A Tutorial for HCMMCNVs

1. Introduction

HCMMCNVs is a browser-based software for detecting copy number variants (CNVs) using whole exome sequencing (WES) technology developed by R Shiny graphical user interface (GUI). In this tutorial, we will go through the installation and usage of each module step by step using 10 mini processed bam files from Cancer Cell Line Encyclopedia (CCLE). The HCMMCNVs software is publicly available at https://github.com/lunching/HCMM_CNVs. This tutorial can be found at ??? (TBA). Each module of HCMMCNVs will be introduced in later sections.

2. How to start

This is instruction of how to install and start HCMMCNVs shiny software (also available at https://github.com/lunching/HCMM_CNVs).

Requirement:

- R (>= 3.4.1)
- Shiny (>=1.2.0)

How to install shiny package:

1. Open R.
2. User can install the shiny package by the following command in R:
   install.package("shiny")

How to install and run HCMMCNVs

1. Open R.
2. Run HCMMCNVs by the following commands in R:
   library(shiny)
   shiny::runGitHub("HCMM_CNVs","lunching")
   (The first tab of HCMMCNVs setting page will pop-up, see Figure 1)

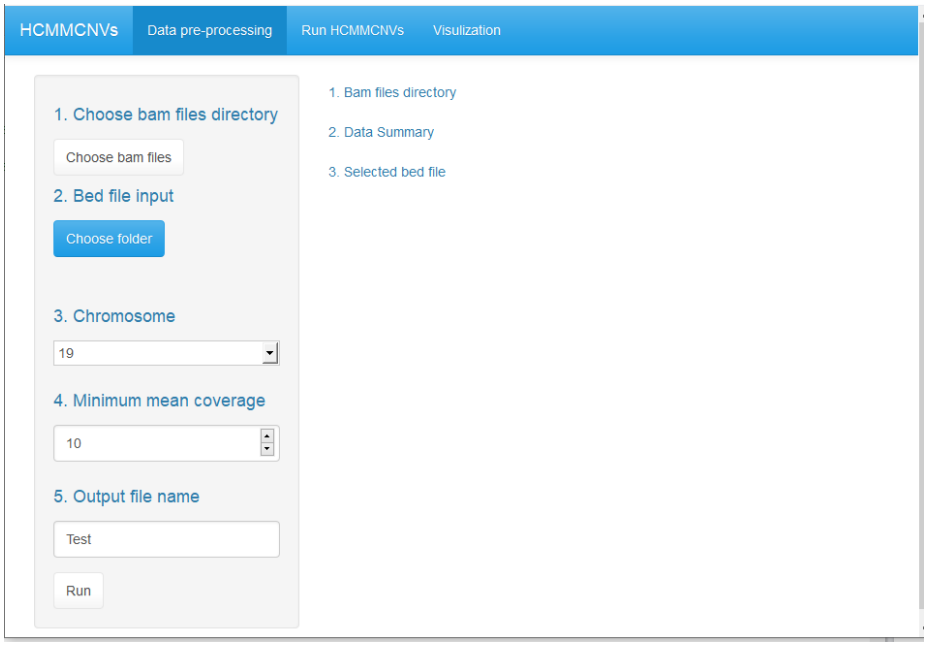


Figure 1: HCMMCNVs shiny software GUI setting page

3. HCMMCNVs setting page

After starting HCMMCNVs, there are three tabs: (1) Data pre-processing; (2) Run HCMMCNVs and (3) Visualization on the top of this page (see Figure 1). For question and bug report, please contact Lun-Ching Chang (changl@fau.edu) or leave your comment at github page (https://github.com/lunching/HCMM_CNVs).

4. Prepare data

In this section, we will introduce how to prepare input data sets (bam files) and functions in each tab by providing toy mini bam files (can be downloaded at github page: https://github.com/lunching/HCMM_CNVs/tree/master/Toy%20example).

4.1. Input data sets for data pre-processing

- **Processed bam files**

  A BAM file (.bam) is the binary version of a SAM file. A SAM file (.sam) is a tab-delimited text file that contains sequence alignment data. These formats are described on the SAM Tools web site: http://samtools.github.io/hts-specs/. (see more details in https://software.broadinstitute.org/software/igv/BAM).

