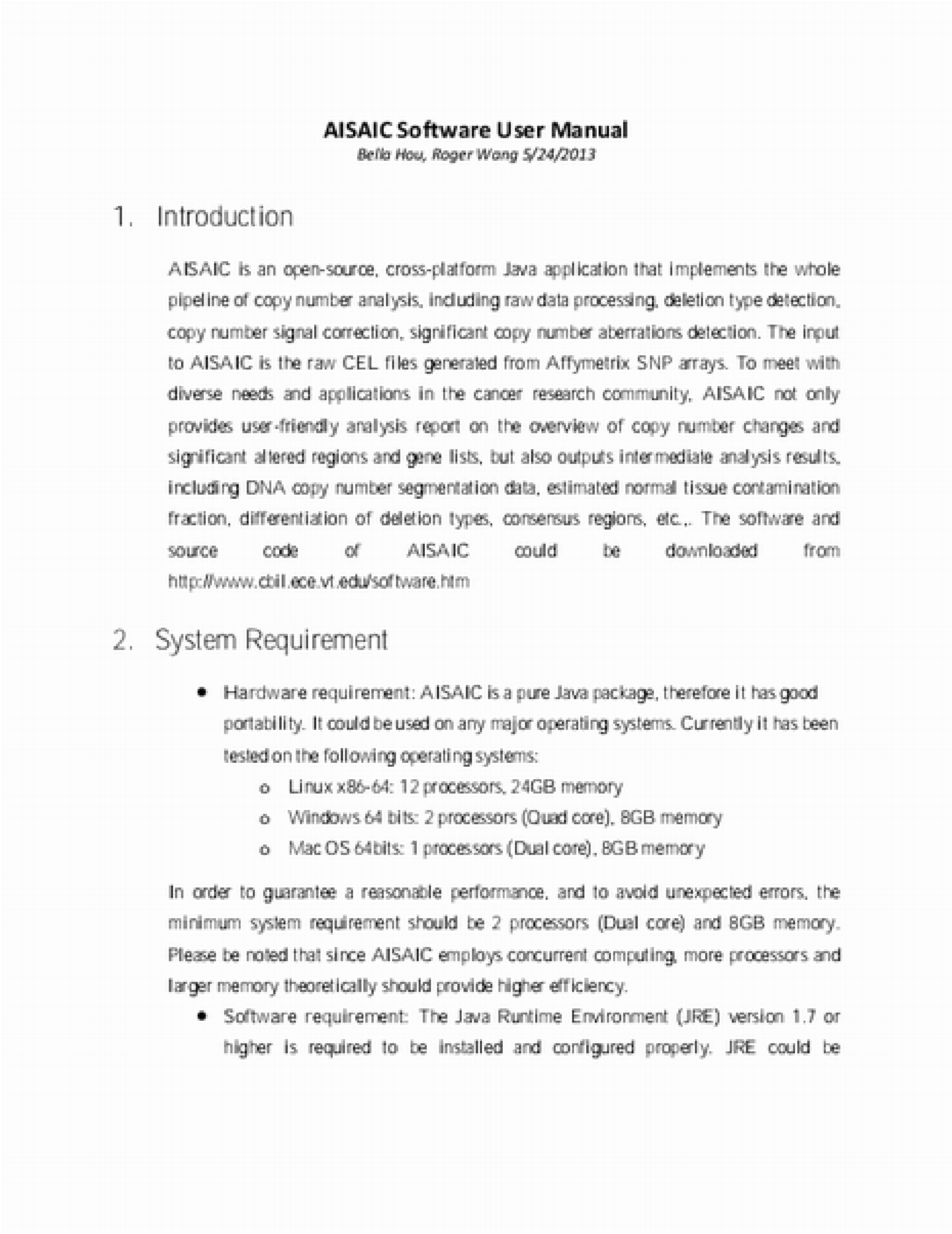

AISAIC Software User Manual

Bella Hou, Roger Wang 5/24/2013

1. Introduction

AISAIC is an open-source, cross-platform Java application that implements the whole pipeline of copy number analysis, including raw data processing, deletion type detection, copy number signal correction, significant copy number aberrations detection. The input to AISAIC is the raw CEL files generated from Affymetrix SNP arrays. To meet with diverse needs and applications in the cancer research community, AISAIC not only provides user-friendly analysis report on the overview of copy number changes and significant altered regions and gene lists, but also outputs intermediate analysis results, including DNA copy number segmentation data, estimated normal tissue contamination fraction, differentiation of deletion types, consensus regions, etc.,. The software and source code of AISAIC could be downloaded from http://www.cbil.ece.vt.edu/software.htm

