BACOM User's Manual

1 Introduction

BACOM is a cross-platform and open source Java application. Bayesian Analysis of COpy number Mixtures (BACOM) is a statistically-principled *in silico* approach to accurately estimate genomic deletions and normal tissue contamination, and accordingly recover the true copy number profile in cancer cells. The BACOM software instantiates not only the algorithm of Bayesian analysis of copy number mixtures but also other relevant processing steps, including extraction of raw copy number signals from CEL files, iterative data normalization, identification of AB loci, copy number detection and segmentation, probe sets annotation, differentiation of deletion types, estimation of normal tissue fraction, and correction of normal tissue contamination. Interested readers can download freely the software and source code at `http://www.cbil.ece.vt.edu/software.htm`.

2 System Requirement

The BACOM has been tested on the following operating systems:

- Linux x86-64, 6GB or more memory.
- Windows 64 bit, 6GB or more memory.

Note that it is important to meet the memory requirement and use 64 bit operating systems.

The Java Runtime Environment (JRE) version 1.6 or higher is required to be installed and configured properly. Sun JRE can be downloaded at `http://www.java.com`.

3 Usage

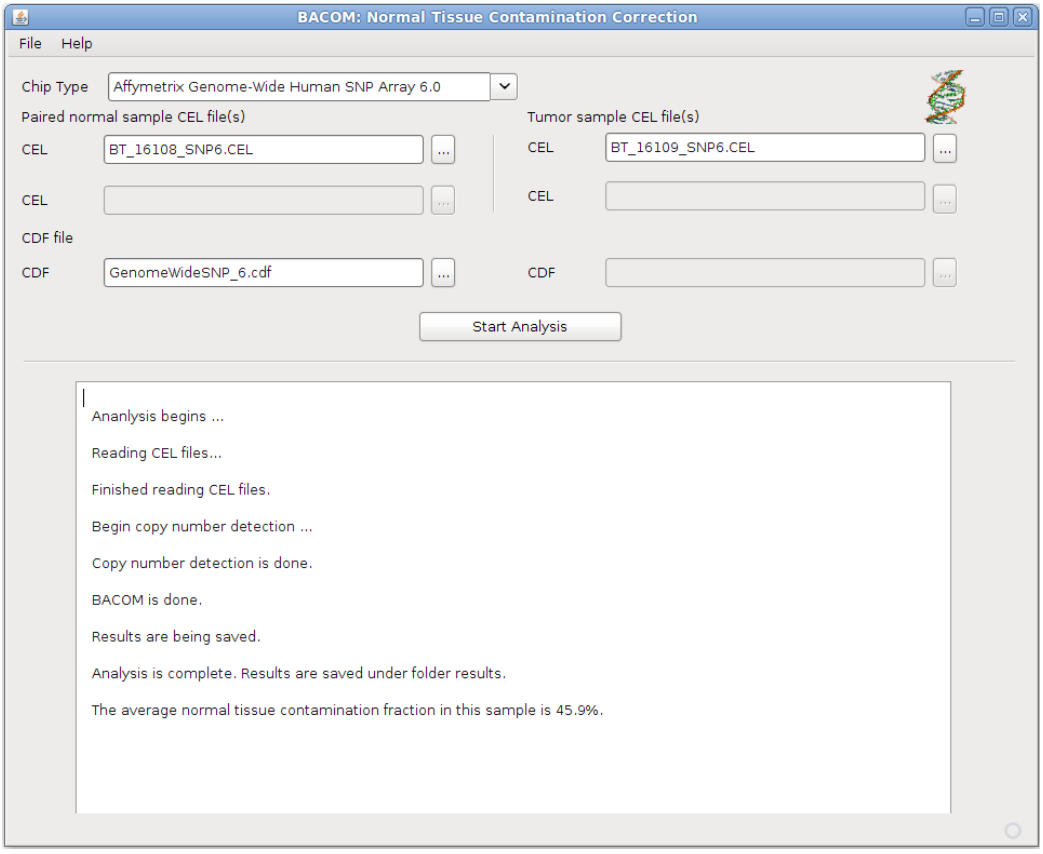

Figure 1: A screen shot of the BACOM software (Affymetrix Genome-Wide Human SNP Array 6.0).

To run BACOM, under Windows, double-click `BACOM.bat`; under Linux, double-click BACOM.sh. After the program is launched, a window similar to Figure 1 will show up.

