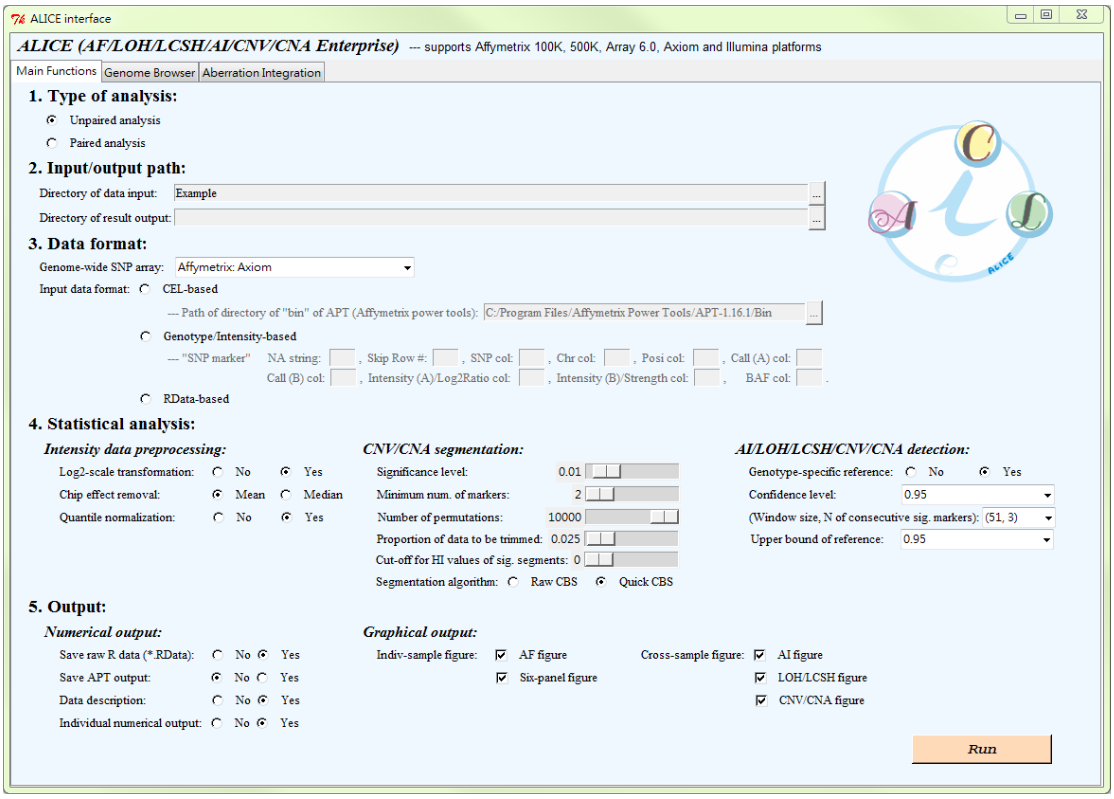
User Manual for ALICE software

1. Introduction of ALICE software

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) is a statistical tool providing a set of statistical methods developed for an integrative analysis of allele frequency (AF), allelic imbalance (AI), loss of heterozygosity (LOH), long contiguous stretches of homozygosity (LCSH), and copy number variation/alteration (CNV/CNA) based on SNP probe hybridization intensities and genotypes. It was programmed in R and R-GUI with a user-friendly interface. The software (32-bit and 64-bit), user manual, library files for running Affymetrix Power Tools (APT), annotation files, examples and population-specific database can be downloaded at the ALICE website (http://hcyang.stat.sinica.edu.tw/software/ALICE.html). In this user manual, we list the initiation process of ALICE in Section 2. ALICE provides three components - "Main Functions", "Genome Browser" and "Aberration Integration", each can be executed using the first, second and third tab page of ALICE. The corresponding interfaces are shown in Figure 1, Figure 2 and Figure 3 for the screen width of users' monitors is ≥ 800, and Figure 4, Figure 5 and Figure 6 for the screen width of users' monitors is < 800. The operation procedure and analytic functions of the interfaces of the components are described in Section 3. Moreover, ALICE supports three formats of input data, including CEL-based, Genotype/Intensity-based and RData-based. The common directory setting for all kinds of support input data format is given in Section 4.1, and the rest of directories required for each of the three input data formats are introduced in Section 4.2 to 4.4. The details of the formats and content of result output for the three components are explained in Section 5.1 to 5.3. Three example sets were prepared to illustrate the procedures of running ALICE to analyze different types of input data. For the input data of normal samples genotyped with Array 6.0 and Axiom arrays, the running procedures and parameter settings of ALICE interface for the unpaired analysis are shown in Section 6.1 and 6.2, respectively. Besides the analysis for normal samples, Section 6.3 describes the analysis for admixed samples of tumor cells and the corresponding normal cells of a cancer patient. Regarding the procedure of unpaired analysis for normal samples or cancer patients is the same; we only use Example 3-1 to show how to perform unpaired analysis on the admixed samples in a similar way as in Example 1-1. Besides unpaired analysis, ALICE also provides paired analysis. Example 3-2 demonstrates how to perform paired analysis of the admixed samples.

Figure 1. The first tab page "Main Functions" of ALICE when the screen width of users' monitors is ≥ 800 pixels.


ALICE interface
ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms
Main Functions
Genome Browser
Aberration Integration
Run

1. Type of analysis:

● Unpaired analysis

○ Paired analysis

2. Input/output path:

Directory of data input: Example

Directory of result output:

3. Data format:

Genome-wide SNP array: Affymetrix: Axiom

Input data format: ○ CEL-based

--- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin

○ Genotype/Intensity-based

--- "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) col:
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .

○ RData-based

4. Statistical analysis:

Intensity data preprocessing:

Log2-scale transformation: ○ No ● Yes

Chip effect removal: ● Mean ○ Median

Quantile normalization: ○ No ● Yes

CNV/CNA segmentation:

Significance level: 0.01

Minimum num. of markers: 2

Number of permutations: 10000

Proportion of data to be trimmed: 0.025

Cut-off for HI values of sig. segments: 0

Segmentation algorithm: ○ Raw CBS ● Quick CBS

AI/LOH/LCSH/CNV/CNA detection:

Genotype-specific reference: ○ No ● Yes

Confidence level: 0.95

(Window size, N of consecutive sig. markers): (51, 3)

Upper bound of reference: 0.95

5. Output:

Numerical output:

Save raw R data (*.RData): ○ No ● Yes

Save APT output: ● No ○ Yes

Data description: ○ No ● Yes

Individual numerical output: ○ No ● Yes

Graphical output:

Indiv-sample figure: ☑ AF figure

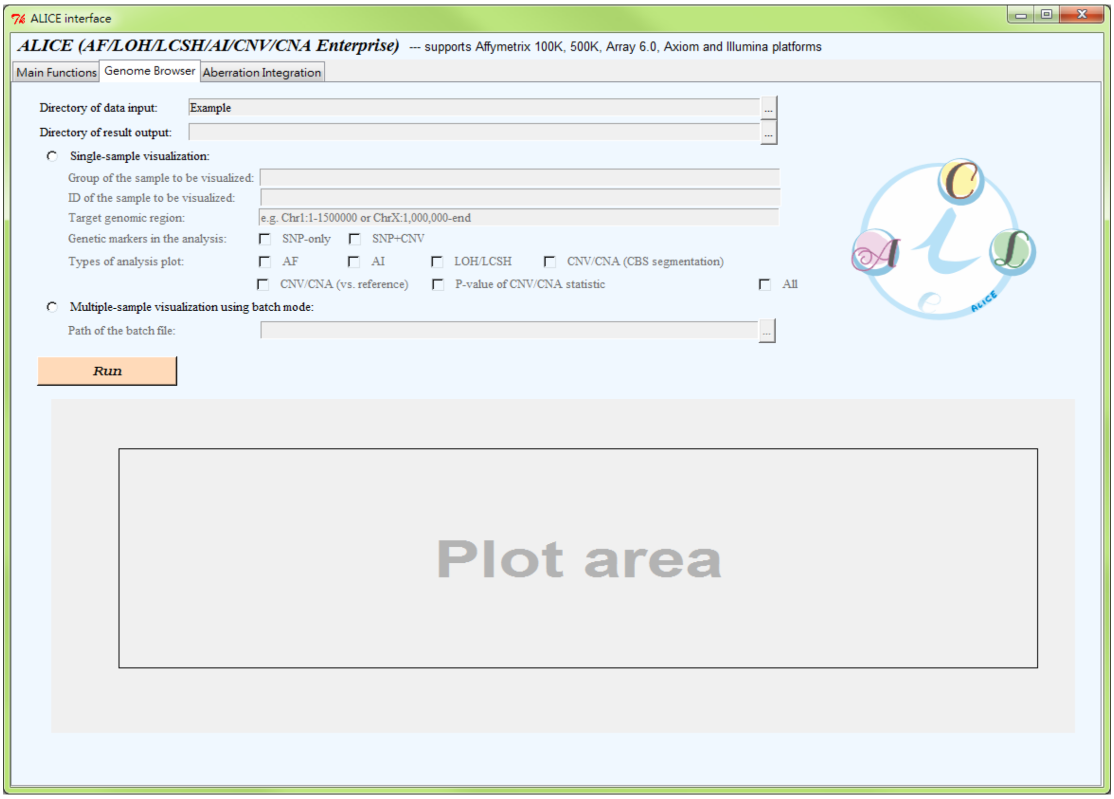
☑ Six-panel figure

Cross-sample figure: ☑ AI figure

☑ LOH/LCSH figure

☑ CNV/CNA figure

Figure 2. The second tab page "Genome Browser" of ALICE when the screen width of users' monitors is ≥ 800 pixels.


ALICE interface
ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms
Main Functions
Genome Browser
Aberration Integration
Run
Plot area

Directory of data input: Example

Directory of result output:

○ Single-sample visualization:

Group of the sample to be visualized:

ID of the sample to be visualized:

Target genomic region: e.g. Chr1:1-1500000 or ChrX:1,000,000-end

Genetic markers in the analysis: □ SNP-only □ SNP+CNV

Types of analysis plot: □ AF □ AI □ LOH/LCSH □ CNV/CNA (CBS segmentation)

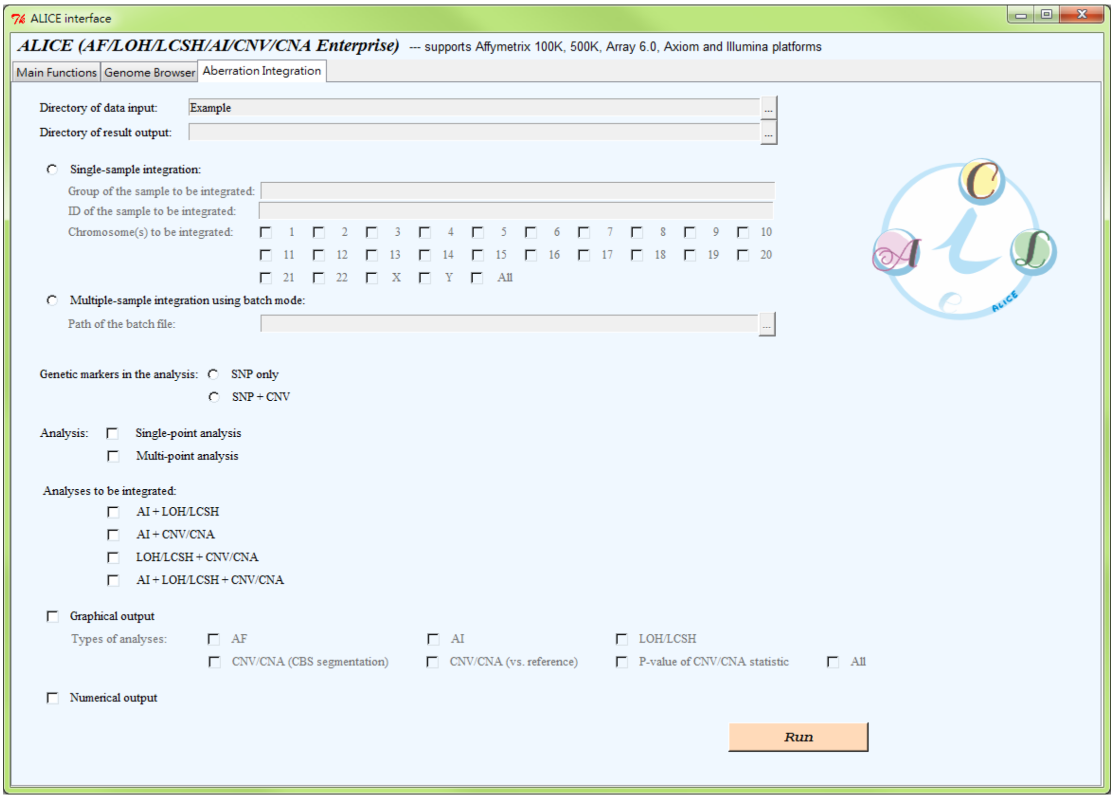
□ CNV/CNA (vs. reference) □ P-value of CNV/CNA statistic □ All

○ Multiple-sample visualization using batch mode:

Path of the batch file:

Figure 3. The third tab page "Aberration Integration" of ALICE when the screen width of users'

monitors is ≥ 800 pixels.



ALICE interface

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Main Functions | Genome Browser | Aberration Integration

Directory of data input: Example ...
Directory of result output: ...

○ Single-sample integration:
Group of the sample to be integrated:
ID of the sample to be integrated:
Chromosome(s) to be integrated: □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10
□ 11 □ 12 □ 13 □ 14 □ 15 □ 16 □ 17 □ 18 □ 19 □ 20
□ 21 □ 22 □ X □ Y □ All

○ Multiple-sample integration using batch mode:
Path of the batch file: ...

Genetic markers in the analysis: ○ SNP only
○ SNP + CNV

Analysis: □ Single-point analysis
□ Multi-point analysis

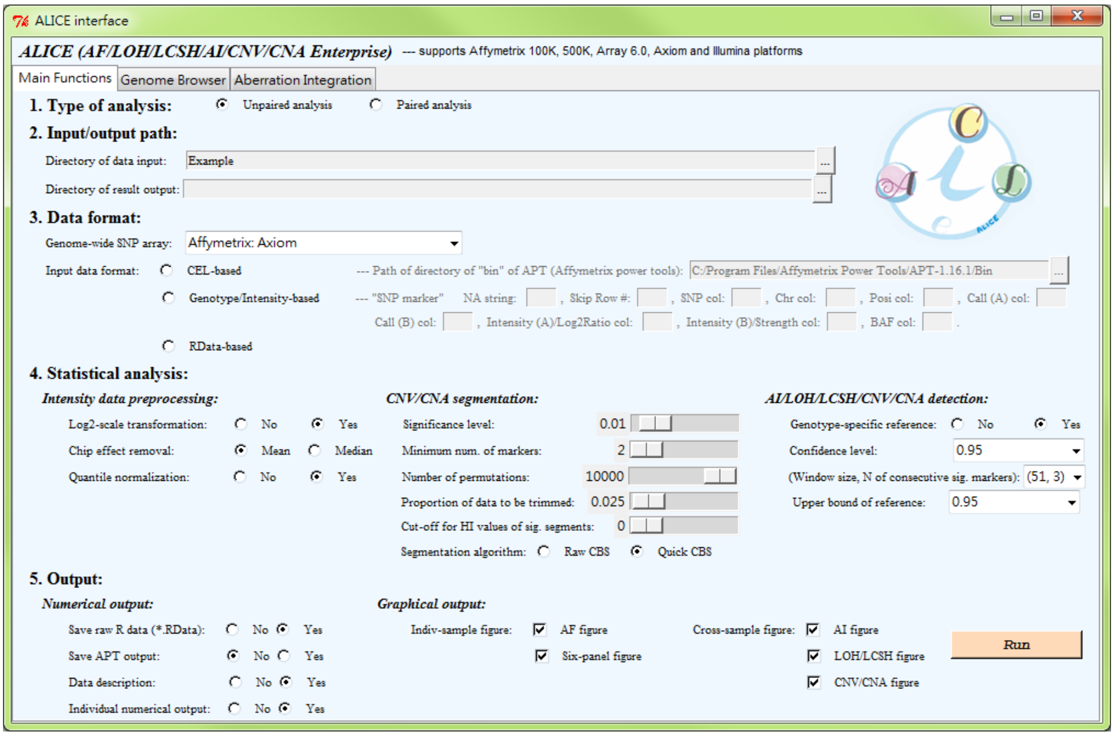
Analyses to be integrated:
□ AI + LOH/LCSH
□ AI + CNV/CNA
□ LOH/LCSH + CNV/CNA
□ AI + LOH/LCSH + CNV/CNA

□ Graphical output
Types of analyses: □ AF □ AI □ LOH/LCSH
□ CNV/CNA (CBS segmentation) □ CNV/CNA (vs. reference) □ P-value of CNV/CNA statistic □ All

□ Numerical output

Run

Figure 4. The first tab page "Main Functions" of ALICE when the screen width of users' monitors is smaller than 800 pixels.



ALICE interface

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Main Functions | Genome Browser | Aberration Integration

1. Type of analysis: ● Unpaired analysis ○ Paired analysis

2. Input/output path:
Directory of data input: Example ...
Directory of result output: ...

3. Data format:
Genome-wide SNP array: Affymetrix: Axiom
Input data format: ○ CEL-based --- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin ...
○ Genotype/Intensity-based --- "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) col:
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .
○ RData-based

4. Statistical analysis:

Intensity data preprocessing:
Log2-scale transformation: ○ No ● Yes
Chip effect removal: ● Mean ○ Median
Quantile normalization: ○ No ● Yes

CNV/CNA segmentation:
Significance level: 0.01
Minimum num. of markers: 2
Number of permutations: 10000
Proportion of data to be trimmed: 0.025
Cut-off for HI values of sig. segments: 0
Segmentation algorithm: ○ Raw CBS ● Quick CBS

AI/LOH/LCSH/CNV/CNA detection:
Genotype-specific reference: ○ No ● Yes
Confidence level: 0.95
(Window size, N of consecutive sig. markers): (51, 3)
Upper bound of reference: 0.95

5. Output:

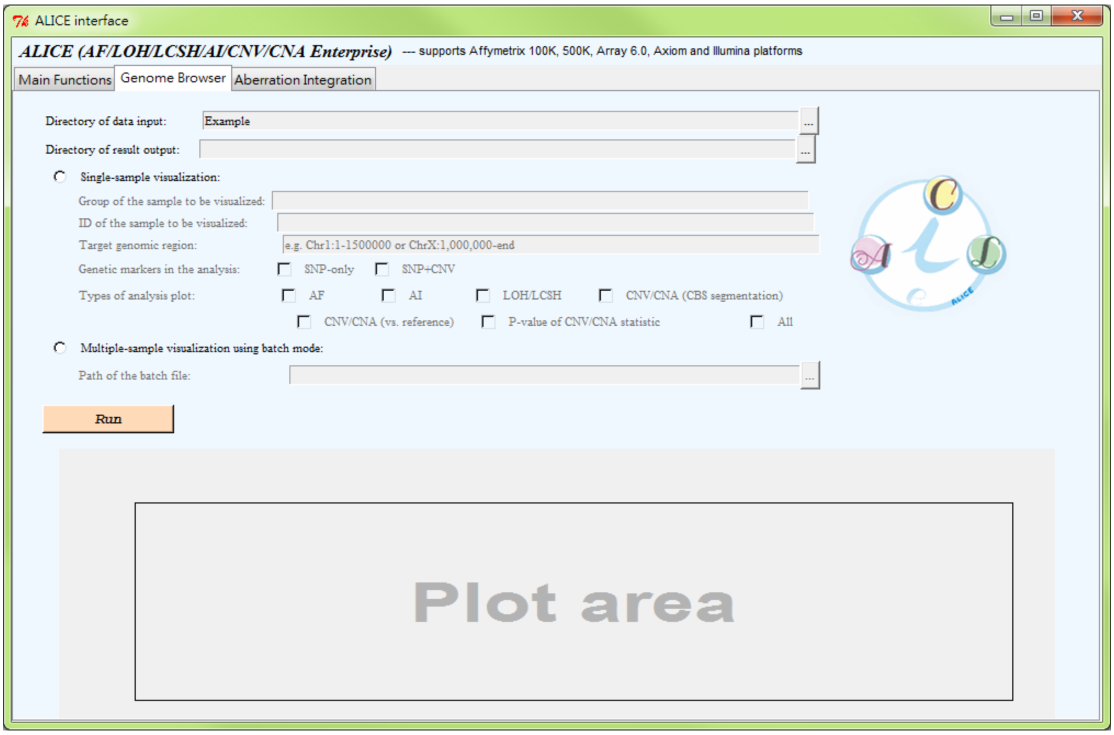
Numerical output:
Save raw R data (*.RData): ○ No ● Yes
Save APT output: ● No ○ Yes
Data description: ○ No ● Yes
Individual numerical output: ○ No ● Yes

Graphical output:
Indiv-sample figure: ☑ AF figure
☑ Six-panel figure
Cross-sample figure: ☑ AI figure
☑ LOH/LCSH figure
☑ CNV/CNA figure

Run

Figure 5. The second tab page "Genome Browser" of ALICE when the screen width of users' monitors

is smaller than 800 pixels.