2. System Requirement

- Hardware requirement: AISAIC is a pure Java package, therefore it has good portability. It could be used on any major operating systems. Currently it has been tested on the following operating systems:
  - Linux x86-64: 12 processors, 24GB memory
  - Windows 64 bits: 2 processors (Quad core), 8GB memory
  - Mac OS 64bits: 1 processors (Dual core), 8GB memory

In order to guarantee a reasonable performance, and to avoid unexpected errors, the minimum system requirement should be 2 processors (Dual core) and 8GB memory. Please be noted that since AISAIC employs concurrent computing, more processors and larger memory theoretically should provide higher efficiency.

- Software requirement: The Java Runtime Environment (JRE) version 1.7 or higher is required to be installed and configured properly. JRE could be

1. Introduction

AISAIC is an open-source, cross-platform Java application that implements the whole pipeline of copy number analysis, including raw data processing, deletion type detection,
copy number signal correction, significant copy number aberrations detection. The input to AISAIC is the raw CEL files generated from Affymetrix SNP arrays. To meet with diverse needs and applications in the cancer research community, AISAIC not only provides user-friendly analysis report on the overview of copy number changes and significant altered regions and gene lists, but also outputs intermediate analysis results, including DNA copy number segmentation data, estimated normal tissue contamination fraction, differentiation of deletion types, consensus regions, etc.,. The software and source code of AISAIC could be downloaded from
http://www.cbil.ece.vt.edu/software.htm

2. System Requirement

- Hardware requirement: AISAIC is a pure Java package, therefore it has good portability. It could be used on any major operating systems. Currently it has been tested on the following operating systems:
  o Linux x86-64: 12 processors, 24GB memory
  o Windows 64 bits: 2 processors (Quad core), 8GB memory
  o Mac OS 64bits: 1 processors (Dual core), 8GB memory
  In order to guarantee a reasonable performance, and to avoid unexpected errors, the minimum system requirement should be 2 processors (Dual core) and 8GB memory. Please be noted that since AISAIC employs concurrent computing, more processors and
  larger memory theoretically should provide higher efficiency.
- Software requirement: The Java Runtime Environment (JRE) version 1.7 or higher is required to be installed and configured properly. JRE could be

downloaded at http://www.oracle.com/technetwork/java/javase/downloads/jre7-downloads-1880261.html

3. Installation

Step 1: Download a Java IDE such as Netbeans or Eclipse and an up-to-date Java Development Kit (JDK). This manual uses the Netbeans IDE and Java SE 7 or above.

Netbeans: https://netbeans.org/downloads/index.html (Download any of the options, although only the simplest version is necessary)

Java SE 7: http://www.oracle.com/technetwork/java/javase/downloads/index.html (Download the Java Platform (JDK) 7u21 or whatever is the most up-to-date version. It is the first download link.)

Step 2: Install both software using the installation wizards. It is recommended to install the JDK first because Netbeans will then automatically use the JDK as the base platform; otherwise, you will have to browse for the folder the JDK was installed to so Netbeans will know which Java Platform to use.

Step 3: If not already done, obtain the software package (AISAIC) and related data and test files. Unzip the package to any folder you like.

Step 4: Open up the Netbeans application. Once it has finished loading, click 'File' and 'Open Project'. Find the folder you unzipped the software package to and expand it. You will notice that projects are illustrated as a cup of coffee in Netbeans. The project should be called "AISAIC" and have a coffee cup image next to it. Open it by double-clicking or single-clicking and then clicking 'Open Project'.

Step 5: In the projects pane on the (default) left side of the Netbeans window, you will now see the 'AISAIC' project. (**if there are warning symbols on some of the Source Packages, follow the subsequent procedure). Expand the project by clicking the '+' button. Now, expand the 'Libraries' folder. You should see JDK 1.7 (Default) or something similar depending on the JDK version you downloaded earlier. Netbeans may have automatically added the necessary libraries from the software package; it will be named the full pathname (e.g. home/ Desktop/AISAIC/libraries). If you do not see this folder, right- click the 'Libraries' folder and click 'Add JAR/Folder...'. Then navigate to where you unzipped the project. You will see a folder called 'libraries' that contains 'affymetrix' and 'umontreal' folders; add the 'libraries' folder to the project libraries.