  HCMMCNVs will automatically detect all bam files and bai files (indexed bam file). For generating "bai" format, use the following command under linux:

  samtools index input.bam

- **Bed files**

  A bed file (.bed) is Browser Extensible Data which provides a flexible way to define the data lines that are displayed in an annotation track. The first three required BED fields with required column names are:

  1. **Chr**: the name of the chromosome (e.g. chr6, chr19, etc.).
  2. **Start**: the start position of the feature in the chromosome.
  3. **End**: the ending position of the feature in the chromosome.

  User can create additional columns in bed files for additional information with preferred user-specified column names. See Figure 2 for the example of the bed file "Demo.bed" (also available at github page: https://github.com/lunching/HCMM_CNVs/tree/master/Toy%20example)

4.2. Example of toy example data set in HCMMCNVs shiny software

10 mini bam files and indexed file (".bai") are provided in HCMMCNVs shiny software at github page: https://github.com/lunching/HCMM_CNVs/tree/master/Toy%20example. Whole exome sequencing (WES) of these 10 bam files are available at Cancer Cell Line Encyclopedia (CCLE): https://portals.broadinstitute.org/ccle.

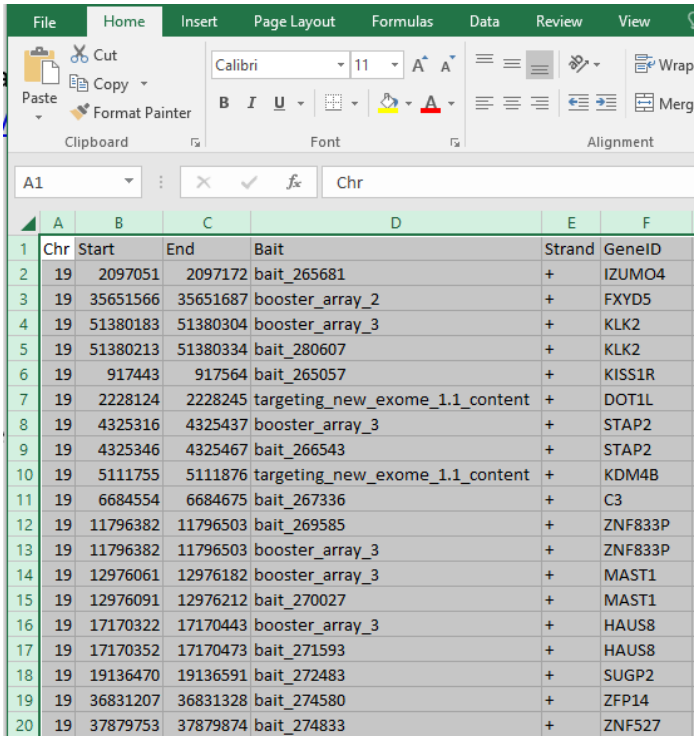
10 mini bam files are generated by 50 exons in chromosome 6 and 19 (see "Demo.bed" at github page under the directory: "**HCMM_CNVs/Toy example/**"). Please do not prepare your preferred "bed" file to work on these 10 mini bam files because these 10 mini bam only contain reads for the regions provided by "Demo.bed" at github page (https://github.com/lunching/HCMM_CNVs/blob/master/Toy%20example/Demo.bed).

The original sized bam files' name from CCLE can be found under the following table and can be downloaded using Genomic Sata Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/):

File names of toy example	File names of original bam file
CCLE_demo_S1.mini.bam	C835.HCC1143.2
CCLE_demo_S2.mini.bam	C835.HCC1954.2
CCLE_demo_S3.mini.bam	C835.K-562.3
CCLE_demo_S4.mini.bam	C836.22Rv1.2
CCLE_demo_S5.mini.bam	C836.ACC-MESO-1.2
CCLE_demo_S6.mini.bam	C836.ALL-SIL.1
CCLE_demo_S7.mini.bam	C836.AML-193.2
CCLE_demo_S8.mini.bam	C836.AMO-1.1
CCLE_demo_S9.mini.bam	C836.BDCM.2
CCLE_demo_S10.mini.bam	C836.BICR_16.1