ALICE interface

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Main Functions | Genome Browser | Aberration Integration

Directory of data input: Example ...

Directory of result output: ...

○ Single-sample visualization:

Group of the sample to be visualized:

ID of the sample to be visualized:

Target genomic region: e.g. Chr1:1-1500000 or ChrX:1,000,000-end

Genetic markers in the analysis: □ SNP-only □ SNP+CNV

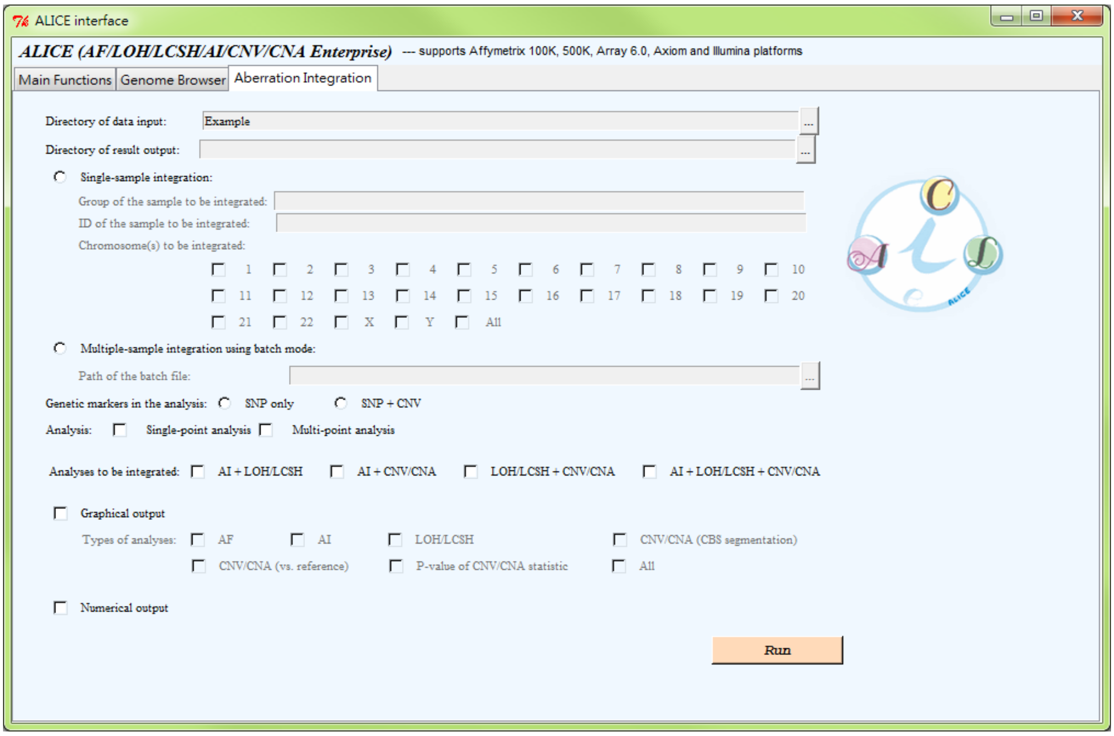
Types of analysis plot: □ AF □ AI □ LOH/LCSH □ CNV/CNA (CBS segmentation) □ CNV/CNA (vs. reference) □ P-value of CNV/CNA statistic □ All

○ Multiple-sample visualization using batch mode:

Path of the batch file: ...

Run

Figure 6. The third tab page “Aberration Integration” of ALICE when the screen width of users’ monitors is smaller than 800 pixels.


ALICE interface

ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Main Functions | Genome Browser | Aberration Integration

Directory of data input: Example ...

Directory of result output: ...

○ Single-sample integration:

Group of the sample to be integrated:

ID of the sample to be integrated:

Chromosome(s) to be integrated:

□ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10

□ 11 □ 12 □ 13 □ 14 □ 15 □ 16 □ 17 □ 18 □ 19 □ 20

□ 21 □ 22 □ X □ Y □ All

○ Multiple-sample integration using batch mode:

Path of the batch file: ...

Genetic markers in the analysis: ○ SNP only ○ SNP + CNV

Analysis: □ Single-point analysis □ Multi-point analysis

Analyses to be integrated: □ AI + LOH/LCSH □ AI + CNV/CNA □ LOH/LCSH + CNV/CNA □ AI + LOH/LCSH + CNV/CNA

□ Graphical output

Types of analyses: □ AF □ AI □ LOH/LCSH □ CNV/CNA (CBS segmentation) □ CNV/CNA (vs. reference) □ P-value of CNV/CNA statistic □ All

□ Numerical output

Run

2. ALICE Software initialization

ALICE software was developed using R software and wrapped as an executable file. The procedure of initiating ALICE is listed as follows.

- **Step 1**: Download and install R software from the R project website: http://www.r-project.org/.
- **Step 2**: Download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website: http://hcyang.stat.sinica.edu.tw/software/ALICE.html. Select 32-bit or 64-bit version according to your Windows installation.
- **Step 3**: After R software is installed, ALICE software can be executed by double-clicking the icon .

  Note: for the users using OS above MS Windows 7, please run ALICE.exe as administrator.

3. ALICE Interfaces, Functions and Operating Procedures

3.1 ALICE setting for the component "Main Functions"

This component provides the whole-genome AI, LOH/LCSH and CNV/CNA analysis of the data from Affymetrix and Illumina platforms.
There are five main items in this component (Figure 1) as follows.

- **Item 1 – Type of analysis**
  Users can check the icon "Unpaired analysis" and "Paired analysis" to specify the type of analysis.
- **Item 2 – Input/output path**
  Users specify the paths of the directories of data input files and result output files. ALICE will import data from the specified path of "Directories of data input" and save analysis results in the specified path of "Directory of result output". For different types of data input files, users need to construct different directories under the directory "Directories of data input". The details will be introduced in Section 4.
- **Item 3 – Data format**
  (1) **Genome-wide SNP array**: Currently, ALICE supports the analysis of Affymetrix 100K, 500K, Array 6.0 and Axiom as well as Illumina platforms. Users select which type of genome-wide SNP array to analyze from the pull-down menu. For CEL files genotyping using Affymetrix Array 6.0 platform, users select "Affymetrix: Array 6.0 (with CN data)"; otherwise, users select "Affymetrix: Array 6.0 (without CN data)" for Array 6.0 input data.
  (2) **Input data format**: Users select icon to specify the input data format.
    - **CEL-based**: For input data in Affymetrix CEL format, users select icon "CEL-based" and provide the path of directory of "bin" of APT (Affymetrix power tools).
      Note: Prior to run ALICE for CEL format, users should install APT from Affymetrix website:
      http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx.
    - **Genotype/Intensity-based**: For input data in the text file format, which were processed and generated by using Affymetrix or Illumina software (e.g. Genotyping Console, Affymetrix Power Tools or GenomeStudio), users select the icon "Genotype/Intensity-based" and specify the column indices of the required variables. ALICE will

automatically disable the icons of unrequired variables according to the specified *Genome-wide SNP array* platform.

  - **RData-based**: For users who intended to run ALICE again with different parameter settings based on the same input data files that used in the previous ALICE analysis, users select the icon "RData-based". The instruction of how to use the previous-time processed individual RData files as the current-time input data is described in Section 4.4.

- **Item 4 – Statistical analysis**

  (1) **Intensity data processing**: Users select the icons "Yes" or "No" to specify whether the data processing ("Log2-scale transformation", "Chip effect removal" and "Quantile normalization") to apply onto intensity data.

  (2) **CNV/CNA segmentation**: Users specify values for the six parameters used for CBS algorithm.

  - **Significance level**: The range of the parameter used in the test to accept change-points is from 0 to 0.1, with an increment of 0.01.
  - **Minimum num. of markers**: The range of the parameter used in detecting a changed segment is from 2 to 5, with an increment of 1.
  - **Number of permutations**: The range of the parameter used in p-value computation is from 10^3 to 10^5, with an increment of 100.
  - **Proportion of data to be trimmed**: The range of the parameter used in variance calculation for smoothing outliers is from 0 to 1, with an increment of 0.01.
  - **Cut-off for HI values of sig. segments**: The range of the parameter used as the cut-off threshold of HI values of significant segments is from 0 to 3, with an increment of 0.01. The absolute average HI value of a segment lower than the cut-off threshold will be drew in grey color in the 4th panel.
  - **Segmentation algorithm**: Users select the icon "Raw CBS" or "Quick CBS" to specify whether the raw CBS algorithm or our proposed quick CBS algorithm to use.

  (3) **AI/LOH/LCSH/CNV/CNA detection**: Users specify values for the four parameters used in the AI/LOH/LCSH/CNV/CNA calculation.

  - **Genotype-specific reference**: Users select the icon "Yes" or "No" to specify which genotype-specific or non-genotype-specific reference database to use.
  - **Confidence level**: Users specify the value of confidence level to be 0.95, 0.975, or 0.99 using the pull-down menu.

    - **(Window size, N of consecutive significant markers)**: Users select the values of the parameters out of the recommended settings using the pull-down menu. The parameter “Window size” is the number of markers in a window, the parameter “N of consecutive significant markers” is the threshold number of consecutive significant markers.
    - **Upper bound of reference**: Users specify the value of upper bound of the quantile of reference database to be 0.95, 0.975, or 0.99 using the pull-down menu.

- **Item 5 – Output**

  (1) **Numerical output**: Users select the icons “Yes” or “No” to specify whether to save the output files.

    - **Save raw R data (*.RData)**: ALICE processes input data of *CEL-based* and *Genotype/Intensity-based* format and saves the extracted raw data of each individual as an .RData file in a temporary folder. Users select the icon “Save raw R data (*.RData)” to retain the directory; otherwise, ALICE will delete it once the calculation is finished to free up disk space. The instruction of how to use the processed individual RData files as input data is described in Section 4.4.
    - **Save APT output**: This refers to the output files generated during Affymetrix Power Tools calculation. If users select “No”, then the files will be deleted to free up disk space.
    - **Data description**: If users select “Yes”, ALICE will export a text file named “Description.txt” which records the parameter setting that users selected for ALICE analysis.
    - **Individual numerical output**: If users select “Yes”, ALICE will export numerical output during ALICE analysis.

  (2) **Graphical output**: Users check the boxes to specify whether to save *Individual-sample figure*: AF and Six-panel figures, or *Cross-sample figure*: AI, LOH/LCSH and CNV/CNA figures.

After paths, options and parameter values were assigned, users click “Run” icon to submit the job.

Note: ALICE analysis will be terminated if the graphical interface was closed.

3.2 ALICE setting for the component “Genome Browser”

This component provides the genome browser function for a further zoom-in investigation of the results that already being analyzed by ALICE.

There are two main parts in this component (Figure 2) as follows.

- **1. Directory of data input/Directory of result output**

  Users provide the paths of the directories of data input files and result output files used in the ALICE analysis that intended to investigate.

- **2. Single-sample visualization or Multiple-sample visualization using batch mode**

  Users select which type of visualization for the zoom-in investigation, either by “Single-sample visualization” or by “Multiple-sample visualization using batch mode”.

  - **Single-sample visualization**

    Users fill in the following information to specify the region of interest to investigate.

    - **Group of the sample to be visualized**: The filled-in string should be exactly the same as the directory name that ALICE constructed under the *Directory of result output* for the group of interest.
    - **ID of the sample to be visualized**: The filled-in string should be exactly the same as the directory name that ALICE constructed under the *Directory of result output* for the sample of interest.
    - **Target genomic region**: Users should specify the genomic regions of interest in the format “Chr$R$:P_s-P_e”, where R**,** P_s and P_e are the chromosome, starting and ending positions of the genomic region. Users can input “1” and “end” for P_s and P_e to specify the starting and ending position of the region of interest is the beginning and end position of the chromosome R. It is allowed to use comma (“,”) to separate sequences of digits, e.g. 32,000,000. Users also can input “KB” and “MB” for the unit of position, i.e. 32KB = 32,000 bp and 32.5MB = 32500000 bp.
    - **Genomic markers in the analysis**: Users can select the boxes “SNP-only” or “SNP+CN” to specify the marker types of interest. If the input data are not of the format “Affymetrix: Array 6.0 (with CN data)”, the icon “SNP+CN” will not function.
    - **Type of analyses**: Users can check the boxes of the analyses of interest to select which panels to draw in a six-panel figure, or check the box “All” to select all six panels. The boxes “AF”, “AI”, “LOH/LCSH”, “CNV/CNA (CBS segmentation)”, “CNV/CNA (vs. reference)” and “P-value of CNV/CNA statistic” corresponds to the order of the panels from top to bottom in the six-panel figure.

  - **Multiple-sample visualization using batch mode**

    Users provide a batch file of the required information and specify its path through the choose-directory dialog box. The batch file should contain ten tab-delimited columns with specific column names, which are introduced in the following. If a string is enclosed between two double quotation marks (“), it means the name of a column should be specified as the given string; otherwise, the column couldn’t be recognized by ALICE. Each row represents the information for a single sample to be visualized. Table 1 shows an example for a batch file used for the 2nd component. The detail of each column is listed as follows.

- **Column 1 - “Group”**: Each element in this column is the group name of the sample to be visualized.
- **Column 2 - “Sample_ID”**: Each element in this column is the ID of the sample to be visualized.
- **Column 3 - “Posi”**: Each element in this column is the target genomic region to be visualized, which should be in the format of “Chr$R$:P_s-P_e”, where R, P_s and P_e are defined as the same in the aforementioned section about the entry of “Target genomic region”.
- **Column 4 - “Marker_Type”**: Each element in this column is the genomic marker of interest, which should be “SNP” or “SNP+CN” for SNP-only and SNP+CN marker to analyze, respectively.
- **Column 5-10 - “Panel_AF”, “Panel_AI”, “Panel_LOH”, “Panel_CNV_CBS”, “Panel_CNV_ref”, “Panel_CN_pv”**: Each element in the 5th to 10th column is the indicator representing whether the “AF”, “AI”, “LOH”, “CNV (CBS segmentation)”, “CNV (vs. reference)” and “P-value of CN statistics” analysis to be investigated, respectively. An indicator should be “1” or “0” to indicate “Yes” or “No”.

After paths, options and parameter values were assigned, users click “Run” icon to submit the job. ALICE will draw zoom-in multi-panel figure(s), which constituted of the panels of the selected types of analyses, of the target genomic region(s) for the selected sample(s) in the selected group(s).

Note: ALICE analysis will be terminated if the graphical interface was closed.

Table 1. An example batch file for the 2nd component “Genome Browser”.

Group	Sample_ID	Posi	Marker_Type	Panel_AF	Panel_AI	Panel_LOH	Panel_CNV_CBS	Panel_CNV_ref	Panel_CN_pv
Test	Sample_1	Chr1:1-5MB	SNP	1	1	1	1	1	1
Test	Sample_2	Chr20:1KB-end	SNP	1	1	1	1	1	1

3.3 ALICE setting for the component “Aberration Integration”

This component provides the aberration integration function to integrate multiple analyses of results that already analyzed by ALICE, and export graphical and numerical output based on integrated results.

There are seven main parts in this component (Figure 3) as follows.

- **Directory of data input/Directory of result output**

  Users provide the paths of the directories of data input files and result output files that used for the ALICE analysis that intended to investigate.

- **Single-sample integration or Multiple-sample integration using batch mode**

  Users select which type of aberration integration to perform, either by “Single-sample integration” or by “Multiple-sample integration using batch mode”.

  - **Single-sample visualization**

    Users fill in the following information to specify the region of interest to investigate.

    - **Group of the sample to be integrated**: The filled-in string should be exactly the same as the directory name that ALICE constructed under the *Directory of result output* for the group of interest.
    - **ID of the sample to be integrated**: The filled-in string should be exactly the same as the directory name that ALICE constructed under the *Directory of result output* for the sample of interest.
    - **Chromosome(s) to be integrated**: Users select the boxes to specify which chromosome(s) to be integrated, or check the box “All” to select all.

  - **Multiple-sample integration using batch mode**

    Users provide a batch file of the required information and specify its path through the choose-directory dialog box. The batch file should contain three tab-delimited columns with specific column names, which are introduced in the following. If a string is enclosed between two double quotation marks (“), it means the name of a column should be specified as the given string; otherwise, the column couldn’t be recognized by ALICE. Each row represents the information for a single sample to be integrated. Table 2 shows an example for a batch file used for 3rd component. The detail of each column is listed as follows.

    - **Column 1 - “Group”**: Each element in this column is the group name of the sample to be integrated.
    - **Column 2 - “Sample_ID”**: Each element in this column is the ID of the sample to be integrated.
    - **Column 3 - “Chr”**: Each element in this column is the chromosome(s) of the sample to be integrated. Multiple chromosomes should be

separated using ",". For example, the element "1,X,Y" represents chromosomes 1, X and Y to be integrated. Users can use the number 23 and 24 to represent chromosome X and Y, respectively. Users also can use the dash mark "-" to produce a vector slice between two chromosome indices. For example, the element "1-24" represents chromosomes 1, 2, …, 24 to be investigated.

- **Genetic markers in the analysis**

  Users check the icon "SNP only" or "SNP+CN" to select which type of marker to be integrated.
- **Analysis**

  Users check the icon "Single-point result" or "Multi-point result" to select which type of analysis to be integrated.
- **Analyses to be integrated**

  Users check the icons "AI + LOH/LCSH", "AI + CNV/CNA", "LOH/LCSH + CNV/CNA" and/or "AI + LOH/LCSH + CNV/CNA" to select which type of analyses to be integrated.
- **Graphical output**

  Users check the boxes of the type of analyses "AF", "AI", "LOH/LCSH", "CNV/CNA (CBS segmentation)", "CNV/CNA (vs. reference)" and "P-value of CN statistics" to be integrated, , or check the box "All" to select all types of analyses.
- **Numerical output**

  Users check the box to select whether to generate the numerical output of integrated results.

After paths, options and parameter values were assigned, users click "Run" icon to submit the job. ALICE will generate numerical and graphical output file(s) of the selected integrated analyses for the selected chromosome(s) of the selected sample(s) in the selected group(s). For a selected integrated analyses type, a genetic marker is significant if all the results of the selected analyses to be integrated of the marker were significant.

Note: ALICE analysis will be terminated if the graphical interface was closed.

Table 2. An example batch file for the 3rd component "Aberration Integration".