Step 6: If not done so already, expand the 'Source Packages' folder in the 'Projects' pane. If you see warning symbols, go back to Step 5 and make sure you added the correct libraries. Now, you can run the Graphical User Interface (GUI) two ways. For the first method, ignore the other packages and expand the 'gui' package. You will see a .java file called 'AISAICGUI.java'. Open the file by double-clicking or right-click+open. This is called the "Main Class" of the software package. When it is open, simply find the green arrow at the top of the Netbeans window and click it to Run. Alternatively, you can simply press F6 to run the entire project. The GUI should automatically pop-up on your screen.

Figure 1. AISAIC graphical user interface

4. Usage

Step 1. Pass data and parameter.

Users can pass their own data and parameters through the provided graphical user interface, shown below in Figure 1.

The GUI contains four modules.:Analysis Selection, BACOM Module, SAIC Module

and Processing Status.

The first module is "Analysis Selection", shown in Figure 2, where user can choose the analysis they need.

Figure 2. Analysis Selection Module

If "BACOM ONLY" is chosen, all the following operations are constrained in the second module "BACOM Module", shown in Figure 3.

Figure 3. BACOM Module

The input of BACOM analysis is raw copy number data file (.cel) from Affymetrix SNP 6.0.

In this BACOM Module, we allow both "Single sample analysis" and "Multiple samples analysis". To use "Single sample analysis", leave the "Multiple Sample Processing" button unchecked; conversely, if you want to use "Multiple samples analysis", check the button.

• If "Single sample analysis" is chosen, the data and parameters could be passed through following steps:

1. "Input normal sample": a normal raw copy number data file (.CEL file) needs to be selected by clicking button .

2. "Paired tumor sample": the paired tumor raw copy number data file (.CEL file) needs to be selected by clicking button .

3. “CDF file”: the CDF file associated with the specific copy number chip that the data is obtained needs to be selected by clicking button .

4. “BACOM result folder”: the directory under which the results of BACOM will be stored needs to be specified by clicking button .

(Note: the software will create a folder named “results” under the specified directory. There are 3 result files expected, please refer to the “Analysis Result” in this user manual.

5. “Genomic region interested”: if this field is specified (follow the format example in the GUI), the software will output the deletion type (homo-deletion, hemi-deletion or no deletion) of this region in the “Processing Status” on GUI.

- If “Multiple samples analysis” is chosen, the data and parameters could be passed through following steps:

1. “Input Normal sample”: a text file (e.g., NormalSample.txt) that contains the full pathnames of the normal samples on separate lines (Figure 3.) needs to be specified by clicking button .

(Note: the normal samples should be placed in the same folder as the NormalSample.txt file)

Figure 4: Example Text File Setup

2. “Matched Tumor sample”: another text file (e.g., TumorSample.txt) that contains the matched tumor samples with same format as the normal samples text file needs to be specified by clicking button . Here “match” means the normal sample and tumor sample from the same subject should in the same line in their respective text file.

(Note: the tumor samples should be placed in the same folder as the TumorSample.txt file, and tumor samples could be placed under the same folder as normal samples)

3. “CDF file”: a CDF file (.cdf) of the Affymetrix SNP 6.0 chip needs to be selected by clicking button .

4. “BACOM result folder”: the directory under which the results of BACOM will be stored needs to be specified by clicking button .

5. “Genomic region interested”: if this field is specified (follow the example in the GUI), the software will print out the proportion of homo-deletion and hemi-deletion among all the samples in “Processing Status” on GUI.

If “SAIC ONLY” is chosen, all the following operations are constrained in the third module “SAIC Module”, shown in Figure 5.

Figure 5. SAIC module

For SAIC analysis, we provide users with three analysis modes: “Single chromosome **analysis”, “Multiple chromosomes analysis” and “Genome-wide analysis”. Click “Batch** Processing” for Multiple chromosomes analysis, otherwise it is Single chromosome analysis.

- If “Single chromosome analysis” mode is chosen:

1. “Input segmented CNA data”: the segmented copy number data needs to be specified by clicking .

