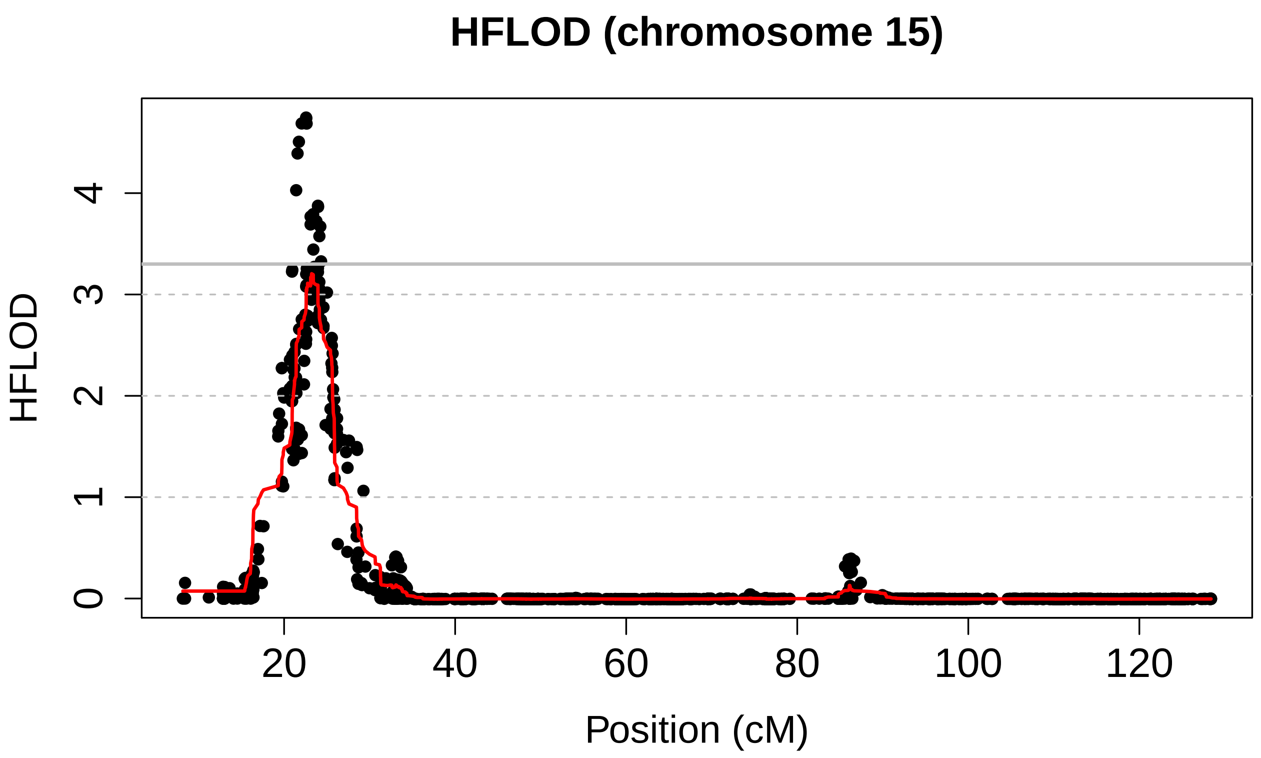
```r
HFLOD.manhattan.plot(F1)
```



- The red lines that you see is the value of the moving average (more information below).