Group	Sample_ID	Chr
Test	Sample_1	1,X,Y
Test	Sample_2	1,3,23-24
Test	Sample_3	1-24

4. Data Input Format for the component “Main Functions”

This section describes the data structure that users should prepare for the first component “Main Functions”. Section 4.1 described the structure of common directories required for all kinds of supported input data format. Section 4.2-4.4 described the directory structure required for the input data formats: “CEL-based”, “Genotype/Intensity-based” and “RData-based”, respectively.

Note: A string with double quotes (“) represents the name of a directory, a file, or a column must be given exactly the same as the string.

4.1 Common directories required for all kinds of input data format

There are at least three directories that users should construct under the *Directory of data input* prior to run ALICE analysis.

1. **Directory “Annotation”**

   This directory should contain four files for the *Genome-wide SNP array* is of “Affymetrix: Array 6.0 (with CN data)” and the *Input data format* is of “CEL-based” and three files for the rest:

   (1) **Annotation for SNP markers**: The file should contain the string “SNPanno” in the filename and thee columns with the specific column names as listed in the follows. Each row is the information of a SNP marker that featured on the SNP array.

   - **Column “SNP_ID”**: Each element in this column is the ID of a SNP marker. It must be character.
   - **Column “Chr”**: Each element in this column is the chromosome of a SNP marker located on. It must be integer between 1 and 24. The number 23 and 24 represents chromosome X and Y, respectively.
   - **Column “PhyPosi”**: Each element in this column is the physical position (bp) of a SNP marker. It must be integer.

   (2) **Annotation for CN markers**: The file should contain the string “CNanno” in the filename and four columns. Each row is the information of a CN marker that featured on the SNP array.

   Note: Users only need to provide this file when the Genome-wide SNP array is of “Affymetrix: Array 6.0 (with CN data)” and the Input data format is of “CEL-based”.

   - **Column “CN_ID”**: Each element in this column is the ID of a CN markers. It must be character.
   - **Column “Chr”**: Each element in this column is the chromosome of

a CN marker located on. It must be integer between 1 and 24. The number 23 and 24 represents chromosome X and Y, respectively.

- **Column "PhyPosi_start"**: Each element in this column is the starting physical position (bp) of a CN marker. It must be integer.
- **Column "PhyPosi_end"**: Each element in this column is the ending physical position (bp) of a CN marker. It must be integer.

(3) **Genomic positions for centromeres**: The file should contain the string "Centromere" in the filename and thee columns. Each row is the information of a centromere.

- **Column "Chr"**: Each element in this column is the chromosome of a centromere located on. It must be integer between 1 and 24. The number 23 and 24 represents chromosome X and Y, respectively.
- **Column "PhyPosi_start"**: Each element in this column is the starting physical position (bp) of a centromere. It must be integer.
- **Column "PhyPosi_end"**: Each element in this column is the ending physical position (bp) of a centromere. It must be integer.

(4) **Gene list**: The file should contain the string "Genelist" in the filename and four columns. Each row is the information of a gene.

- **Column "Chr"**: Each element in this column is the chromosome of a gene located on. It must be integer between 1 and 24. The number 23 and 24 represents chromosome X and Y, respectively.
- **Column "PhyPosi_start"**: Each element in this column is the starting physical position (bp) of a gene. It must be integer.
- **Column "PhyPosi_end"**: Each element in this column is the ending physical position (bp) of a gene. It must be integer.
- **Column "Gene_symbol"**: This column lists the gene symbol of a gene. It must be character.

2. Directory "Ref"

This directory should contain 5 RData files:

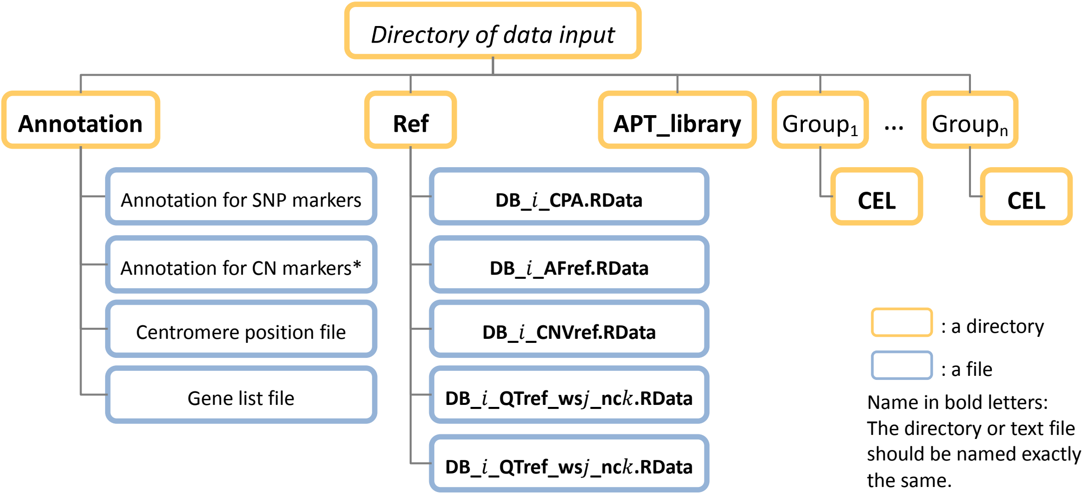
(1) DB_i_CPA.RData,
(2) DB_i_AFref.RData,
(3) DB_i_CNVref.RData,
(4) DB_i_WAPref_ws$j$_nc$k$.RData and
(5) DB_i_QTref_ws$j$_nc$k$.RData.

Among which, i equals 1 to 8 representing 8 kinds of parameter combination, j and k equal the numbers of window size and number of consecutive significant markers. After users click "Run" icon to submit the job, a message window will pop up to remind users which ALICE database to place in this directory. The ALICE database for all supported platforms and input data types are provided on the ALICE website (http://hcyang.stat.sinica.edu.tw/software/ALICE.html) for users to download.

3. **Group directory(ies)**

Users can group the input files by directory, called a group directory, such that ALICE would apply quantile normalization over the sample files in a group directory. It's allowed to construct multiple group directories with different directory name other than "Annotation" and "Ref". ALICE will run analysis of the groups in batch mode by group name in alphabetical order. A group directory that constructed under the *Directory of result output* for saving the analysis results will be named after the group directory name of the source input data.

Figure 7. The structure of *Directory of data input* for *CEL-based* input data format.



*: only required when the *Genome-wide SNP array* is "Affymetrix: Array 6.0 (with CN data)"

4.2 Directories required for CEL-based input data format

ALICE supports the analysis of Affymetrix CEL files. The structure of *Directory of data input* is given in Figure 7. Besides the three common directories mentioned in Section 4.1, users should construct two directories prior to run ALICE analysis.

- **Directory "APT_library"**

  This directory should contain the library files required for running APT (Affymetrix Power Tools). The library files can be downloaded from Affymetrix website or ALICE website. For different platform, the required library files are slightly different. Please refer to the user manual of APT for more details about library files. (http://media.affymetrix.com/support/developer/powertools/changelog/index.html).

- **Directory “CEL”**

  Under a group directory, users construct a directory named “CEL” and save the input data of *.CEL file format in it.

4.3 Directories required for Genotype/Intensity-based input data format

The structure of *Directory of data input* for *Genotype/Intensity-based* input data format is given in Figure 8. Besides the three common directories mentioned in Section 4.1, the directory(ies) that users should construct prior to run ALICE analysis and the content of the input data files are different to different SNP array platform.

1. **For Affymetrix 100K/500K platform:**

   For these two platforms, users construct two directories “IndGeno” and “IndPI” to store the genotype and intensity data separately. For a sample genotyped using 100K or 500K platform, there should be two arrays that comprising an array set (for 100K: Hind and Xba, for 500K: Nsp and Sty) for the sample.

   Note: The text files of two arrays of a sample should be adjacent to each other in the list.

   - **Directory “IndGeno”**: The directory should contain text files (*.txt) of genotype and genomic information of two arrays of the samples. In each text file, there are at least 4 columns should be contained and each row is the information of a SNP that featured on the SNP array. Users specify the string representing NA (i.e. not available) and the number of skip rows of headers in the entry “NA string” and “Skip Row #”, respectively.

     i. **A column for SNP IDs**: Each element in this column is the ID of a SNP marker. It must be character. Users specify the index of the column in the entry “SNP col”.

     ii. **A column for chromosomes**: Each element in this column is the chromosome of a SNP marker. Users specify the index of the column in the entry “Chr col”.

     iii. **A column for physical positions**: Each element in this column is the physical position of a SNP marker. It must be integer. Users specify the index of the column in the entry “Posi col”.

     iv. **A column for genotype calls**: Each element in this column is the genotype call of a SNP marker. It must be character. Users specify the index of the column in the entry “Call (A) col”.

- **Directory “IndPI”**: The directory should contain text files (*.txt) of probe intensities of two arrays of the sample. In each text file, there are at least 57 columns should be contained.
  - i. **The 2nd column**: Each element in this column is the ID of a SNP marker. It must be character.
  - ii. **The 3rd – 58th columns**: The elements in the 56 column of a row are the intensities of 56 probes of a SNP marker. It must be numeric.

2. **Affymetrix Array 6.0 platform:**

   For this platform, users construct a directory “IndGeno” to store the text files (*.txt) of the samples.

- **Directory “IndGeno”**: The directory should contain text files (*.txt) of genotypes, genomic information and intensities of the samples. In each text file, there are at least 6 columns should be contained. Users specify the string representing NA (i.e. not available) and the number of skip rows of headers in the entry “NA string” and “Skip Row #”, respectively.
  - i. **A column for SNP IDs**: Each element in this column is the ID of a SNP marker. It must be character. Users specify the index of the column in the entry “SNP col”.
  - ii. **A column for chromosomes**: Each element in this column is the chromosome of a SNP marker. Users specify the index of the column in the entry “Chr col”.
  - iii. **A column for physical positions**: Each element in this column is the physical position of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Posi col”.
  - iv. **A column for genotype calls**: Each element in this column is the genotype call of a SNP marker. It must be character. Users specify the index of the column in the entry “Call (A) col”.
  - v. **A column for intensities of allele A**: Each element in this column is the intensity of allele A of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(A)/Log2Ratio col”.
  - vi. **A column for intensities of allele B**: Each element in this column is the intensity of allele B of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(B)/Strength col”.

3. **Affymetrix Axiom platform:**

For this platform, users construct a directory “IndGeno” to store the text files (*.txt) of the samples.

- **Directory “IndGeno”**: The directory should contain text files (*.txt) of genotypes, $\log_2$ ratios and strengths of the samples. In each text file, there are at least 4 columns should be contained. Users specify the string representing NA (not available) and the number of skip rows of headers in the entry “NA string” and “Skip Row #”, respectively.
  - i. **A column for SNP IDs**: Each element in this column is the ID of a SNP marker. It must be character. Users specify the index of the column in the entry “SNP col”.
  - ii. **A column for genotype calls**: Each element in this column is the genotype call of a SNP marker. It must be character. Users specify the index of the column in the entry “Call (A) col”.
  - iii. **A column for intensities of allele A**: Each element in this column is the intensity of allele A of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(A)/Log2Ratio col”.
  - iv. **A column for intensities of allele B**: Each element in this column is the intensity of allele B of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(B)/Strength col”.

4. **Illumina platform:**

For this platform, users construct a directory “IndGeno” to store the text files (*.txt) of the samples.

- **Directory “IndGeno”**: The directory should contain text files (*.txt) of genotypes, genomic information and intensities of the samples. In each text file, there are at least 7 columns should be contained. Users specify the string representing NA (not available) and the number of skip rows of headers in the entry “NA string” and “Skip Row #”, respectively.
  - i. **A column for SNP IDs**: Each element in this column is the ID of a SNP marker. It must be character. Users specify the index of the column in the entry “SNP col”.
  - ii. **A column for chromosomes**: Each element in this column is the chromosome of a SNP marker. Users specify the index of the column in the entry “Chr col”.
  - iii. **A column for physical positions**: Each element in this column is the physical position of a SNP marker. It must be numeric. Users

specify the index of the column in the entry “Posi col”.

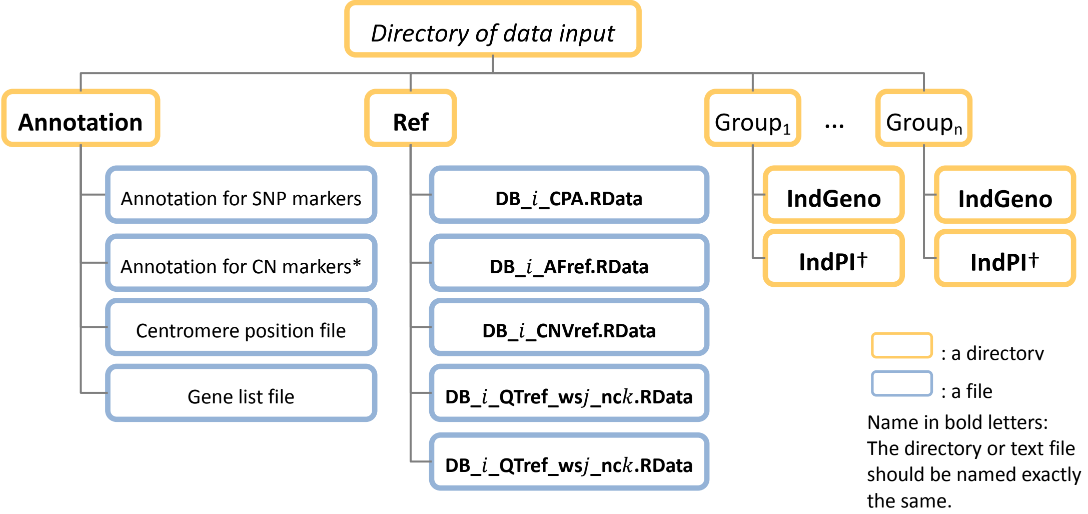
iv. **A column for genotype calls**: Each element in this column is the genotype call of a SNP marker. It must be character. Users specify the index of the column in the entry “Call (A) col”. If the genotype calls are separated into two columns, then users specify the index of column of genotype call of allele A and B in the entry “Call (A) col” and “Call (B) col”, respectively.

v. **A column for intensities of allele A**: Each element in this column is the raw intensity of allele A of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(A)/Log2Ratio col”.

vi. **A column for intensities of allele B**: Each element in this column is the raw intensity of allele B of a SNP marker. It must be numeric. Users specify the index of the column in the entry “Intensity(B)/Strength col”.

vii. **A column for allele frequencies of allele B (BAFs)**: Each element in this column is the allele intensity of allele B of a SNP marker, i.e. the BAF value that calculated by Illumina official software (e.g. GenomeStudio). It must be numeric between 0 and 1. Users specify the index of the column in the entry “BAF col”.

Figure 8. The structure of *Directory of data input* for *Genotype/Intensity-based* input data format.



*: only required when the *Genome-wide SNP array* is “Affymetrix: Array 6.0 (with CN data)”

†: not required only when the genome-wide SNPs array is “Affymetrix: Axiom”

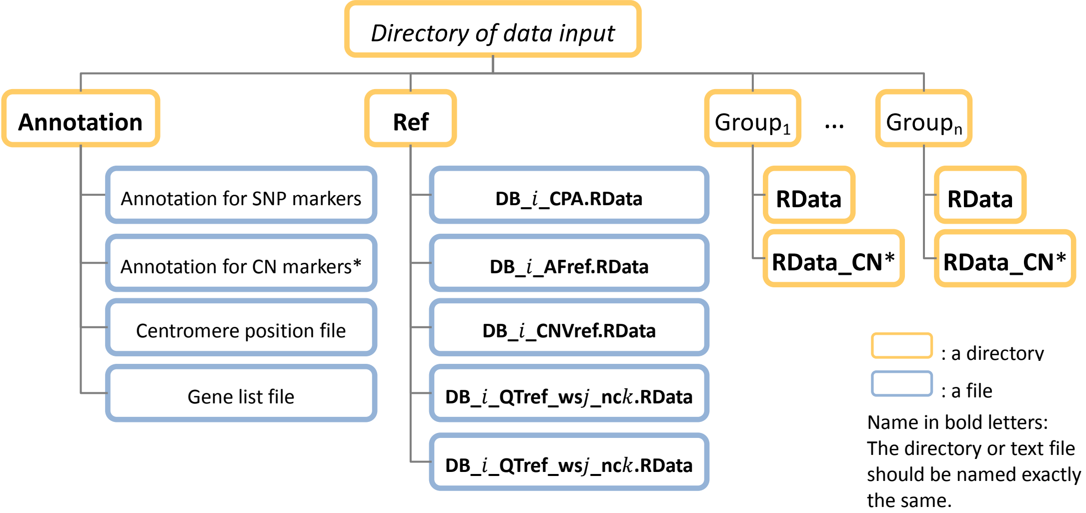
4.4 Directory required for RData-based input data format

For users who intended to run ALICE again with different parameter settings based on the same input data files, the use of raw R data from the previous ALICE analysis as the input data in the current analysis could skip the procedure of feature extraction and save the processing time. Prior to run ALICE analysis, users construct a directory "RData" under a group directory to store the raw R data (*.RData) of SNP markers of the samples in the group. If the *Genome-wide SNP array* is of "Affymetrix: Array 6.0 (with CN data)", users also should construct a directory "RData_CN" under a group directory to store the raw R data (*.RData) of CN markers of the samples in the group. How to copy raw R data of SNP and CN markers from the previous ALICE analysis to the directories of the current one are described as follows. The structure of *Directory of data input* for *RData-based* input data format is given in Figure 9. Note: Users need to select the icon "Yes" for "Save raw R data (*.RData)" on the graphical user interface in the previous ALICE analysis; otherwise, the directories "TEMP_Indiv" and "TEMP_Indiv_CN" will be deleted once ALICE analysis is done.

- **Directory "TEMP_Indiv" for SNP markers**: The raw R data of SNP markers that processed in the previous ALICE analysis were saved under the group directory(ies) under the "TEMP_Indiv" directory under the *Directory of result output* specified in the analysis. Users should copy the group directory(ies) of interest from that directory to the *Directory of data input* that specified for the current ALICE analysis. Then, users construct a directory "RData" under a group directory and move the RData files under that directory. For example, the path of *Directory of result output* specified in the previous ALICE analysis is "C:\ALICE\PreviousAnalysis\Output", and the path of *Directory of data input* specified for the current ALICE analysis is "C:\ALICE\CurrentAnalysis\Input". For the group "Demo" of interest, users copy the raw R data files of SNP markers from "C:\ALICE\PreviousAnalysis\Output\TEMP_Indiv\Demo" to "C:\ALICE\CurrentAnalysis\Input\Demo\RData".
- **Directory "TEMP_Indiv_CN" for CN markers**: The procedure for CN markers is similar to the one for SNP markers, in addition that a suffix "_CN" is added to the directory names used for SNP markers. Specifically, the raw R data of CN markers were saved under the group directory(ies) under the "TEMP_Indiv_CN" directory under the *Directory of result output* specified for the previous ALICE analysis. Users should copy the group directory(ies) of interest from that directory to the *Directory of data input* that specified for the current ALICE analysis. Then, users construct a directory "RData_CN" under a group directory and move the RData files under that directory. For

example, the path of *Directory of result output* specified for the previous ALICE analysis is “C:\ALICE\PreviousAnalysis\Output”, and the path of *Directory of data input* specified for the current ALICE analysis is “C:\ALICE\CurrentAnalysis\Input”. For the group “Demo” of interest, users copy the raw R data files of CN markers from “C:\ALICE\PreviousAnalysis\Output\TEMP_Indiv_CN\Demo” to “C:\ALICE\CurrentAnalysis\Input\Demo\RData_CN”.

Figure 9. The structure of *Directory of data input* for *RData-based* input data format.