(Note: the data should be a N*M matrix with N loci and M samples,and each entry is the copy number value after normalization and segmentation)

2. “SAIC detection results folder”: the directory under which the SAIC results will be stored needs to be specified by clicking button . Please refer to the “Analysis Results” section in this manual for detailed description of the results.

- If “Multiple chromosomes analysis” is chosen:

1. “Input segmented CNA data”: the directory where the segmented copy number data matrix of 22 chromosomes stored needs to be specified by clicking . Please be noticed that the file name for the 22 chromosomes should follow the rules: Chr1, Chr2, Chr3... Chr22.

3. “SAIC detection results folder”: the directory under which the SAIC results will be stored needs to be specified by clicking button . Please refer to the “Analysis Results” section in this manual for detailed description of the results.

- If “Genome-wide analysis” is chosen:

1. “Input segmented CNA data”: a segmented genome-wide copy number data matrix file needs to be selected by clicking button .

2. “SAIC detection results folder”: the directory under which the SAIC results will be stored needs to be specified by clicking button . Please refer to the “Analysis Results” section in this manual for detailed description of the results

- “Log-ratio amplification threshold”:the threshold for detecting copy number amplification needs to be set from this field.
- “Log-ratio deletion threshold”: the threshold for detecting copy number deletion needs to be set from this field.
- “Number of permutation”: the number of permutation in the null hypothesis estimation needs to be set from this field.
- “Correlation coefficient threshold”: the threshold for CNA unit construction needs to be set from this field.
- “Adopt SCA-excluding permutation scheme”: the default setting is chosen since this step is one of the character of SAIC algorithm. In case some users do not prefer this step they can unclick it.

- “Adopt quick SAIC”: if chosen, a down-sampling scheme is utilized to shrink the size of the input data matrix, and improve the efficiency of the algorithm. Extensive experiments suggest that this down-sampling can significantly speed up the algorithm with little accuracy lost.

If “BACOM+SAIC” is chosen, i.e., the whole pipeline analysis of AISAIC, the raw copy number data need to be passed from “BACOM Module” as introduced above. And the parameters for SAIC analysis also need to be specified in “SAIC Module”. The software package will take care of data format conversion from BACOM results to the input format required by SAIC. And please be note that if “BACOM+SAIC” is chosen, the default “Analysis mode” of BACOM is “Multiple samples analysis” and the default “Analysis mode” of SAIC is “Multiple chromosomes analysis”. If users prefer “Genome-wide analysis”, they can still choose it.

When all the data and parameters are ready, users can start the analysis by clicking “Start analyzing”, the processing status and some results will be shown in the last module, “Processing Status”, shown below in Figure 6.

Figure 6. Processing Status Module 5. Analysis Results

Analysis Results will depend on which analysis is chosen.

- If “BACOM Only” is chosen, only the output of BACOM analysis will be stored in a folder named “results” under the directory indicated by the user. For each sample, there should be 3 results files expected: “*_outputBACOM_locations”, which contains the deletion segments and the probability of homo-deletion and hemi-deletion; “*_outputBACOMresults” which contains the estimated normal tissue contamination fraction in this sample; And “*_outputdata”, which contains the normalized copy number value, segmented copy number value and copy number value after normal tissue contamination correction for each SNP loci.

- **If “SAIC Only” is chosen, for each chromosome, or whole genome, there are 4** files with file names “Result_ampSCA”, “Result_delSCA”, “Result_ampGene” and “Result_delGene” should be expected. Therefore if “Multiple chromosomes analysis” is chosen, there should be 4*22 results files expected. “Result_ampSCA”, “Result_delSCA” contain the detected deleted SCAs and amplified SCAs respectively. Each row contains the information of a detected SCA, including its genomic location and corresponding U-score, P-value. “Result_ampGene” and “Result_delGene” contain the genes covered by the amplified SCAs and deleted SCAs.
- If “BACOM+SAIC” is chosen, output of both BACOM and SAIC will be stored in the same folder named “results”. In addition, the intermediate results that converted from BACOM output to the input for SAIC, i.e., the segmented copy number data matrix of all the 22 chromosomes, named as “Chr1”, “Chr2”... “Chr22”, and one genome-wide copy number data matrix, named as “GenomeWide” will also be stored in the same folder. Another important output of “BACOM+SAIC” is the proportion of homo-deletion and hemi-deletion of the SCAs among all the analyzed samples, this result could be found in “*.delSCA” files.