If you prefer to use your own bam files, please make sure the name of the chromosome is “chrx” instead of “x”. If you have bam files only have “x”, you can use following commands in linux to change the chromosome notation:

*for file in *.bam; do filename=`echo $file | cut -d "." -f 1`; samtools view -H $file | sed -e 's/SN:\([0-9XY]\)/SN:chr\1/' -e 's/SN:MT/SN:chrM/' | samtools reheader - $file > ${filename}_chr.bam; done*



	A	B	C	D	E	F
1	Chr	Start	End	Bait	Strand	GeneID
2	19	2097051	2097172	bait_265681	+	IZUMO4
3	19	35651566	35651687	booster_array_2	+	FXYD5
4	19	51380183	51380304	booster_array_3	+	KLK2
5	19	51380213	51380334	bait_280607	+	KLK2
6	19	917443	917564	bait_265057	+	KISS1R
7	19	2228124	2228245	targeting_new_exome_1.1_content	+	DOT1L
8	19	4325316	4325437	booster_array_3	+	STAP2
9	19	4325346	4325467	bait_266543	+	STAP2
10	19	5111755	5111876	targeting_new_exome_1.1_content	+	KDM4B
11	19	6684554	6684675	bait_267336	+	C3
12	19	11796382	11796503	bait_269585	+	ZNF833P
13	19	11796382	11796503	booster_array_3	+	ZNF833P
14	19	12976061	12976182	booster_array_3	+	MAST1
15	19	12976091	12976212	bait_270027	+	MAST1
16	19	17170322	17170443	booster_array_3	+	HAUS8
17	19	17170352	17170473	bait_271593	+	HAUS8
18	19	19136470	19136591	bait_272483	+	SUGP2
19	19	36831207	36831328	bait_274580	+	ZFP14
20	19	37879753	37879874	bait_274833	+	ZNF527

Figure 2: An example of bed file

5. Process HCMMCNVs

In this section, we introduce how to upload the data sets into the HCMMCNVs shiny software using provided toy example

(https://github.com/lunching/HCMM_CNVs/tree/master/Toy%20example) step by step in each tab.

5.1. Data pre-processing

Step 1: Select the bam file folder

On the data pre-processing tab, click "Choose bam files" on left panel under 1. Choose bam file directory (Figure 3-I), then select the folder where you place your bam files (Figure 4-I), then click "Select" (Figure 4-II).

Step 2: Select the bed file

On the data pre-processing tab, click "Choose folder" on left panel under 2. Bed file input (Figure 3-II), then select the bed file, then click "open".

Step 3: Choose chromosome

On the data pre-processing tab, select the autosomal chromosome on the left panel under 3. Chromosome (Figure 3-III).

Step 4: Select threshold for minimum mean coverage (default is 10)

On the data pre-processing tab, enter the number of minimum mean coverage on left panel under 4. Minimum mean coverage (Figure 3-IV). The regions from provided bed file will be filtered if the mean coverage among all samples less than selected number (default is 10).

Step 5: Enter the file name for processed coverage matrix

On the data pre-processing tab, enter the file name of the output on the left panel under 5. Output file name (Figure 3-V). The output with "RData" format "Cov_matrix_[file name].RData" will be generate under the folder of selected bam file directory in step 1. The "RData" coverage matrix can be use as input under second tab: Run HCMMCNVs (Figure 5). Adjusted coverage matrix and input bed file will be saved in the "Cov_matrix_[file name].RData". User can use it for additional analysis.