*: only required when the *Genome-wide SNP array* is “Affymetrix: Array 6.0 (with CN data)”

5. Result Output Format for the component “Main Functions”.

This section describes the structure and format of the output files that generated during ALICE analysis. Section 5.1 described the structure of *Directory of result output* for the component “Main Functions”. Section 5.2 described the format of graphical output for the component “Genome Browser”. Section 5.3 described the format of graphical and numerical output for the component “Aberration Integration”.

5.1 Result Output Format for the component “Main Functions”.

The structure of *Directory of result output* is shown in Figure 10.

1. **Directory “APT_output”**

   This directory saves the output files generated during Affymetrix Power Tools (APT) analysis, therefore, this directory exists only when the input data format is of *CEL-based*. If users chose not to save the results, this directory will be deleted once ALICE analysis is done.

2. **Directory “TEMP” and “TEMP_CN”**

   ALICE constructs a directory “TEMP” (“TEMP_CN”) to temporarily store the intermediate results of SNP (CN) markers during ALICE calculation, and will be deleted once ALICE calculation is finished

3. **Directory “TEMP_Indiv” and “TEMP_Indiv_CN”**

   Under the directory “TEMP_Indiv” (“TEMP_Indiv_CN”), ALICE constructs a group directory for each group of input data to store the raw R data of SNP (CN) markers of each individual in the group that generated after feature extraction procedure. The name of a group directory is given according to the directory name of the source group directory of input data. If users chose “No” for the icon “Save raw R data (*.RData)” on the ALICE interface, the directories “TEMP_Indiv” and “TEMP_Indiv_CN” will be deleted once ALICE calculation is finished.

4. **Description.txt**

   This text file records the parameter setting that users selected on the interface for ALICE analysis.

5. **log.txt**

   This file updates the progress of ALICE analysis for users to monitor the progress in a real time.

6. **Group directory(ies)**

   It is allowed to have multiple group directories constructed under the *Directory of data input*. Besides the differences in the directory names, the structure of the directories and files are the same for each group directory. Under a group directory, ALICE constructs a directory for each individual sample in the source group directory of input data, named sample directory. The name of a sample directory is given by capitalizing the ID of the source sample after the removal of any of the strings (case-insensitive): Txt, Cel, Rdata, Hind, Xba, Nsp, Sty, Birdseed, Brlmm, Br, CN5, CNCHP. For example, the ID of source sample data is “Sample1_BR.txt”, and then the name of the sample directory will be “SAMPLE1”. The structure of the directories and files under a group directory is shown in Figure 11.

(1) **samplelist.txt**: The file contains the indices, IDs and genetic genders of the individual samples in the group directory. Each row is the information of a sample. The content of the three columns are described in the following.

- **Column 1 – “Index”**: The indices of samples.
- **Column 2 – “Sample_ID”**: The IDs of samples.
- **Column 3 – “Gender”**: The estimated genetic genders of samples. If the ratio of heterozygous SNPs over called SNPs on chromosome X that featured on the array is smaller than 0.1, the genetic gender is estimated to be male (“M”); otherwise, it is female (“F”).

(2) Sample direcoty(ies)

i. Directory “FigData”

This directory stores the .RData files of the analysis results that ALICE generated for any further analysis by using Component 2 and 3 of ALICE.

ii. ***SampleID*_SixPanel.tiff**

The six-panel figure of whole-genome SNP markers is named by the ID of source sample with suffix “_SixPanel.tiff”.

iii. Chromosome-level numerical output

The chromosome-level files include “Chr01.txt”, “Chr02.txt”, …, “Chr22.txt” and “ChrX.txt”. The columns in a chromosome-level file are introduced in the following. Each row of the file is the information of a SNP marker.

- **Columns “SNP_ID”, “Chr”, “PhyPosi”, “CumPP” and “Genotype”**: The IDs, chromosomes, physical positions, cumulative positions and genotypes of SNP markers, respectively.
- **Column “AF”**: The allele frequencies of allele A of the SNP marker. If the Genome-wide SNP array is from Affymetrix

platforms, the AF estimates are CPA+LIM-adjusted.

- **Column "QC_rm_index"**: The index representing the state of a SNP in the quality control (QC) procedure.
  - **The index = 0**: The SNP passed the QC procedure;
  - **The index = 1**: no reference database for the SNP;
  - **The index = 2**: unreliable reference database, i.e. the call rate of reference samples used to construct AF database for the SNP is < 95%, or the availability of HI values of reference samples used to construct CNV/CNA database for the SNP is < 95%;
  - **The index = 3**: the intensity of either one or both alleles of the SNP marker is unavailable;
  - **The index = 4**: the SNP marker is located on chromosome Y for a female sample.

  If the value of "QC_rm_index" of a SNP marker is not 0, i.e. not passed the quality controls, then the SNP will not be analyzed. That is, besides the columns of genomic information ("SNP_ID", "Chr", "PhyPosi", "CumPP", "Genotype") and "QC_rm_index", the rest columns of the SNP will be "NA" (the string stands for unavailable in R).

  Note: A marker is excluded from the report of ALICE if it is a control marker designed by the array platform, or located on an unknown location or mitochondrion.
- **Columns "CNV_s_value"**: The HI values of SNP markers.
- **Columns "AI_s_index" and "LOH_s_index"**: The aberrant indices of SNP markers from single-point AI and LOH/LCSH analysis, respectively.
- **Columns "CNV_s_stat", "CNV_s_pv" and "CNV_s_index"**: The statistic values, p-values and aberrant indices of SNP markers from single-point CNV/CNA analysis, respectively.
- **Columns "AI_m_prop", "AI_m_pv" and "AI_m_index"**: The WAP values, p-values and aberrant indices of SNP markers from multi-point AI analysis, respectively.
- **Columns "LOH_m_prop", "LOH_m_pv" and "LOH_m_index"**: The WAP values, p-values and aberrant indices of SNP markers from multi-point LOH/LCSH analysis, respectively.
- Columns "**CNV_m_prop_Gain**", "**CNV_m_pv_Gain**", "**CNV_m_prop_Del**", "**CNV_m_pv_Del**", and "**CNV_m_index**":

The WAP values and p-values for the detection of CN gain, WAP value and p-values for the detection of CN deletion, and aberrant indices of SNP markers from multi-point CNV/CNA analysis, respectively.

iv. Individual-level segment files

The individual-level segment files include the following types.

- **Based on single-point analysis results**
  - **"Indiv_AI_single.txt", "Indiv_LOH_single.txt" and "Indiv_CNV_single.txt"**: The segment file contains the segments detected using the single-point AI, LOH/LCSH and CNV/CNA analysis of ALICE, respectively;
- **Based on multi-point analysis results**
  - **"Indiv_AI_multi.txt", "Indiv_LOH_multi.txt" and "Indiv_CNV_multi.txt"**: The segment file contains the segments detected using the multi-point AI, LOH/LCSH and CNV/CNA analysis of ALICE, respectively;
  - **"Indiv_CNV_multi_CBS.txt"**: The segment file contains the segments detected using the specified CBS algorithm.

The content of the columns in a segment file are described in the following. Each row is the information of a segment.

- **Column "Chr"**: The chromosomes that the segments are located on.
- **Column "Start_PhyPosi" and "End_Marker"**: The physical positions of the starting and ending marker of segments, respectively.
- **Column "Start_Marker" and "End_PhyPosi"**: The IDs of the starting and ending marker of segments, respectively.
- **Column "Total_Num_Of_Sig_Markers"**: The total numbers of consecutive significant markers that the segments comprised of.
- **Column "AssociatedGene"**: The list of the gene(s) that the locations of segments associated with. If there is more than one gene, the genes are separated by comma (,).
- **Column "Window_size"**: The window size that users specified for the multi-point analysis. This column will be generated only when the segment files are based on

multi-point analysis.

- **Column “Abnormality”**: The abnormality types of segments detected in the multi-point CNV/CNA analysis. This column will be generated only when the segment files are based on multi-point CNV/CNA analysis.
- **Column “Index_CNV_AIorLOH”**: The indices of AI/LOH/LCSH+CNV/CNA aberrant of segments, which equal to 1 if the markers comprises a segment are significant in the multi-point CNV/CNA analysis and more than half of them are also significant in the multi-point AI or LOH/LCSH analysis; otherwise, the value is 0. This column will be generated only when the segment files are based on multi-point CNV/CNA analysis.

v. Directory “SNP_CN_output”

This directory contains the results integrated SNP markers with CN markers of a sample. ALICE adds the suffix “(SNP+CN)” to the name of a figure or text file to label it’s an integrated result. This directory will be generated only when the input data format is Array 6.0 *CEL-based* platform.

I. ***SampleID*_SixPanel(SNP+CN).tiff**: The six-panel figure of whole-genome SNP and CN markers of the sample with ID noted in the figure name. The markers are sorted according to their chromosomes and physical positions.

II. Chromosome-level numerical output files

The chromosome-level files includes “Chr01(SNP+CN).txt”, “Chr02(SNP+CN).txt”, …, “Chr22(SNP+CN).txt” and “ChrX(SNP+CN).txt”. In addition to the columns that aforementioned for the chromosome-level file of SNP markers, a chromosome-level file of SNP and CN markers contains 3 additional columns, as described in the following. Each row is the information of a marker. The markers are sorted according to their chromosomes and physical positions.

- ■ **Column “Marker”**: The marker types of markers (SNP or CN).
- ■ **Columns “SNP_ID”**: The IDs of SNP markers. If the marker is a CN marker, then the element will be “NA”.
- ■ **Columns “SNP_ID”, “CN_ID”**: The IDs of CN markers. If

the marker is a SNP marker, then the element will be "NA".

III. **Individual-level segment files**

The columns are the same as the aforementioned description for the segment files based on SNP markers only.

(3) **Cross-sample figures**

There are several optional figures that uses select to output.

i. **"AI_s_index_all.tiff" ("AI_m_index_all.tiff")**: The cross-sample figure represents the single-point (multi-point) AI analysis results across all samples in a group directory.

ii. **"LOH_m_index_all.tiff"**: The cross-sample figure represents the multi-point LOH/LCSH analysis results across all samples in a group directory.

iii. **"CNV_s_index_all.tiff" ("CNV_m_index_all.tiff")**: The cross-sample figure represents the single-point (multi-point) CNV/CNA analysis results on SNP markers across all samples in a group directory.

iv. **"CNV_s_index_all (SNP+CN).tiff" ("CNV_m_index_all(SNP+CN).tiff")** : The cross-sample figure represents the single-point (multi-point) CNV/CNA analysis results on SNP and CN markers across all samples in a group directory.

(4) **Cross-sample and cross-genome numerical output:**

For each cross-sample figure, there are two numerical text files will be generated.

i. **Cross-genome numerical output files**: For a cross-sample figure, the text file is named after the name of the figure with the suffix "_Prop_CrossMarkers". For example, the file name "AI_m_index_all_Prop_CrossMarkers.txt" is the cross-genome numerical report of the cross-sample figure "AI_m_index_all.tiff". A cross-genome numerical output file contains at least 4 columns, which are introduced in the following. Each row is the information of a sample provided in a group directory.

- **Column "Index"**:The indices of sample in a group directory.
- **Columns "Sample_ID" and "Gender"**: The sample IDs and estimated genetic genders of samples, respectively.
- **Column "Prop_CrossChr"**: Each element in the column is the proportion (%) of a marker of AI or LOH/LCSH over whole-genome markers. ALICE only generates this column

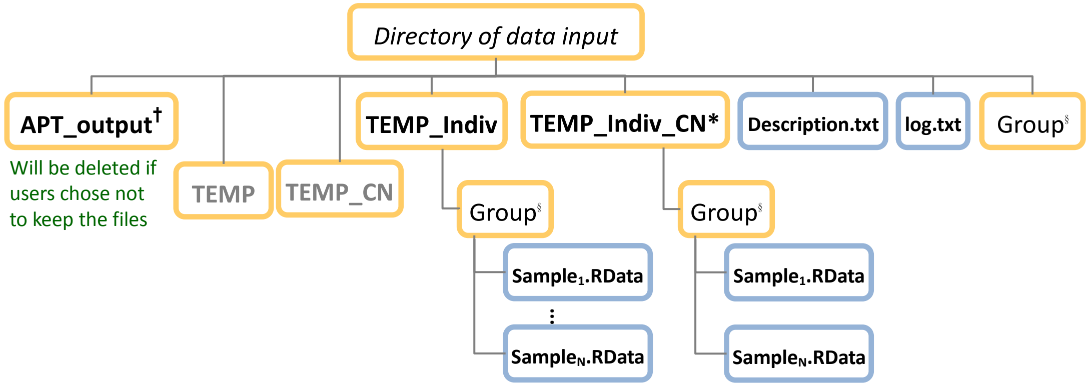
when the aberration is of AI or LOH/LCSH.

- **Columns "Prop_Gain_CrossChr" and "Prop_Loss_CrossChr"**: Each element in the column "Prop_Gain_CrossChr" and "Prop_Loss_CrossChr" is the proportion (%) of a marker of CN-gain and CN-loss over whole-genome markers, respectively. ALICE only generates the columns when the aberration is of CN-gain or CN-loss, respectively.

ii. **Cross- sample numerical output files**: For a cross-sample figure, the text file is named after the name of the figure with the suffix "_Prop_CrossSamples". For example, the file name "AI_m_index_all_Prop_CrossSamples.txt" is the cross-sample numerical report of the cross-sample figure "AI_m_index_all.tiff". A cross-sample numerical output file contains at least 4 columns, which are introduced in the following. Each row is the information of a marker that featured on the SNP array.

- **Columns "SNP_ID" and "CN_ID"**: If a marker is a SNP marker, the ID will be listed in the column "SNP_ID"; otherwise, it will be listed in the column "CN_ID". ALICE only generates the columns when the *Genome-wide SNP array* is of "Affymetrix: Array 6.0 (with CN data)" and *Input data format* is of "CEL-based".
- **Columns "Chr" and "PhyPosi"**: The chromosomes and physical positions of markers.
- **Column "Prop_CrossSamples"**: Each element in the column is the proportion (%) of a sample with aberrant marker of AI or LOH/LCSH over the samples in the same group directory. ALICE only generates this column when the aberration is of AI or LOH/LCSH.
- **Columns "Prop_Gain_CrossSamples" and "Prop_Loss_CrossSamples"**: Each element in the column "Prop_Gain_CrossSamples" and "Prop_Loss_CrossSamples" is the proportion (%) of a sample with aberrant marker of CN-gain and CN-loss over the samples in the same group directory, respectively. ALICE only generates the columns when the aberration is of CN-gain or CN-loss.

Figure 10. The structure of *Directory of result output*.



- Name in bold letters: The directory or text file should be named exactly the same.
- Name colored in grey: The directory will be deleted once ALICE calculation is done.

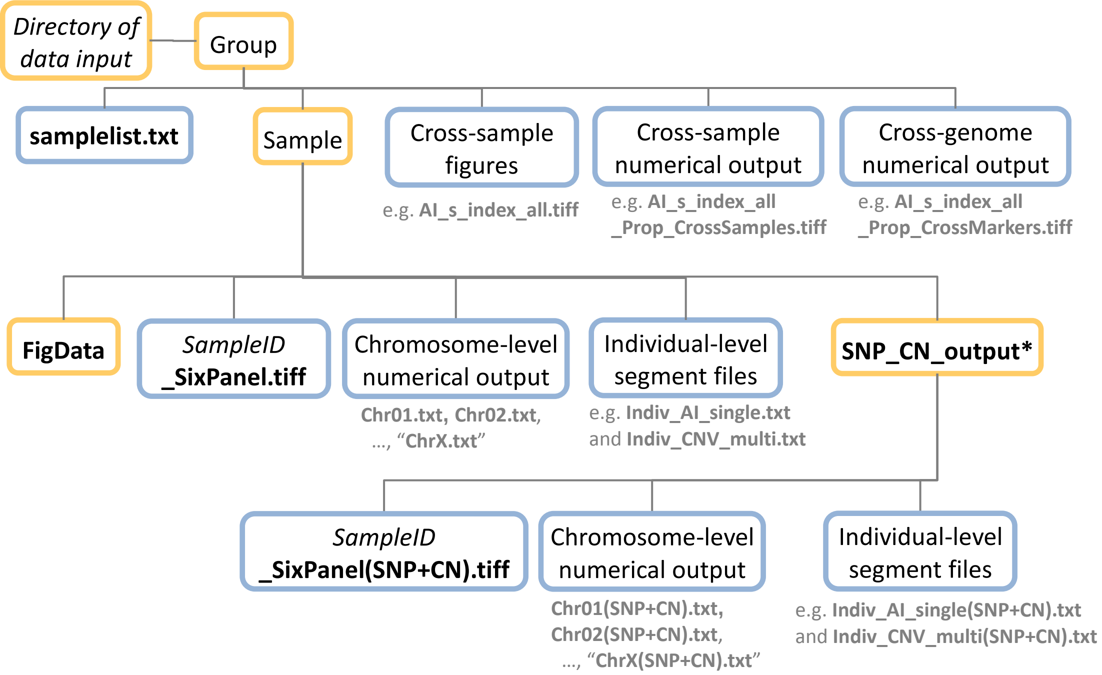
*: only constructed when the *Genome-wide SNP array* is "Affymetrix: Array 6.0 (with CN data)"

†: only constructed when the input data format is *CEL-based*

§: Allowed multiple group directories existed under the *Directory of data input*.

🟨 : a directory
🟦 : a file

Figure 11. The structure of a group directory.


Directory of data input
Group
samplelist.txt
Sample
Cross-sample figures
e.g. AI_s_index_all.tiff
Cross-sample numerical output
e.g. AI_s_index_all_Prop_CrossSamples.tiff
Cross-genome numerical output
e.g. AI_s_index_all_Prop_CrossMarkers.tiff
FigData
SampleID_SixPanel.tiff
Chromosome-level numerical output
Chr01.txt, Chr02.txt, ..., "ChrX.txt"
Individual-level segment files
e.g. Indiv_AI_single.txt and Indiv_CNV_multi.txt
SNP_CN_output*
SampleID_SixPanel(SNP+CN).tiff
Chromosome-level numerical output
Chr01(SNP+CN).txt, Chr02(SNP+CN).txt, ..., "ChrX(SNP+CN).txt"
Individual-level segment files
e.g. Indiv_AI_single(SNP+CN).txt and Indiv_CNV_multi(SNP+CN).txt

- Name in bold letters: The directory or text file should be named exactly the same.
- Name colored in grey: The directory will be deleted once ALICE calculation is done.

*: only constructed when the *Genome-wide SNP array* is "Affymetrix: Array 6.0 (with CN data)"

🟨 : a directory
🟦 : a file

5.2 Result Output Format for the component "Genome Browser".

For each sample that users specified to perform visualization, ALICE generates a zoom-in figure and saves it to the target sample directory under the *Directory of result output.* For the target sample in the sample directory I, the zoom-in multi-panel figure is named as “I_Chr$R$_P_s-P_e.tiff”, where R, P_s and P_e are the chromosome, starting and ending positions of the genomic region of interest that specified for the sample.

5.3 Result Output Format for the component “Aberration Integration”.

For each sample that users specified to perform aberration integration, ALICE re-generates graphical and numerical output based on the integrated results. The output files will be saved to the target sample directory under the *Directory of result output.*

- **Graphical output**

  Given users chose to perform integration onto the aberrations S based on the results of chromosome R from the multi-point analysis for the sample in the sample directory I, the zoom-in multi-panel figure is named as “I_Chr R_P_s-P_e(S).tiff”, where P_s and P_e are the starting and ending positions of the chromosome and S = “AI_LOH”, “AI_CNV”, “LOH_CNV” and “AI_LOH_CNV” for the selected *Analyses to be integrated* is “AI+LOH/LCSH”, “AI+CNV/CNA”, “LOH/LCSH+CNV/CNA” and “AI+LOH/LCSH+CNV/CNA”, respectively. In the figure, the top bars in the selected panels will be colored according to the integrated detection results. If the marker is not significant in all of the selected *Analyses to be integrated*, then it will be colored in blue in the top bar.