5.2 HFLOD for a chromosome

```
HFLOD.plot.chr(F1, chr=15)
```

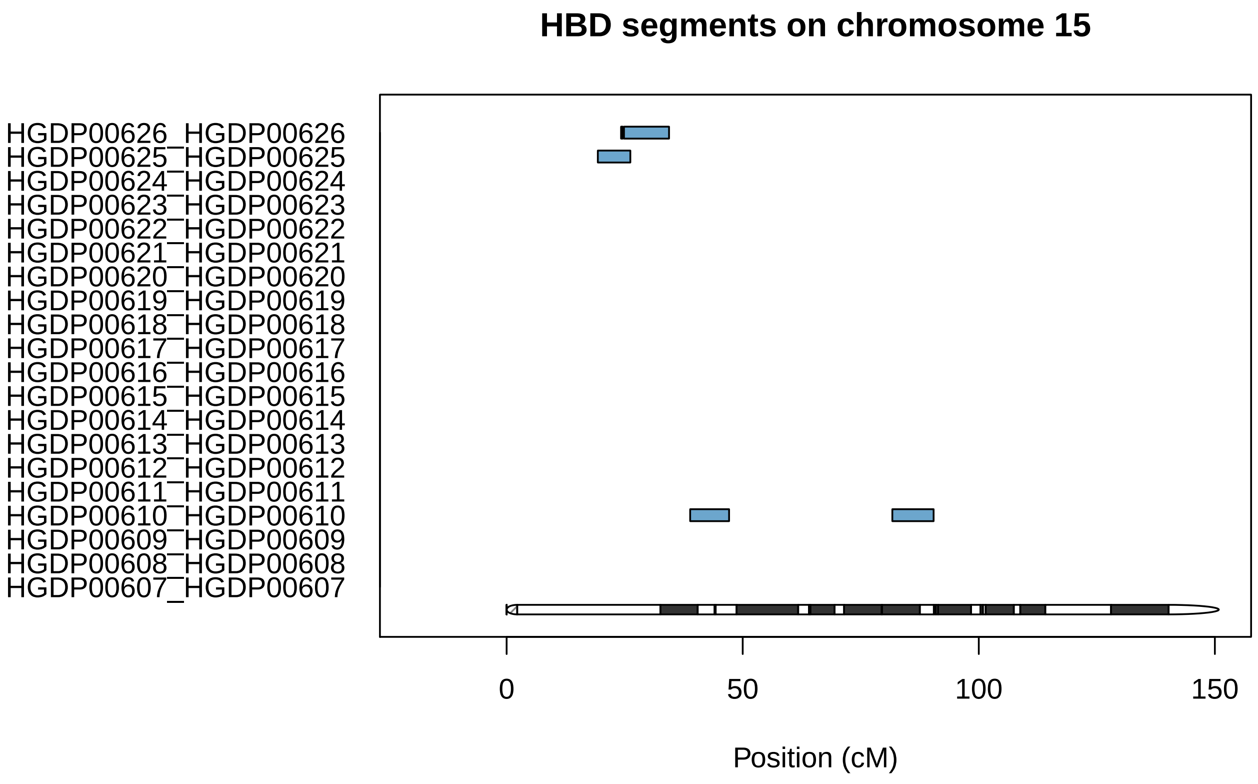


- As you can see you have a red line plotted, it is the moving average of the HFLOD, calculated on moving windows of 50 markers with the rollmean function of the R package zoo. This allows checking the consistency of HFLOD calculations (i.e. checking the fact that a high HFLOD score is not due to one submap only). A moving average is computed to remove the impact of a submap with a false positive signal.

- For method "Hotspots by segments", the use of the moving average does not make much sense as results are already average over the snps of a segment between hospots regions. We recommend using MA=FALSE.

5.3 HBD plot for a chromosome

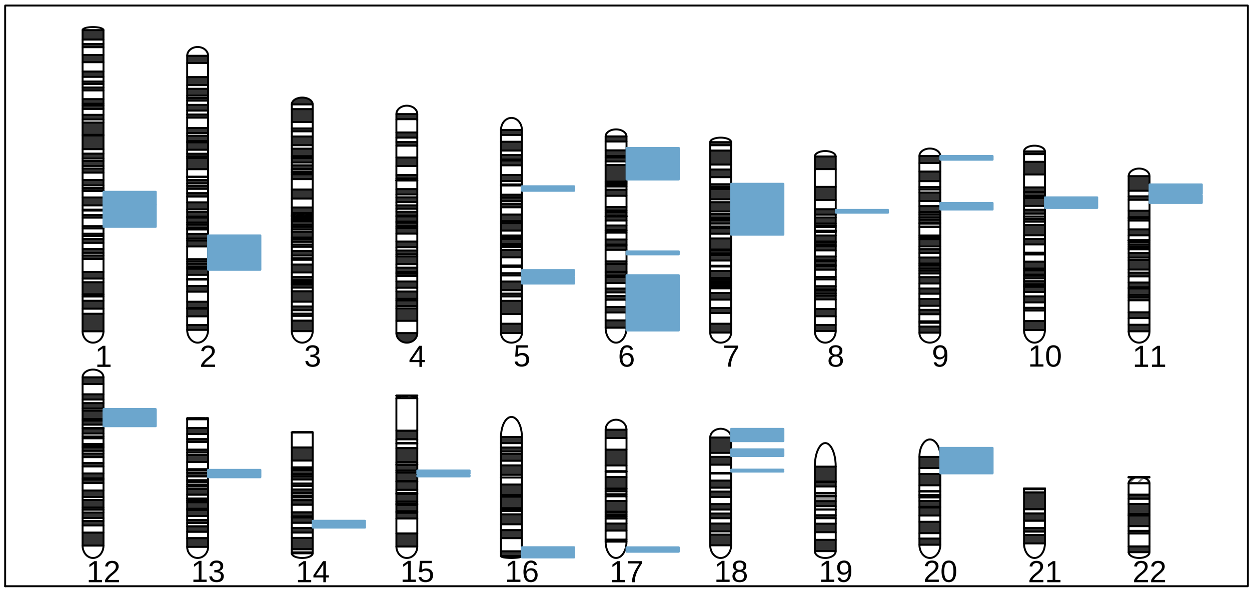
```
HBD.plot.chr(F1, chr=15)
```



5.4 HBD plot for an individual

```
HBD.plot.id(F1, individual.id = "HGDP00649", family.id = "HGDP00649")
```

HBD segments of HGDP00649_HGDP00649


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22