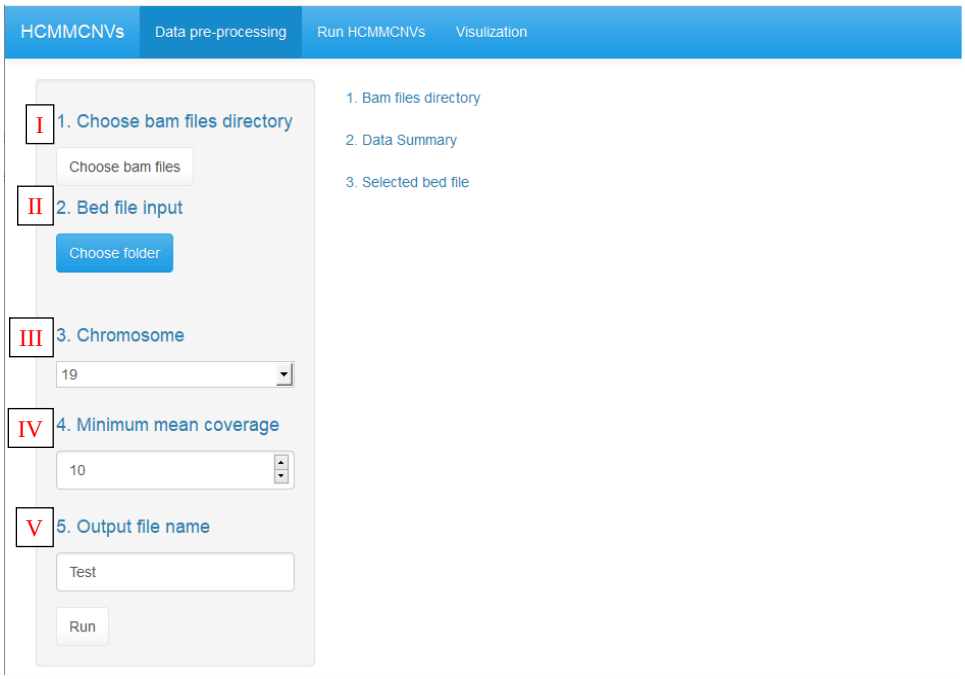


Figure 3: GUI for the data pre-processing tab

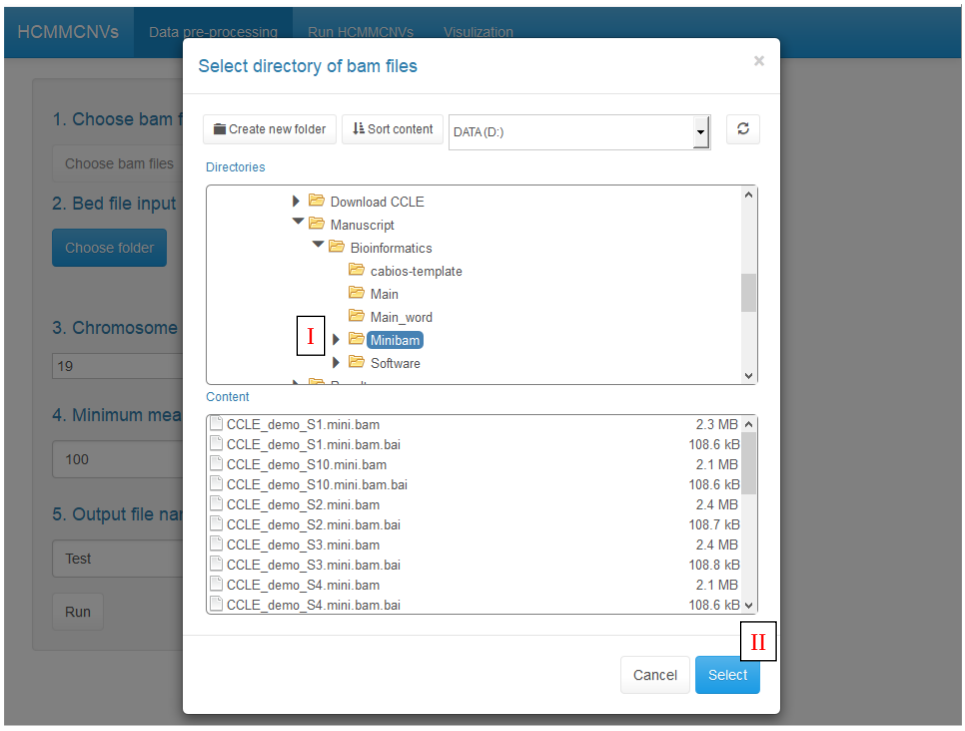


Figure 4: Uploading page for bam file directory

5.2. Run HCMMCNVs

Step 1: Select the coverage data

On the run HCMMCNVs tab, click "Choose Coverage RData" on the left panel under 1. Load the coverage RData (Figure 5-I) generated from first tab: data pre-processing (Figure 3).