- **Numerical output**

  Given users chose to perform integration onto the aberrations S based on the results of chromosome R from the T analysis of the marker type M for the sample in the sample directory I, the segment file of integrated aberrations is named as "Indiv_S_T(M)_Chr$R$.txt”, where S = “AI_LOH”, “AI_CNV”, “LOH_CNV” and “AI_LOH_CNV” for the selected *Analyses to be integrated* is “AI+LOH/LCSH”, “AI+CNV/CNA”, “LOH/LCSH+CNV/CNA” and “AI+LOH/LCSH+CNV/CNA”, respectively, T = “s” or “m” for single-point and multi-point result analysis and M = “SNPonly” or “SNP+CN”. If a marker is not significant in all of the selected *Analyses to be integrated*, the aberrant index will be changed to 0. Then, ALICE will re-scan on the genome-wide scale to generate the segment file based on the integrated results of selected aberrations.

6. Examples

This section described the procedures of running ALICE on the examples provided on ALICE website. Firstly, we describe the analysis of normal samples. Two normal samples genotyped with Array 6.0 and Axiom platforms were used in Example set 1 and Example set 2, respectively, to illustrate the procedures for the two platforms. Specifically, Example set 1 displays the analysis procedure of RData-based data of SNP and CN markers genotyped using Affymetrix Array 6.0 platform, and Example set 2 shows the analysis procedure of Genotype/Intensity-based data of SNP markers genotyped using Affymetrix Axiom platform. Secondly, we focus on the analysis of cancer patients. In Example set 3, the data set contains the RData files of a metastatic small-cell lung cancer cell line and the corresponding blood cell line from a lung cancer patient, and an admixed sample at an admixture proportion of 50% (i.e. 50% of cancer cell line and 50% of blood cell line of the patient). ALICE provides unpaired and paired analysis. The procedure of unpaired analysis for cancer patients is the same for normal samples. Therefore, we only give an example of running unpaired analysis with the first component in Example 3-1. On the other hand, Example 3-2 demonstrates the procedure of paired analysis for the admixed sample by treating the corresponding blood cell line of the cancer patient as the matched control to them. The data set used in Example set 3 was genotyped using Affymetrix Axiom platform. Here we used “C:/ALICE” as the working directory, which users can change it to any destination directory.

6.1 Example set 1

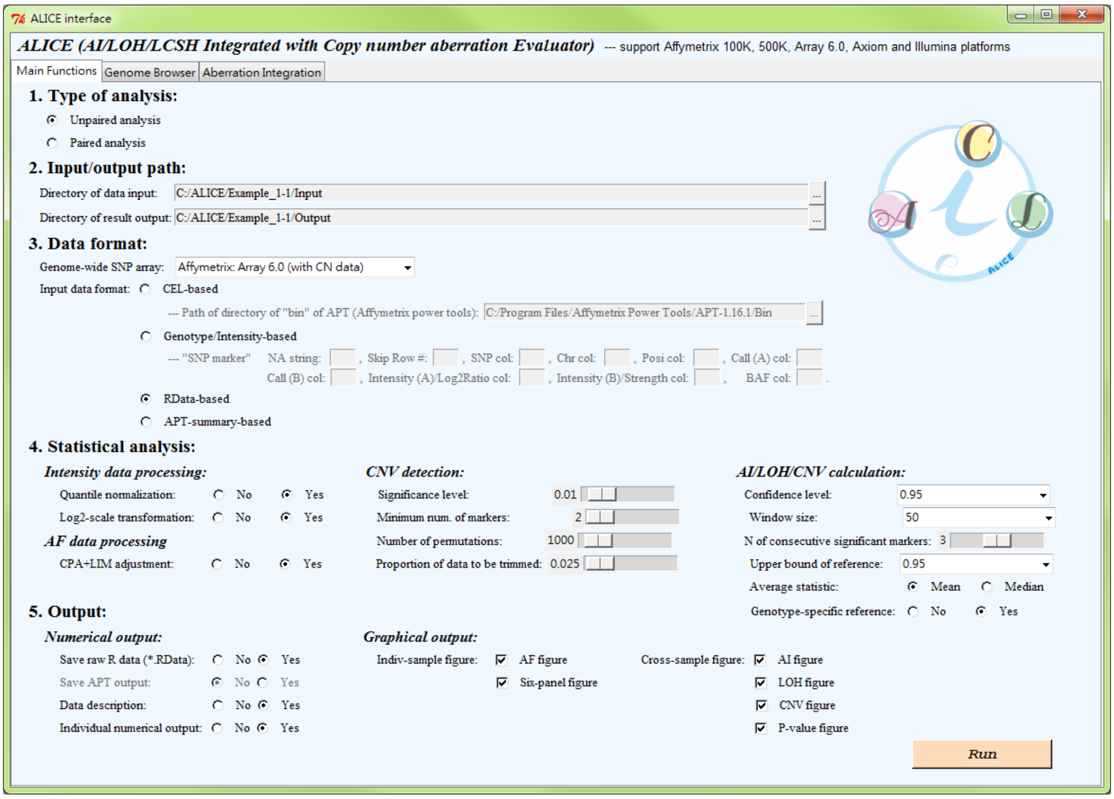
The first example set includes three examples: Example 1-1, Example 1-2 and Example 1-3. Example 1-1 analyzes two samples that genotyped with Affymetrix Array 6.0 platform using the components “Main Functions” to run ALICE analysis. Focused on the aberrant regions found in the results from Example 1-1, Example 1-2 illustrates how to zoom-in the regions using the components “Genome Browser” and Example 1-3 illustrates how to investigate the regions by integrating aberrations using the components “Aberration Integration”. The further investigations of Example 1-2 and 1-3 are done by using *Single-sample visualization*.

6.1.1 Example 1-1 for the unpaired analysis of normal samples genotyped with Affymetrix Array 6.0 platform using the component “Main Functions”

The procedures of analyzing the RData-based input data files of two samples genotyped using Affymetrix Array 6.0 platform are listed in the following.

- Step 1 – Download the zip file of the data used for Example 1-1 “Example_1-1.zip” from ALICE website and decompress the files into the working directory.
- Step 2 – Download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of the component “Main Functions” will pop up (see Figure 1).
- Step 4 – Type “Example 1” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 1-1. ALICE will automatically set the parameters required for the example (see Figure 12).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_1-1\Output”. The whole-genome six-panel figures of the two samples from the graphical output are shown in Figure 13 and Figure 14.

Figure 12. Parameter setting for Example 1-1.



ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

1. Type of analysis:

● Unpaired analysis
○ Paired analysis

2. Input/output path:

Directory of data input: C:/ALICE/Example_1-1/Input
Directory of result output: C:/ALICE/Example_1-1/Output

3. Data format:

Genome-wide SNP array: Affymetrix: Array 6.0 (with CN data)
Input data format: ○ CEL-based
--- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin
○ Genotype/Intensity-based
--- "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) col:
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .
● RData-based
○ APT-summary-based

4. Statistical analysis:

Intensity data processing:
Quantile normalization: ○ No ● Yes
Log2-scale transformation: ○ No ● Yes

AF data processing
CPA+LIM adjustment: ○ No ● Yes

CNV detection:
Significance level: 0.01
Minimum num. of markers: 2
Number of permutations: 1000
Proportion of data to be trimmed: 0.025

AI/LOH/CNV calculation:
Confidence level: 0.95
Window size: 50
N of consecutive significant markers: 3
Upper bound of reference: 0.95
Average statistic: ● Mean ○ Median
Genotype-specific reference: ○ No ● Yes

5. Output:

Numerical output:
Save raw R data (*.RData): ○ No ● Yes
Save APT output: ● No ○ Yes
Data description: ○ No ● Yes
Individual numerical output: ○ No ● Yes

Graphical output:
Indiv-sample figure: ☑ AF figure
☑ Six-panel figure
Cross-sample figure: ☑ AI figure
☑ LOH figure
☑ CNV figure
☑ P-value figure

Figure 13. The whole-genome six-panel figures of the first sample (Sample1) used in Example 1-1.

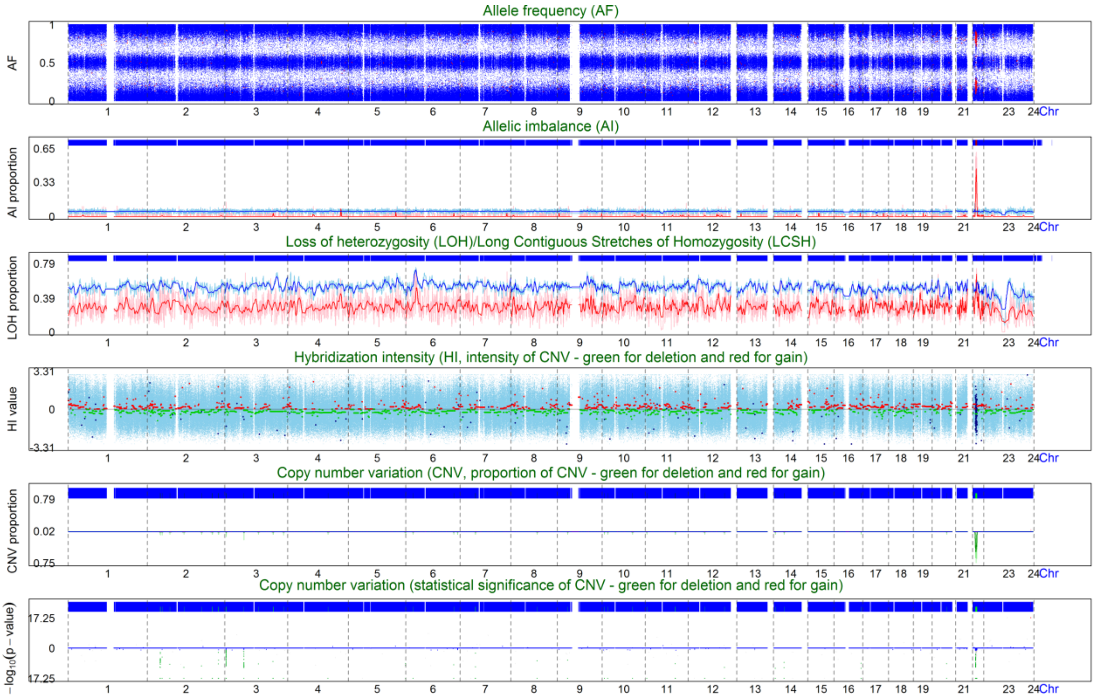
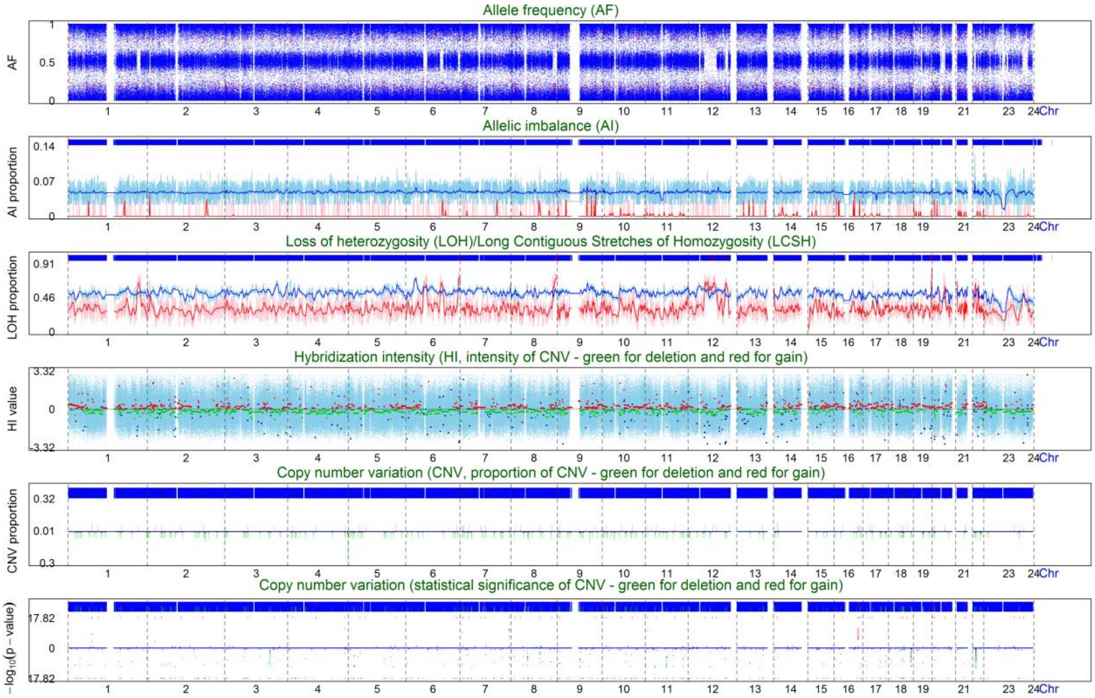


Figure 14. The whole-genome six-panel figures of the second sample (Sample2) used in Example 1-1.

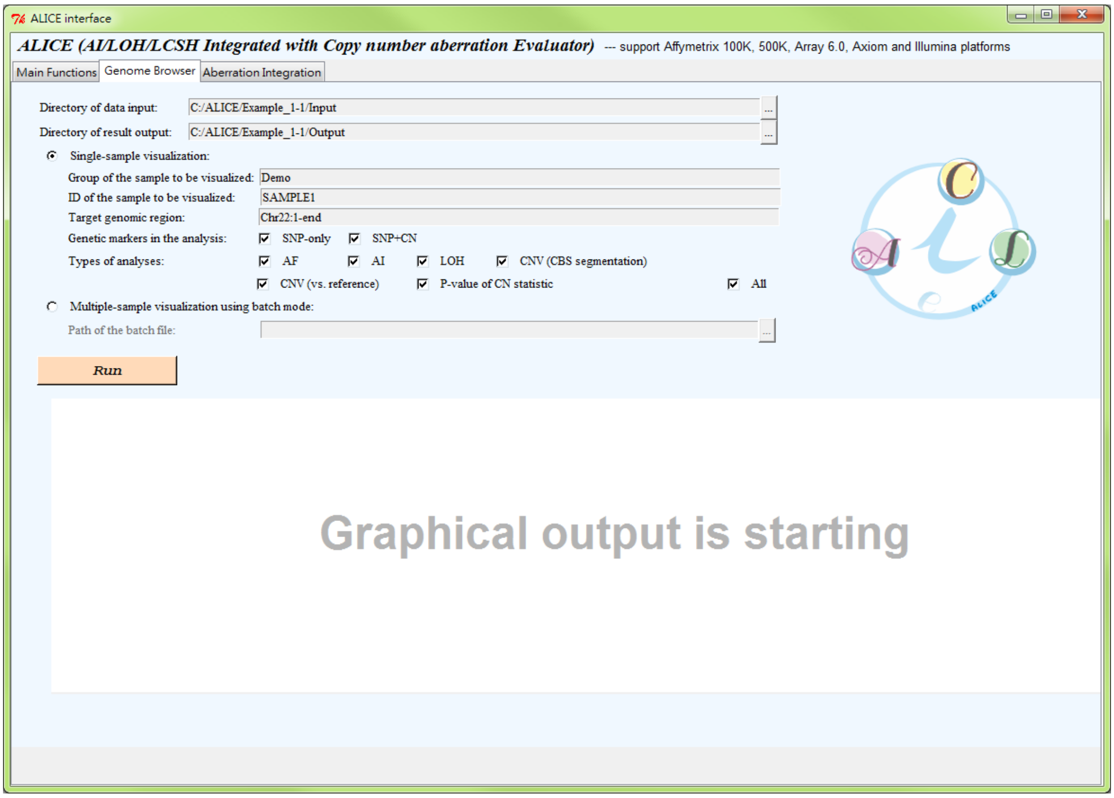


6.1.2 Example 1-2 for the unpaired analysis of normal samples genotyped with Affymetrix Array 6.0 platform using the component “Genome Browser”

From the results of Example 1-1, we observed AI, LOH/LCSH and CN aberrations occurred in the 22$_{th}$ chromosome of the first sample. Aimed to zoom in the chromosome through *Single-sample visualization*, the procedures of using functions provided in the 2nd component “Genome Browser” are listed in the following.

- Step 1 – If users already run Example 1-1, then this step can be skipped. If not, users need to download the zip file “Example_1-2.zip” from ALICE website, which is a compressed file which contains the least data required for Example 1-2. Users decompress the files into the working directory.
- Step 2 – If users already downloaded “ALICE.exe”, then this step can be skipped. If not, users download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of ALICE will pop up. Users click the tab “Genome Browser” to switch to the interface (see Figure 2).
    - Step 4 – Type “Example 1” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 1-2. ALICE will automatically set the parameters required for the example (see Figure 15).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_1-1\Output\Demo\SAMPLE1” if users already run analysis of Example 1-1. Otherwise, the results will be saved to “C:\ALICE\Example_1-2\Output\Demo\ SAMPLE1” if users run this example based on “Example_1-2.zip”. The six-panel figure of the target genomic region of the first sample (Sample1) from the graphical output is shown in Figure 16.

Figure 15. Parameter setting for Example 1-2.



ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Directory of data input: C:/ALICE/Example_1-1/Input

Directory of result output: C:/ALICE/Example_1-1/Output

● Single-sample visualization:

Group of the sample to be visualized: Demo

ID of the sample to be visualized: SAMPLE1

Target genomic region: Chr22:1-end

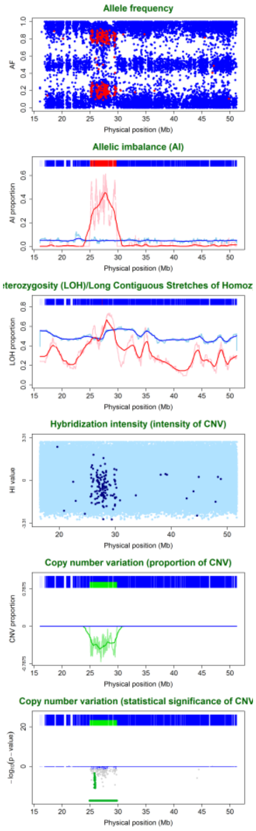
Genetic markers in the analysis: ☑ SNP-only ☑ SNP+CN

Types of analyses: ☑ AF ☑ AI ☑ LOH ☑ CNV (CBS segmentation) ☑ CNV (vs. reference) ☑ P-value of CN statistic ☑ All

○ Multiple-sample visualization using batch mode:

Path of the batch file:

Figure 16. The six-panel figure of the target genomic region of the first sample (Sample1).

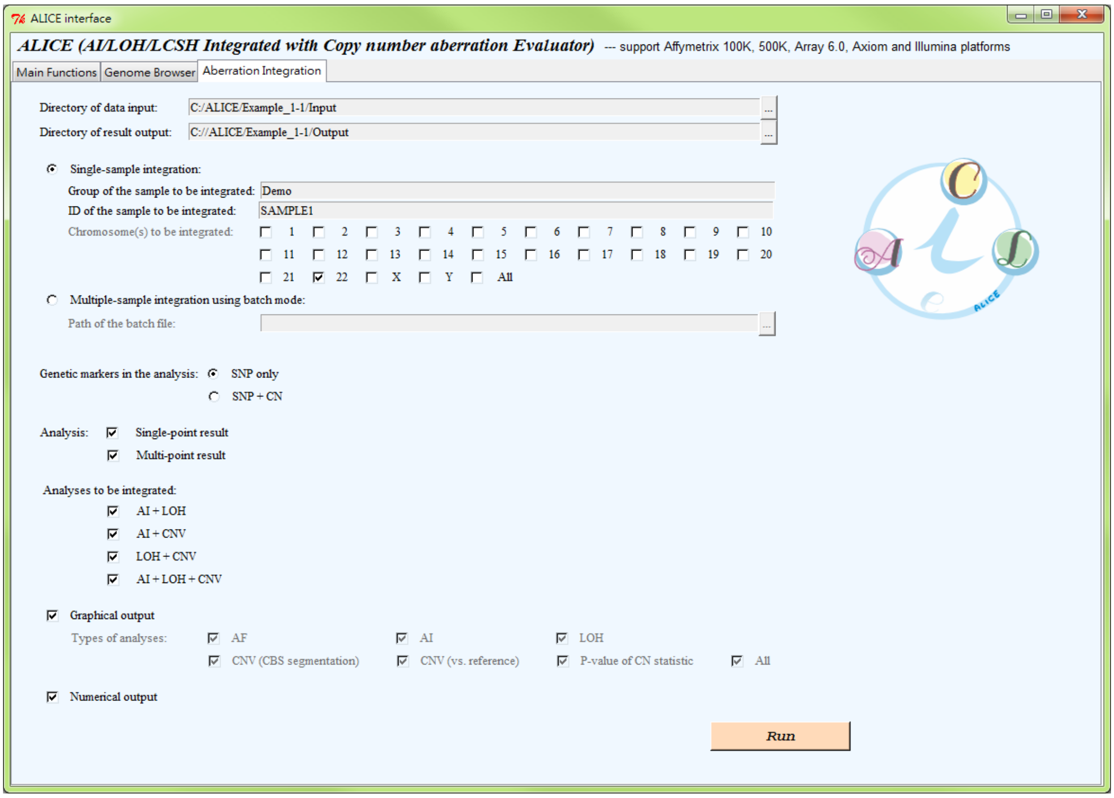


6.1.3 Example 1-3 for the unpaired analysis of normal samples genotyped with Affymetrix Array 6.0 platform using the component “Aberration Integration”

Furthermore, we integrated two or three out of AI, LOH/LCSH and CNV/CNA aberrations that identified on the 22_{th} chromosome of the first sample based on the results of Example 1-1. The procedures through *Single-sample visualization* using the functions provided in the 3^{rd} component “Aberration Integration” to achieve the goal are listed in the following.

- Step 1 – If users already run Example 1-1, then this step can be skipped. If not, users need to download the zip file “Example_1-3.zip” from ALICE website, which is a compressed file which contains the least data required for Example 1-3. Users decompress the files into the working directory.
- Step 2 – If users already downloaded “ALICE.exe”, then this step can be skipped. If not, users download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 –Right-click on the ALICE executable file to run it as administrator, then the interface of ALICE will pop up. Users click the tab “Aberration Integration” to switch to the interface (see Figure 3).
- Step 4 – Type “Example 1” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 1-3. ALICE will automatically set the parameters required for the example (see Figure 17).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_1-1\Output\Demo\SAMPLE1” if users already run analysis of Example 1-1. Otherwise, the results will be saved to “C:\ALICE\Example_1-3\Output\Demo\ SAMPLE1” if users run this example based on “Example_1-3.zip”. The six-panel figures of the target genomic region of the first sample (Sample1) from the graphical output, which integrated the analyses of AI and CNV, are shown in Figure 18.

Figure 17. Parameter setting for Example 1-3.



ALICE interface

ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Directory of data input: C:/ALICE/Example_1-1/Input

Directory of result output: C://ALICE/Example_1-1/Output

● Single-sample integration:

Group of the sample to be integrated: Demo

ID of the sample to be integrated: SAMPLE1

Chromosome(s) to be integrated: □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10 □ 11 □ 12 □ 13 □ 14 □ 15 □ 16 □ 17 □ 18 □ 19 □ 20 □ 21 ☑ 22 □ X □ Y □ All

○ Multiple-sample integration using batch mode:

Path of the batch file:

Genetic markers in the analysis: ● SNP only ○ SNP + CN

Analysis: ☑ Single-point result ☑ Multi-point result

Analyses to be integrated:

☑ AI + LOH

☑ AI + CNV

☑ LOH + CNV

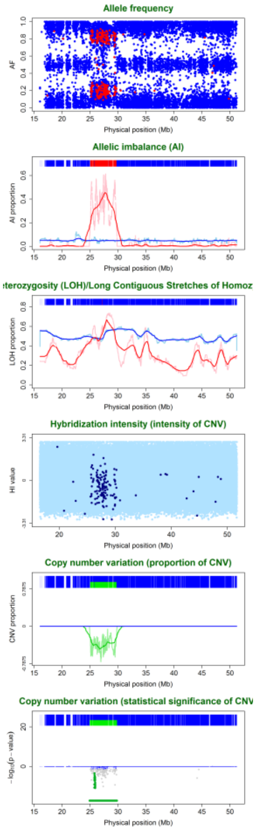
☑ AI + LOH + CNV

☑ Graphical output

Types of analyses: ☑ AF ☑ AI ☑ LOH ☑ CNV (CBS segmentation) ☑ CNV (vs. reference) ☑ P-value of CN statistic ☑ All

☑ Numerical output

Figure 18. The six-panel figures of the target genomic region of the first sample (Sample1).



6.2 Example set 2

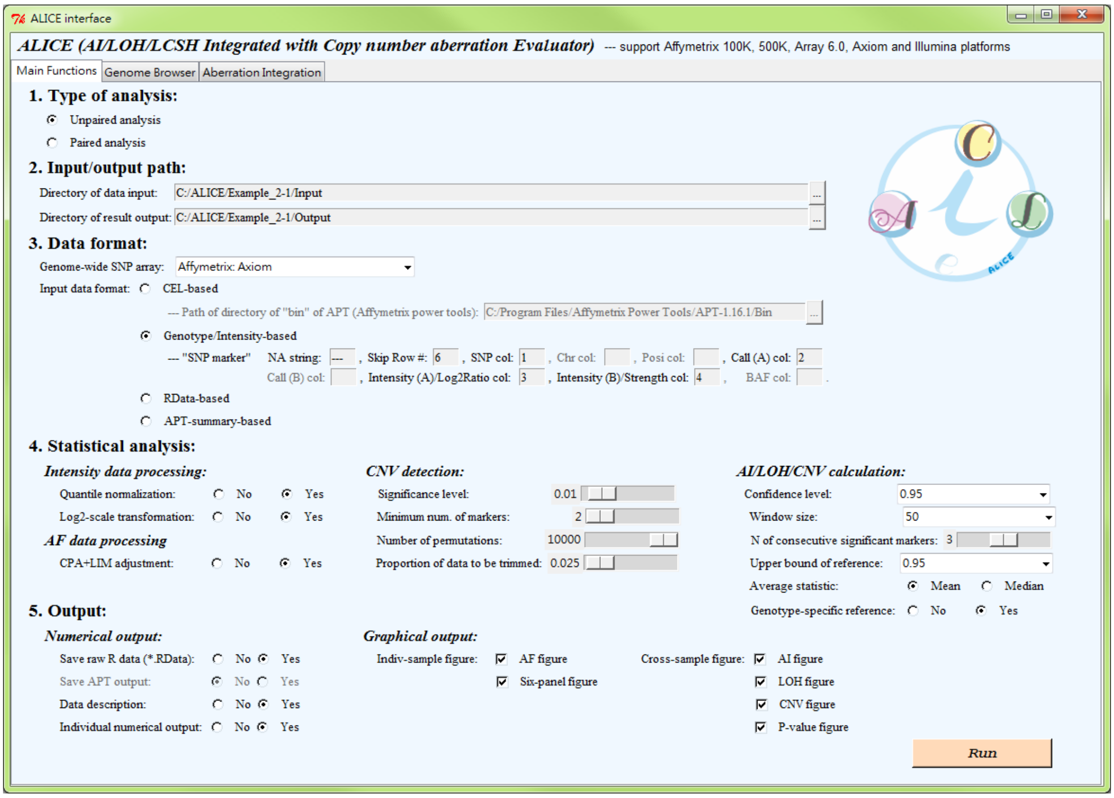
The second example set includes three examples: Example 2-1, Example 2-2 and Example 2-3. This example set based on the same two samples that used in first example set, but genotyped them using Affymetrix Axiom platform. The analysis procedures are similar as used in the first example set. Example 2-1 analyzes the two samples using the components “Main Functions” to run ALICE analysis. Focused on the aberrant regions found in the results from Example 2-1, Example 2-2 illustrates how to zoom-in the regions using the components “Genome Browser” and Example 2-3 illustrates how to investigate the regions by integrating aberrations using the components “Aberration Integration”. The further investigations of Example 2-2 and 2-3 are done by using *Multi-sample visualization using batch mode*.

6.2.1 Example 2-1 for the unpaired analysis of normal samples genotyped with Affymetrix Axiom platform using the component “Main Functions”

The procedures of analyzing the Genotype/Intensity-based input data files of two samples genotyped using Affymetrix Axiom platform are listed in the following.

- Step 1 – Download the zip file of the data used for Example 2-1 “Example_2-1.zip” from ALICE website and decompress the files into the working directory.
- Step 2 – If users already downloaded “ALICE.exe”, then this step can be skipped. If not, users download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of the component “Main Functions” will pop up (see Figure 1).
- Step 4 – Type “Example 2” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 2-1. ALICE will automatically set the parameters required for the example (see Figure 19).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_2-1\Output”. The whole-genome six-panel figures of the two samples from the graphical output are shown in Figure 20 and Figure 21.

Figure 19. Parameter setting for Example 2-1.



ALICE interface

ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Main Functions | Genome Browser | Aberration Integration

1. Type of analysis:

● Unpaired analysis
○ Paired analysis

2. Input/output path:

Directory of data input: C:/ALICE/Example_2-1/Input
Directory of result output: C:/ALICE/Example_2-1/Output

3. Data format:

Genome-wide SNP array: Affymetrix: Axiom
Input data format: ○ CEL-based
--- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin
● Genotype/Intensity-based
--- "SNP marker" NA string: --- , Skip Row #: 6 , SNP col: 1 , Chr col: , Posi col: , Call (A) col: 2
Call (B) col: , Intensity (A)/Log2Ratio col: 3 , Intensity (B)/Strength col: 4 , BAF col: .
○ RData-based
○ APT-summary-based

4. Statistical analysis:

Intensity data processing:
Quantile normalization: ○ No ● Yes
Log2-scale transformation: ○ No ● Yes
AF data processing
CPA+LIM adjustment: ○ No ● Yes

CNV detection:
Significance level: 0.01
Minimum num. of markers: 2
Number of permutations: 10000
Proportion of data to be trimmed: 0.025

AI/LOH/CNV calculation:
Confidence level: 0.95
Window size: 50
N of consecutive significant markers: 3
Upper bound of reference: 0.95
Average statistic: ● Mean ○ Median
Genotype-specific reference: ○ No ● Yes

5. Output:

Numerical output:
Save raw R data (*.RData): ○ No ● Yes
Save APT output: ● No ○ Yes
Data description: ○ No ● Yes
Individual numerical output: ○ No ● Yes

Graphical output:
Indiv-sample figure: ☑ AF figure
☑ Six-panel figure
Cross-sample figure: ☑ AI figure
☑ LOH figure
☑ CNV figure
☑ P-value figure

Figure 20. The whole-genome six-panel figures of the first sample (Sample1) used in Example 2-1.

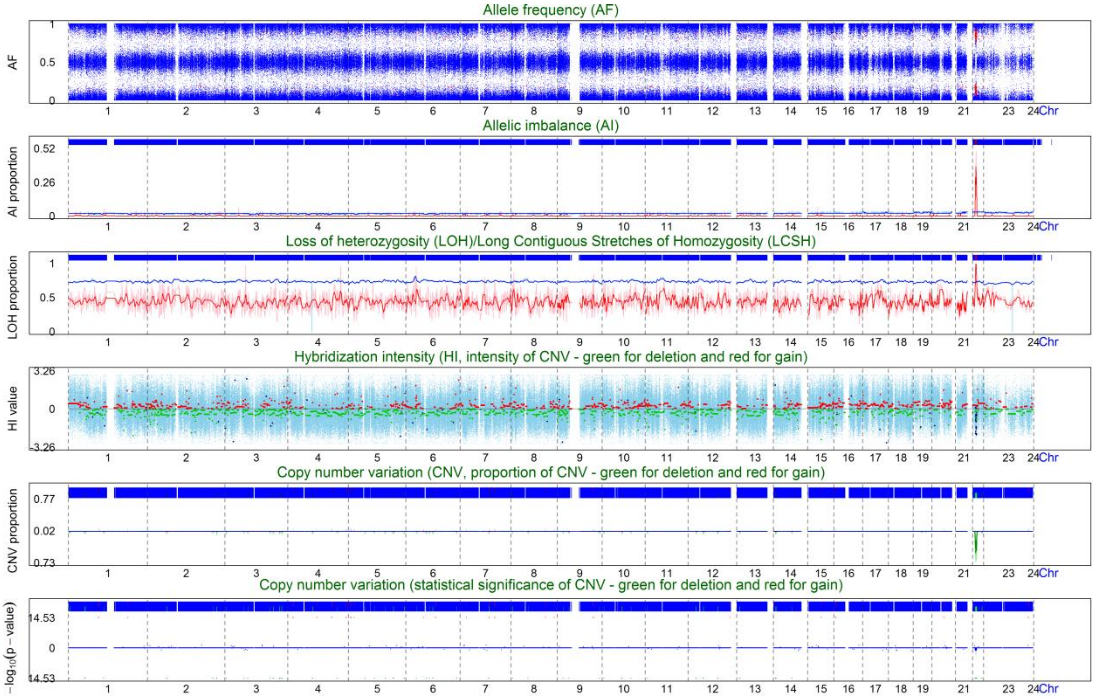
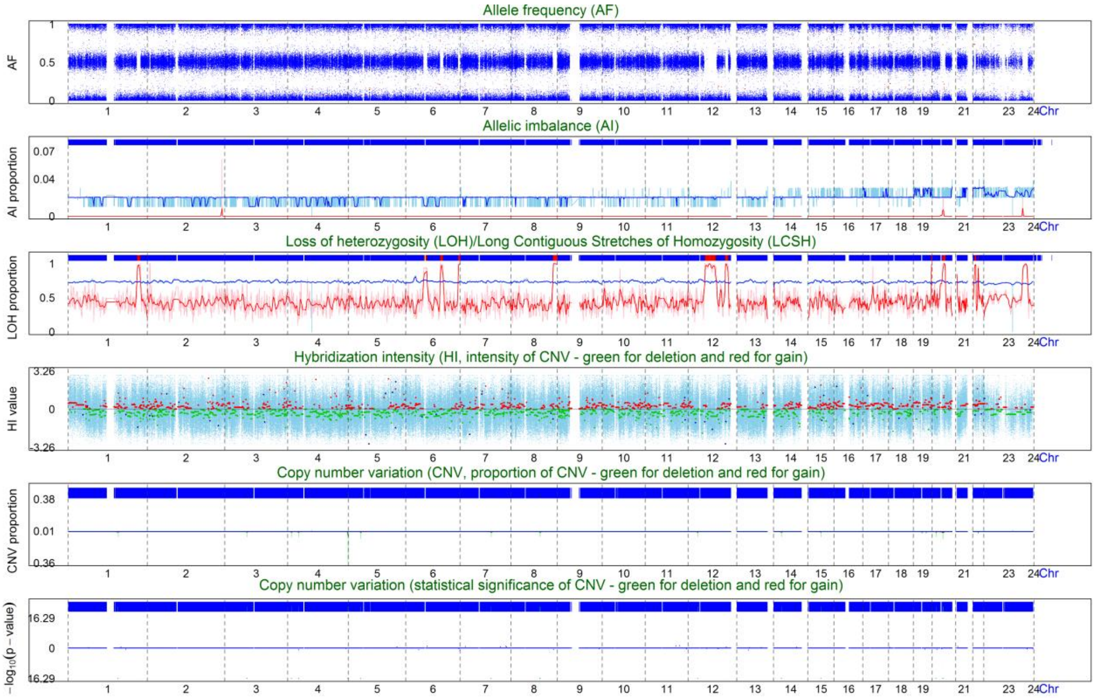


Figure 21. The whole-genome six-panel figures of the second sample (Sample2) used in Example 2-1.



6.2.2 Example 2-2 for the unpaired analysis of normal samples genotyped with Affymetrix Axiom platform using the component “Genome Browser”

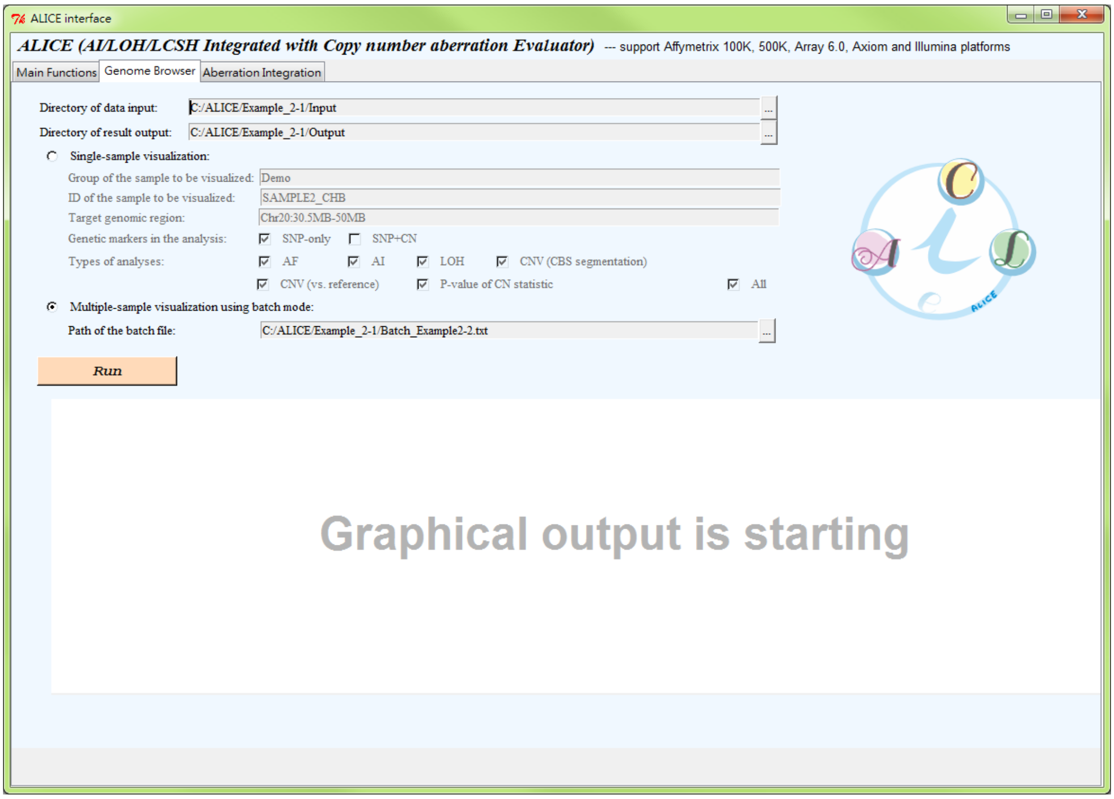
From the results of Example 2-1, we observed AI, LOH/LCSH and CNV/CNA aberrations occurred in the 23,800,000-32,000,000 bp of the 22_{th} chromosome of the first sample and 30500000-5000000 bp of the 20_{th} chromosome of the second sample. The procedures of using functions provided in the component “Genome Browser” to zoom-in the chromosomes in batch mode are listed in the following.