Step 2: Choose number of clusters for hierarchical clustering method

On the run HCMMCNVs tab, enter the number of clusters (suggested number of clusters from 2 to 4, default is 3) for hierarchical clustering method on the left panel under 2. Hierarchical Clustering: number of clusters (Figure 5-II).

Step 3: Enter the file name for processed HCMMCMVs

On the run HCMMCNVs tab, enter the file name of the output on the left panel under 3. Output file name (Figure 5-III). The output with "RData" format "CBS_[file name].RData" will be generate under the directory of the HCMMCNVs shiny software. Segmentation means generated by circular binary segmentation (CBS) in R package and mixture model will be save in the "CBS_[file name].RData". The "RData" can be use as input under third tab: Visulization and user can use the "RData" for additional analysis.

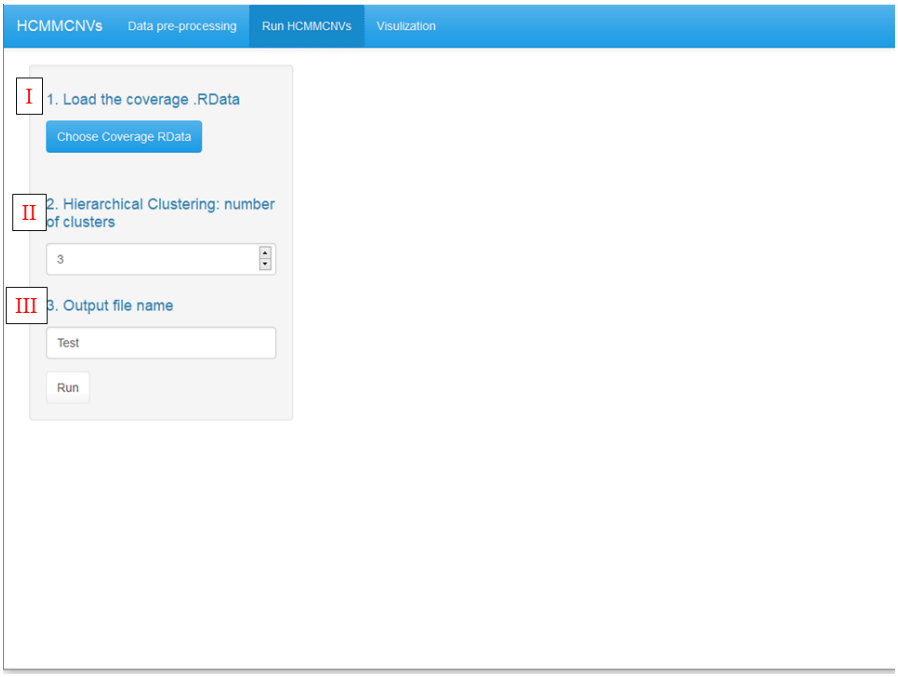


Figure 5: GUI for the HCMMCNVs tab

5.3. Visualization

Step 1: Select the output from circular binary segmentation

On the Visualization tab, click "Choose CBS results" on the left panel under 1. Load the CBS result (Figure 6-I) generated from second tab: run HCMMCNVs (Figure 5).

Step 2: Select sample to plot segmentation mean

On the Visualization tab, file names of uploaded bam files will appear under the check box (Figure 6-II) once user upload the CBS result generated from second tab.

Step 3: Plot

On the Visualization tab, click "Plot" on the left panel (Figure 6-III), the segmentation mean versus chromosome positions will appear on the right (Figure 6-IV). User can also save the figure using the download tab (Figure 6-V).

Figure 6: GUI for the Visualization tab