- Step 1 – If users already run Example 2-1, then this step can be skipped. If not, users need to download the zip file “Example_2-2.zip” from ALICE website, which is a compressed file which contains the least data required for Example 2-2. Users decompress the files into the working directory.
- Step 2 –If users already downloaded “ALICE.exe”, then this step can be skipped. Download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of ALICE will pop up. Users click the tab “Genome Browser” to switch to the interface (see Figure 2).
- Step 4 – To visualize multiple samples using batch mode, users need to prepare a batch file as mentioned in Section 3.2. If users already run Example_2-1, then

users download the batch file “Batch_Example2-2.txt” from ALICE website and save it under the folder " C:\ALICE\Example_2-1". If users run this example based on “Example_2-2.zip”, then this step can be skipped due to the batch file has already included in the compressed file.

- Step 5 – Type “Example 2” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 2-2. ALICE will automatically set the parameters required for the example (see Figure 22).
- Step 6 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_2-1\Output” if users already run analysis of Example 2-1. Otherwise, the results will be saved to “C:\ALICE\Example_2-2\Output” if users run this example based on “Example_2-2.zip”. From the graphical output, we showed the six-panel figure of the target genomic region on the 22th and 20th chromosome of the first and second sample (Sample1 and Sample 2) in Figure 23 (a) and (b), respectively.

Figure 22. Parameter setting for Example 2-2.



ALICE interface

ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Directory of data input: C:/ALICE/Example_2-1/Input

Directory of result output: C:/ALICE/Example_2-1/Output

○ Single-sample visualization:

Group of the sample to be visualized: Demo

ID of the sample to be visualized: SAMPLE2_CHB

Target genomic region: Chr20:30.5MB-50MB

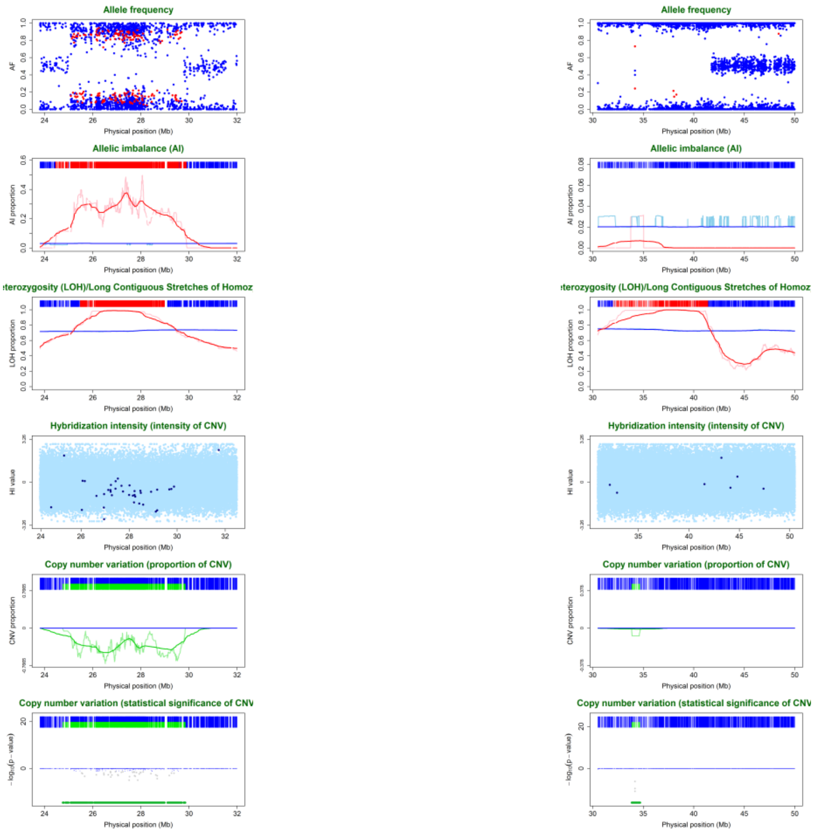
Genetic markers in the analysis: ☑ SNP-only □ SNP+CN

Types of analyses: ☑ AF ☑ AI ☑ LOH ☑ CNV (CBS segmentation) ☑ CNV (vs. reference) ☑ P-value of CN statistic ☑ All

● Multiple-sample visualization using batch mode:

Path of the batch file: C:/ALICE/Example_2-1/Batch_Example2-2.txt

Figure 23. The six-panel figures of the target genomic region of the first sample (Sample1).



(a) Sample 1.

(b) Sample 2.

6.2.3 Example 2-3 for the unpaired analysis of normal samples genotyped with Affymetrix Axiom platform using the component “Aberration Integration”

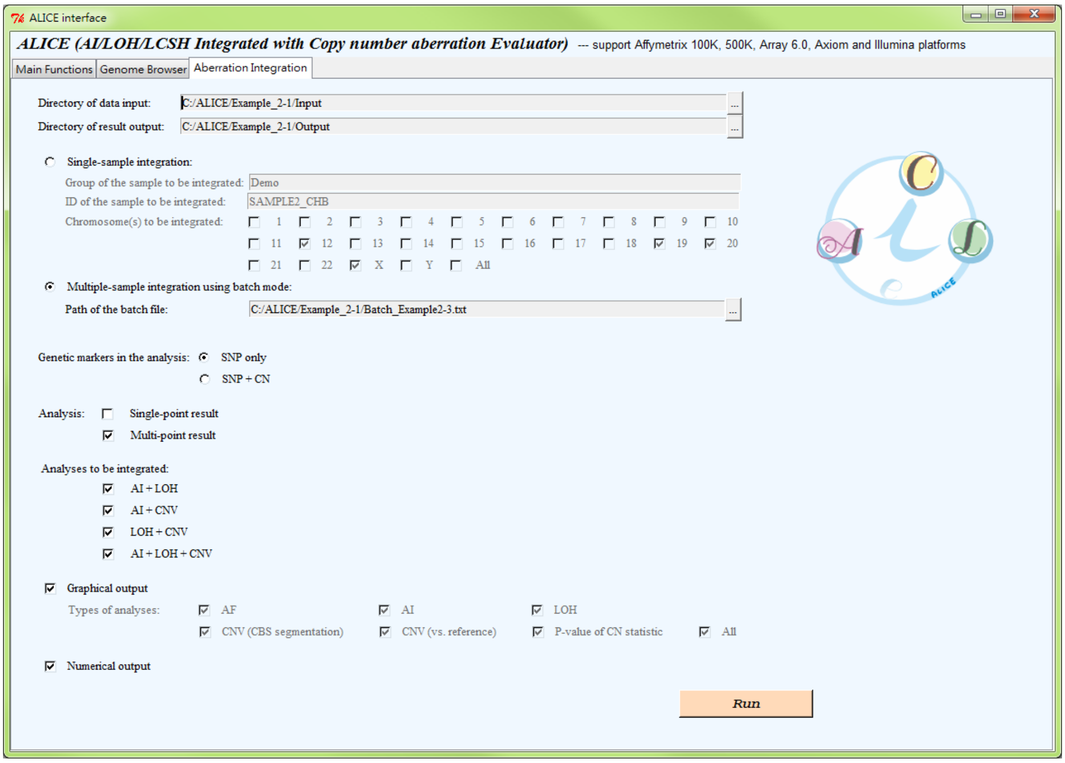
Furthermore, we integrated two or three out of AI, LOH/LCSH and CNV/CNA aberration based on the results of the two samples used in Example 2-1. For the first sample, the integration is performed on the 22th chromosome. For the second sample, the integrations are performed on the chromosomes 12, 19-20 and X. The procedures of using functions provided in the 3rd component “Aberration Integration” to achieve the goal are listed in the following.

- Step 1 – If users already run Example 2-1, then this step can be skipped. If not, users need to download the zip file “Example_2-3.zip” from ALICE website, which is a compressed file which contains the least data required for Example 2-3. Users decompress the files into the working directory.
- Step 2 – If users already downloaded “ALICE.exe”, then this step can be skipped. If not, users download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then

the interface of ALICE will pop up. Users click the tab “Aberration Integration” to switch to the interface (see Figure 3).

- Step 4 – To visualize multiple samples using batch mode, users need to prepare a batch file as mentioned in Section 3.2. If users already run Example_2-1, then users download the batch file “Batch_Example2-2.txt” from ALICE website and save it under the folder " C:\ALICE\Example_2-1". If users run this example based on “Example_2-3.zip”, then this step can be skipped due to the batch file has already included in the compressed file.
- Step 5 – Type “Example 2” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 2-3. ALICE will automatically set the parameters required for the example (see Figure 24).
- Step 6 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_2-1\Output\Demo\SAMPLE1” if users already run analysis of Example 2-1. Otherwise, the results will be saved to “C:\ALICE\Example_2-3\Output\Demo\ SAMPLE1” if users run this example based on “Example_2-3.zip”. From the graphical output, we showed the six-panel figure of the 22th and 20th chromosome of the first and second sample (Sample1 and Sample 2) in Figure 25 (a) and (b), respectively.
-

Figure 24. Parameter setting for Example 2-3.



ALICE (AI/LOH/LCSH Integrated with Copy number aberration Evaluator) --- support Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

Directory of data input: C:/ALICE/Example_2-1/Input

Directory of result output: C:/ALICE/Example_2-1/Output

○ Single-sample integration:

Group of the sample to be integrated: Demo

ID of the sample to be integrated: SAMPLE2_CHB

Chromosome(s) to be integrated: □ 1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9 □ 10 □ 11 ☑ 12 □ 13 □ 14 □ 15 □ 16 □ 17 □ 18 ☑ 19 ☑ 20 □ 21 □ 22 ☑ X □ Y □ All

● Multiple-sample integration using batch mode:

Path of the batch file: C:/ALICE/Example_2-1/Batch_Example2-3.txt

Genetic markers in the analysis: ● SNP only ○ SNP + CN

Analysis: □ Single-point result ☑ Multi-point result

Analyses to be integrated:
☑ AI + LOH
☑ AI + CNV
☑ LOH + CNV
☑ AI + LOH + CNV

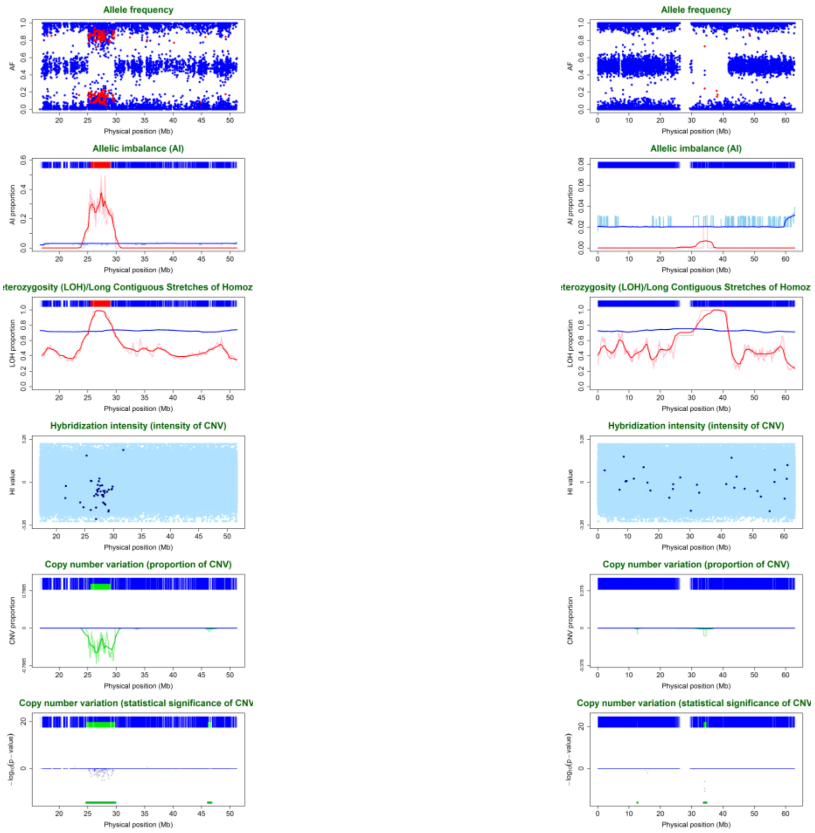
☑ Graphical output

Types of analyses: ☑ AF ☑ AI ☑ LOH ☑ CNV (CBS segmentation) ☑ CNV (vs. reference) ☑ P-value of CN statistic ☑ All

☑ Numerical output

Figure 25. The chromosomal six-panel figures of the first sample (Sample1).

(a) Sample 1, Chr 22. (b) Sample 2, Chr 20.



6.3 Example set 3

This example set illustrates the analysis procedures of RData files of three admixed samples from a cancer patient genotyped using Affymetrix Axiom platform. Example 3-1 describes the procedure of running unpaired analysis with the first component of ALICE. Example 3-2 demonstrates the procedure of running paired analysis with the first component of ALICE.

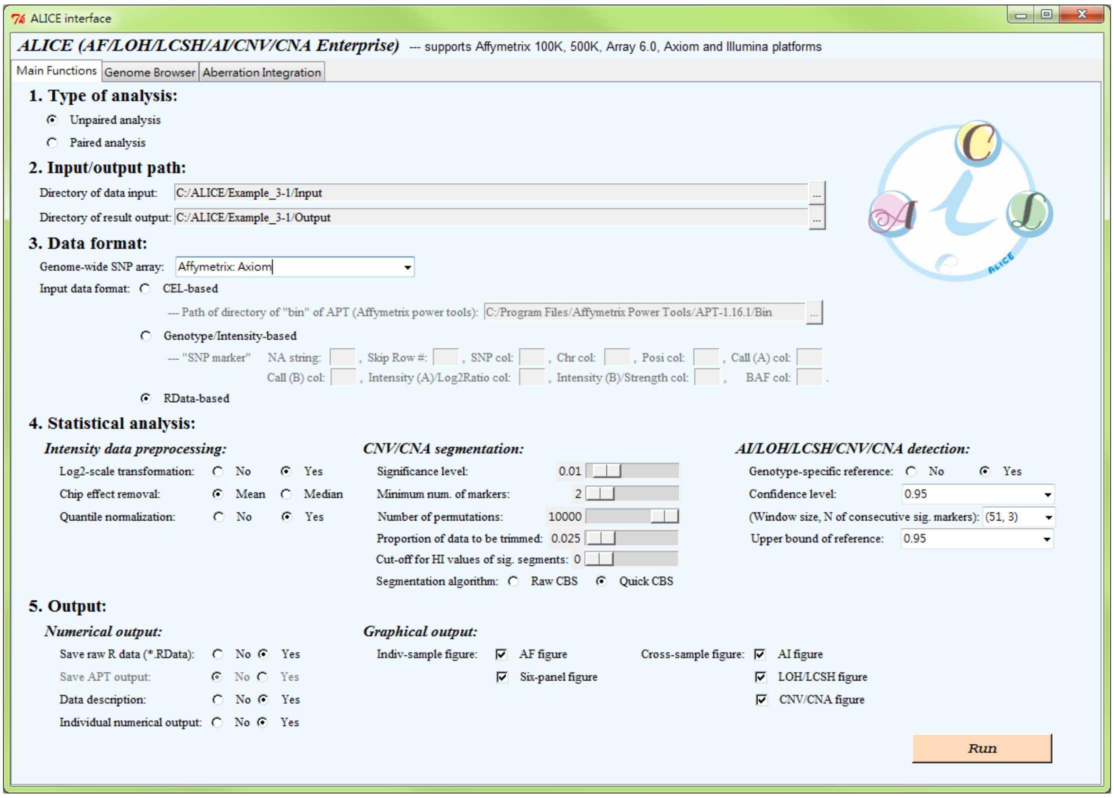
6.3.1 Example 3-1 for the unpaired analysis of cancer samples genotyped with Affymetrix Axiom platform using the component "Main Functions"

The procedures of analyzing the RData-based input data files of three admixed samples genotyped using Affymetrix Axiom platform are listed in the following.

- Step 1 – Download the zip file of the data used for Example 3-1 "Example_3-1.zip" from ALICE website and decompress the files into the working directory.
- Step 2 – Download the compressed file of ALICE v. 1.0, which contains ALICE executable file "ALICE(32-bit).exe" or "ALICE(64-bit).exe", from ALICE website and save it to the working directory.

- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of the component “Main Functions” will pop up (see Figure 1).
- Step 4 – Type “Example 3_1” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 3-1. ALICE will automatically set the parameters required for the example (see Figure 26).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the directory “C:\ALICE\Example_3-1\Output”. The whole-genome six-panel figures of the three admixed samples from the graphical output are shown in Figure 27, Figure 28 and Figure 29.

Figure 26. Parameter setting for Example 3-1.



ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

1. Type of analysis:

● Unpaired analysis

○ Paired analysis

2. Input/output path:

Directory of data input: C:/ALICE/Example_3-1/Input

Directory of result output: C:/ALICE/Example_3-1/Output

3. Data format:

Genome-wide SNP array: Affymetrix: Axiom

Input data format: ○ CEL-based

--- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin

○ Genotype/Intensity-based

--- "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) col:
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .

● RData-based

4. Statistical analysis:

Intensity data preprocessing:

Log2-scale transformation: ○ No ● Yes

Chip effect removal: ● Mean ○ Median

Quantile normalization: ○ No ● Yes

CNV/CNA segmentation:

Significance level: 0.01

Minimum num. of markers: 2

Number of permutations: 10000

Proportion of data to be trimmed: 0.025

Cut-off for HI values of sig. segments: 0

Segmentation algorithm: ○ Raw CBS ● Quick CBS

AI/LOH/LCSH/CNV/CNA detection:

Genotype-specific reference: ○ No ● Yes

Confidence level: 0.95

(Window size, N of consecutive sig. markers): (51, 3)

Upper bound of reference: 0.95

5. Output:

Numerical output:

Save raw R data (*.RData): ○ No ● Yes

Save APT output: ● No ○ Yes

Data description: ○ No ● Yes

Individual numerical output: ○ No ● Yes

Graphical output:

Indiv-sample figure: ☑ AF figure ☑ Six-panel figure

Cross-sample figure: ☑ AI figure ☑ LOH/LCSH figure ☑ CNV/CNA figure

Figure 27. The whole-genome six-panel figures of the blood cell line sample (Sample_CHB_000) from a lung cancer patient used in Example 3-1.

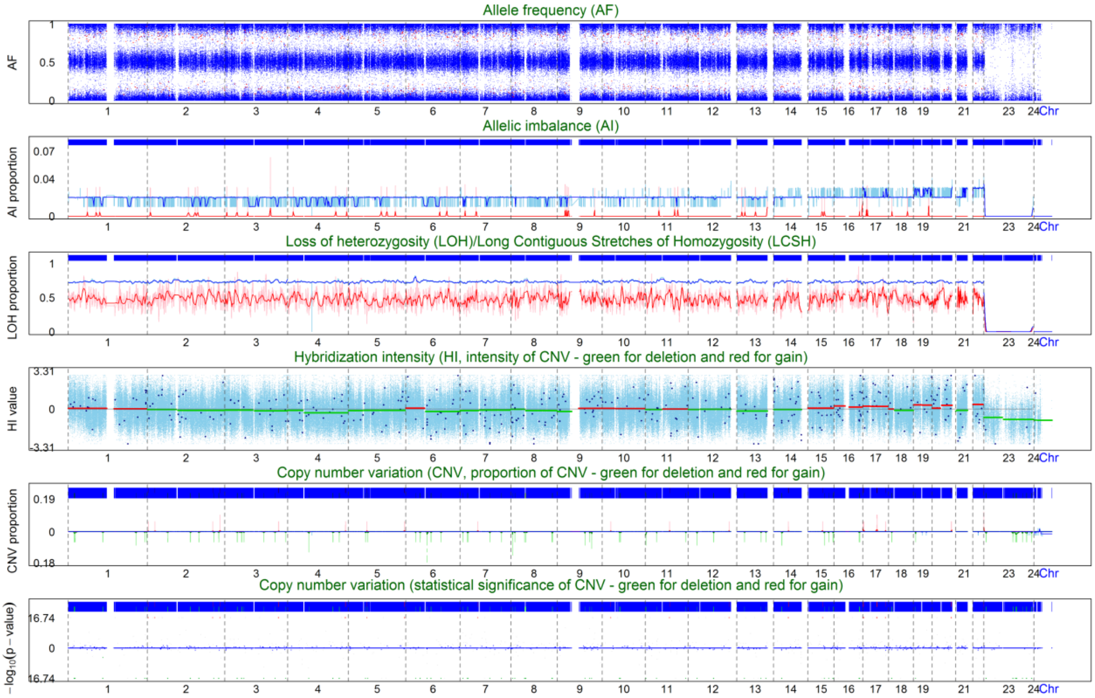


Figure 28. The whole-genome six-panel figures of the admixed sample at an admixture proportion of 50% (Sample_CHB_050) from a lung cancer patient used in Example 3-1.

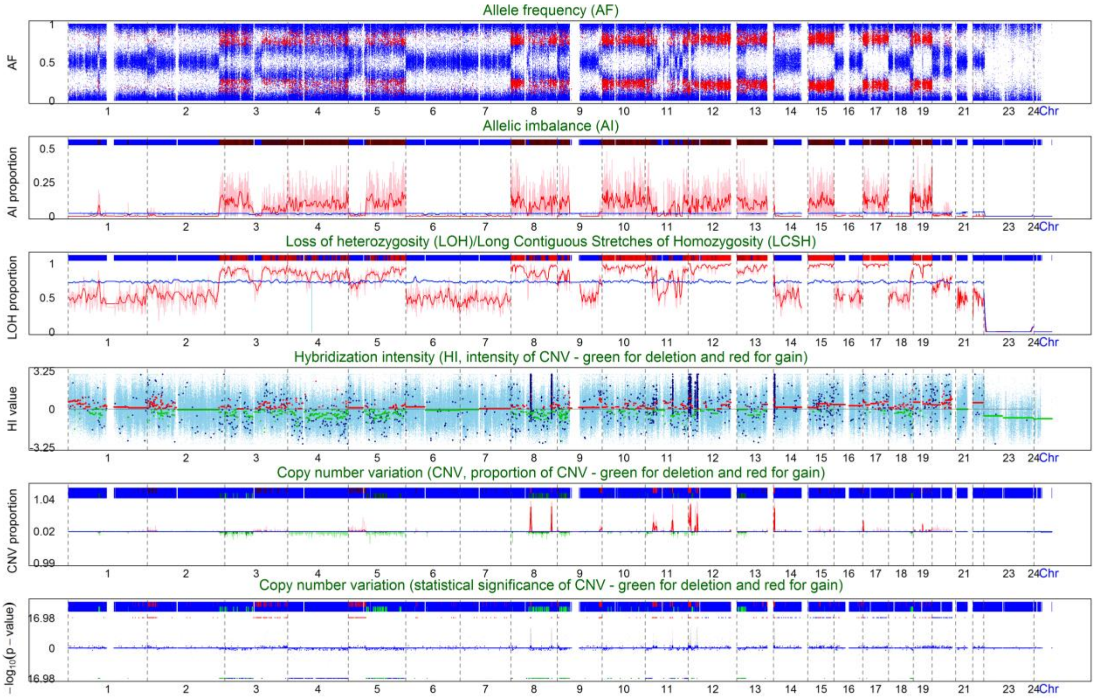
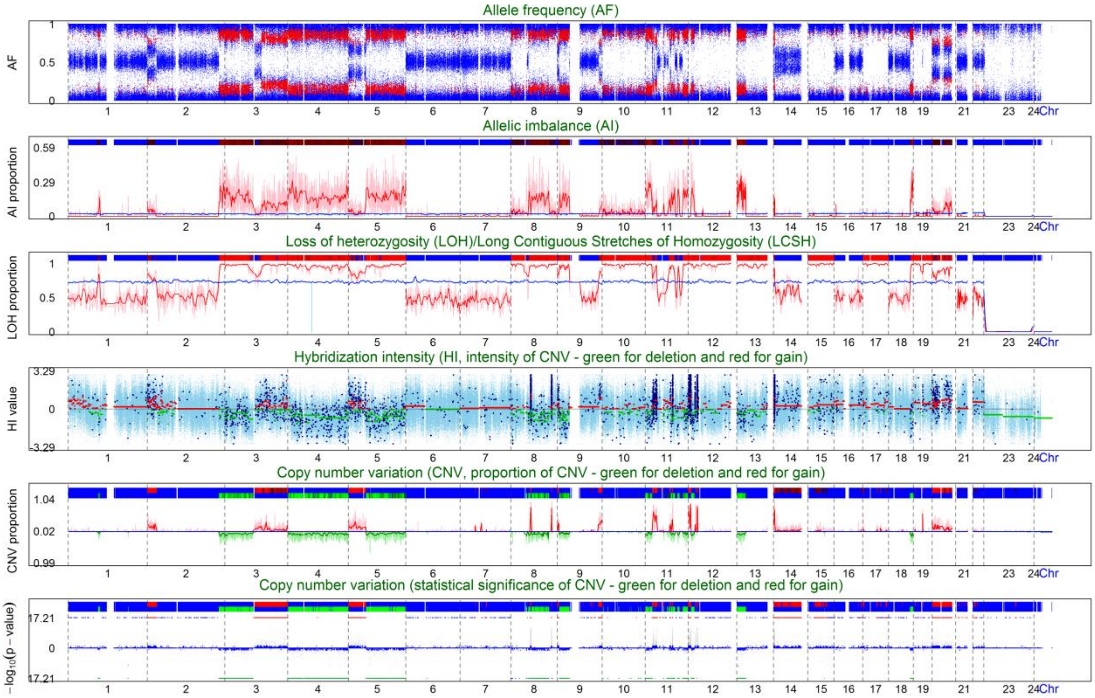


Figure 29. The whole-genome six-panel figures of the cancer cell line sample (Sample_CHB_100) from a lung cancer patient used in Example 3-1.



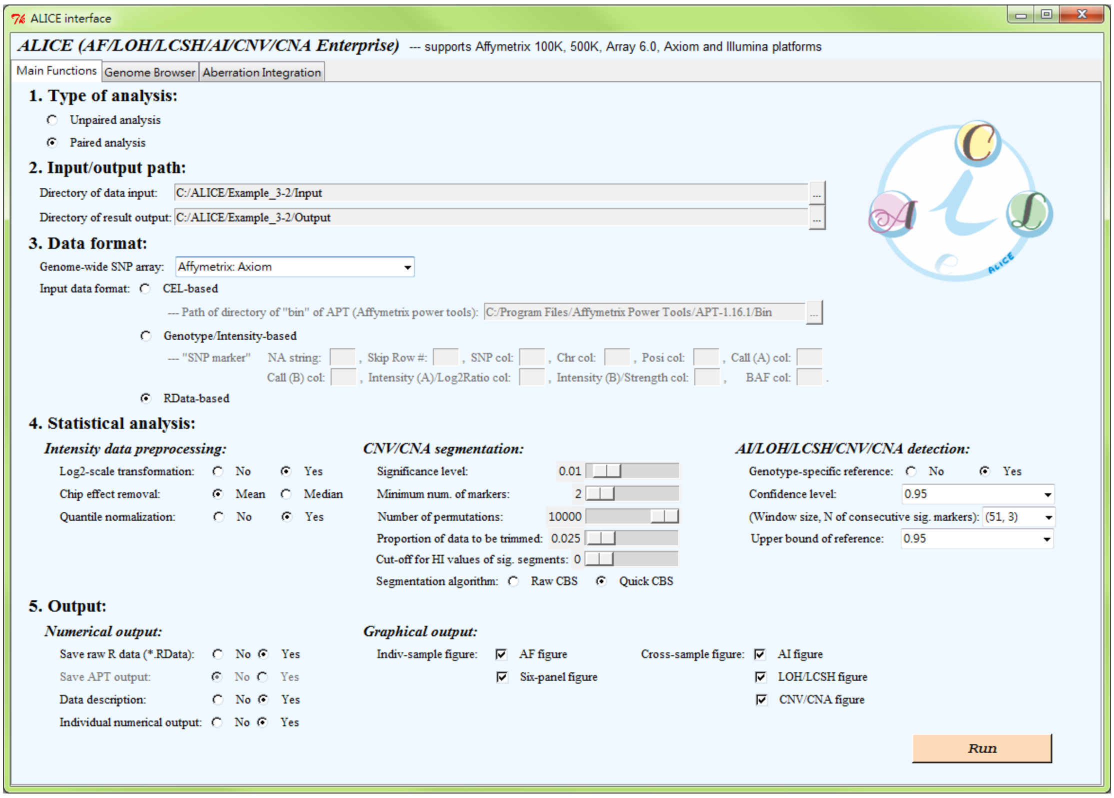
6.3.2 Example 3-2 for the paired analysis of cancer samples genotyped with Affymetrix Axiom platform using the component “Main Functions”

The procedures of analyzing the RData-based input data files of three admixed samples genotyped using Affymetrix Axiom platform are listed in the following.

- Step 1 – Download the zip file of the data used for Example 3-2 “Example_3-2.zip” from ALICE website and decompress the files into the working directory.
- Step 2 – Download the compressed file of ALICE v. 1.0, which contains ALICE executable file “ALICE(32-bit).exe” or “ALICE(64-bit).exe”, from ALICE website and save it to the working directory.
- Step 3 – Right-click on the ALICE executable file to run it as administrator, then the interface of the component “Main Functions” will pop up (see Figure 1).
- Step 4 – Type “Example 3_2” in the entry “Directory of data input” of the item “Input/output path” and click “Run” button to run analysis of Example 3-2. ALICE will automatically set the parameters required for the example (see Figure 30).
- Step 5 – The calculation progress of ALICE analysis will be updated in the text file “C:\ALICE\Logfile_D_T.txt” in a real time, where D and T represent the execution date and time, respectively. The analysis results will be saved in the

directory “C:\ALICE\Example_3-2\Output”. The whole-genome six-panel figures of the admixed sample at an admixture proportion of 50% and 100% from the graphical output are shown in Figure 31 and Figure 32.

Figure 30. Parameter setting for Example 3-2.



ALICE (AF/LOH/LCSH/AI/CNV/CNA Enterprise) --- supports Affymetrix 100K, 500K, Array 6.0, Axiom and Illumina platforms

1. Type of analysis:

○ Unpaired analysis

● Paired analysis

2. Input/output path:

Directory of data input: C:/ALICE/Example_3-2/Input

Directory of result output: C:/ALICE/Example_3-2/Output

3. Data format:

Genome-wide SNP array: Affymetrix: Axiom

Input data format: ○ CEL-based

--- Path of directory of "bin" of APT (Affymetrix power tools): C:/Program Files/Affymetrix Power Tools/APT-1.16.1/Bin

○ Genotype/Intensity-based

--- "SNP marker" NA string: , Skip Row #: , SNP col: , Chr col: , Posi col: , Call (A) col:
Call (B) col: , Intensity (A)/Log2Ratio col: , Intensity (B)/Strength col: , BAF col: .

● RData-based

4. Statistical analysis:

Intensity data preprocessing:

Log2-scale transformation: ○ No ● Yes

Chip effect removal: ● Mean ○ Median

Quantile normalization: ○ No ● Yes

CNV/CNA segmentation:

Significance level: 0.01

Minimum num. of markers: 2

Number of permutations: 10000

Proportion of data to be trimmed: 0.025

Cut-off for HI values of sig. segments: 0

Segmentation algorithm: ○ Raw CBS ● Quick CBS

AI/LOH/LCSH/CNV/CNA detection:

Genotype-specific reference: ○ No ● Yes

Confidence level: 0.95

(Window size, N of consecutive sig. markers): (51, 3)

Upper bound of reference: 0.95

5. Output:

Numerical output:

Save raw R data (*.RData): ○ No ● Yes

Save APT output: ● No ○ Yes

Data description: ○ No ● Yes

Individual numerical output: ○ No ● Yes

Graphical output:

Indiv-sample figure: ☑ AF figure

☑ Six-panel figure

Cross-sample figure: ☑ AI figure

☑ LOH/LCSH figure

☑ CNV/CNA figure

Figure 31. The whole-genome six-panel figures of the admixed sample at an admixture proportion of 50% (Sample_CHB_050) from a lung cancer patient used in Example 3-2.

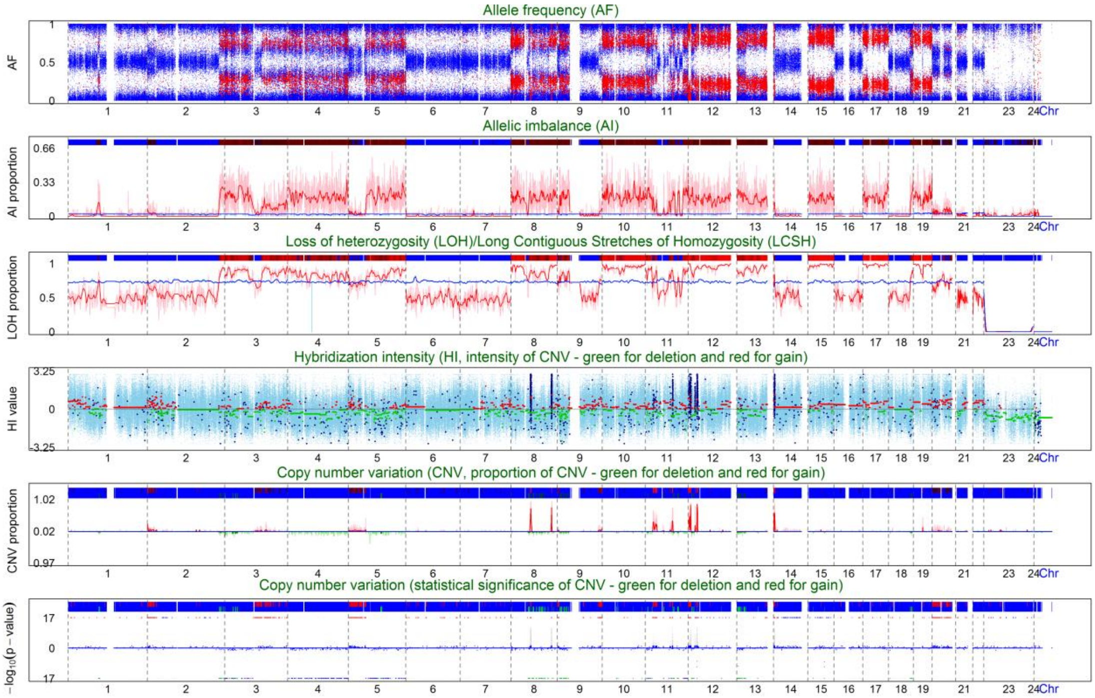
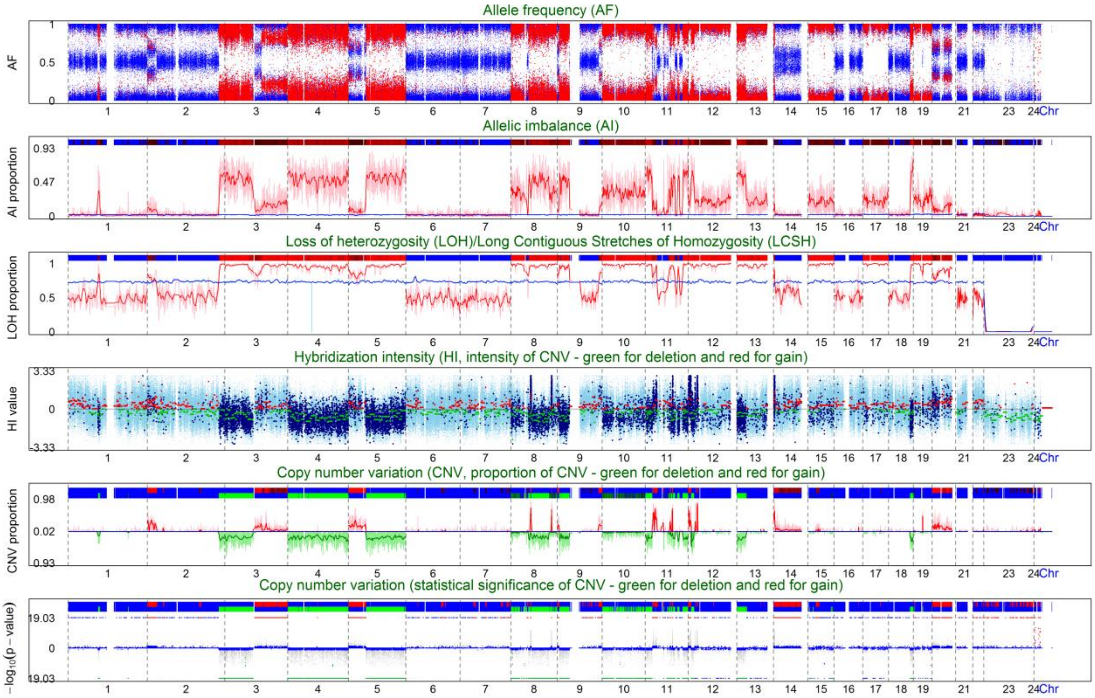


Figure 32. The whole-genome six-panel figures of the cancer cell line sample (Sample_CHB_100) from a lung cancer patient used in Example 3-2.

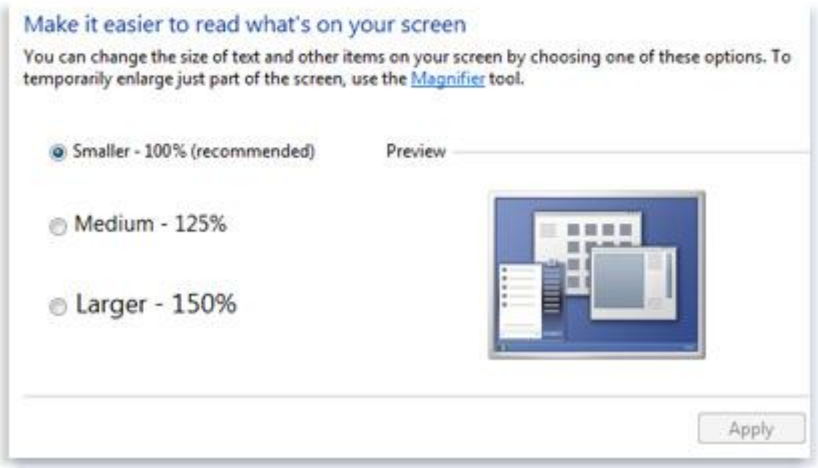


7. Frequently Asked Questions

• Why cannot my laptop/PC display the interface of ALICE software normally?

Windows provides a customized function to change the sizes of text and items on your screen without changing your screen resolution. However, if the sizes of the text and items are not set to be 100%, it will influence the display of the interface of ALICE software. Therefore, please follow the instructions below to change the setting in your laptop/PC using Windows 7.

1. Click the **Start** button, choose **Control Panel**, and click **Appearance and Personalization**.
2. Click **Adjust screen resolution**, choose the option: “**Smaller - 100% (default)**” as shown in the figure below, and click **Apply**.
3. Log off of your Windows account and back.
4. Initiate ALICE software by clicking the icon.



Make it easier to read what's on your screen

You can change the size of text and other items on your screen by choosing one of these options. To temporarily enlarge just part of the screen, use the Magnifier tool.

● Smaller - 100% (recommended)

○ Medium - 125%

○ Larger - 150%

Preview

Please visit the Microsoft website for changing settings in other versions of Windows: http://windows.microsoft.com/en-us/windows7/Make-the-text-on-your-screen-larger-or-smaller?comply