Figure 1 is a screen shot of the BACOM software when analyzing samples assayed using Affymetrix Genome-Wide Human SNP Array 6.0. First, the "Chip Type" is selected in the top drop-down menu, which in this case is Affymetrix Genome-Wide Human SNP Array 6.0. The CEL file on the left panel is the CEL file for the paired normal sample for the analysis and the CEL file on the right panel is the CEL file for the tumor sample. Then, select the CDF file for the corresponding platform. Click the button "Start Analysis" to initiate

the analysis. The processing steps will show up in the white text box at the bottom. The analysis results will be saved in the folder "results" under the BACOM installation path. It usually takes a few minutes to finish the analysis.

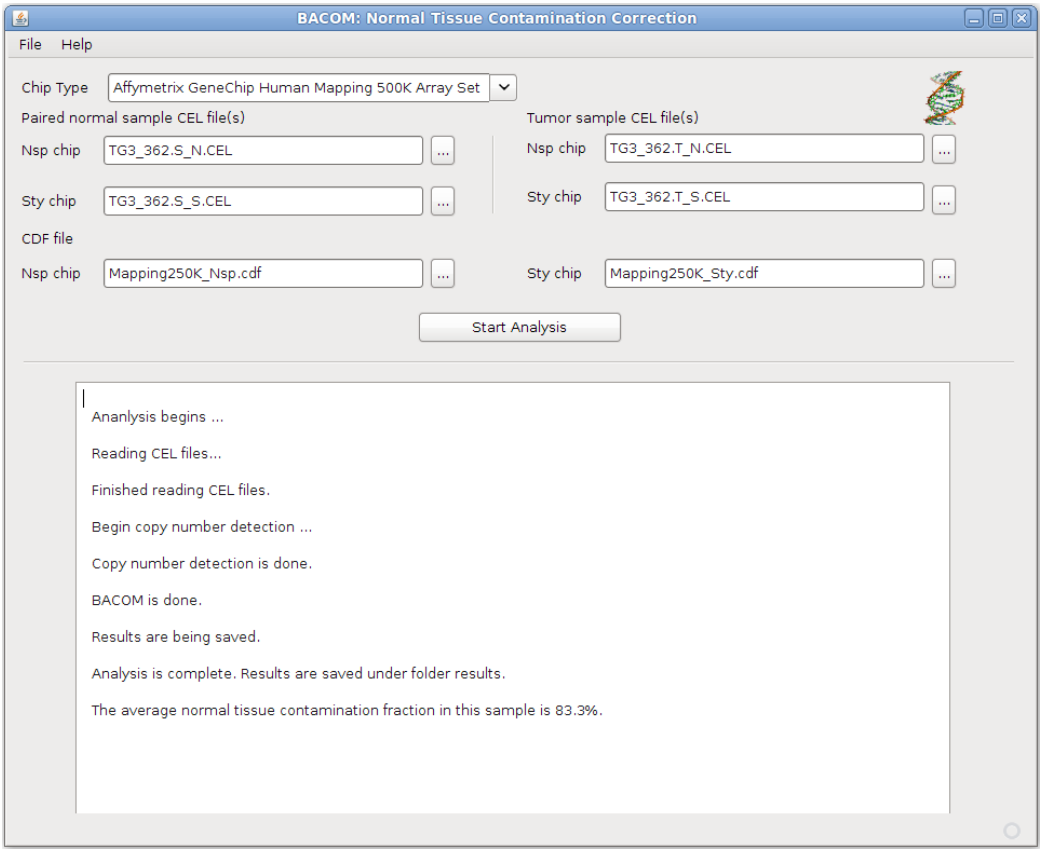

Figure 2: A screen shot of the BACOM software (Affymetrix GeneChip Human Mapping 500K Array Set).

Figure 2 is a screen shot of the BACOM software when analyzing samples assayed using Affymetrix GeneChip Human Mapping 500K Array Set. First, the "Chip Type" is selected in the top drop-down menu, which in this case is Affymetrix GeneChip Human Mapping 500K Array Set. In Affymetrix 500K Array Set, there are two chips for each sample: Nsp chip and Styp chip. Therefore, the CEL files on the left panel are the CEL files (Nsp and Sty) for the paired normal sample for the analysis and the CEL files on the right panel are

the CEL files (Nsp and Sty) for the tumor sample. Accordingly, there are two CDF files for this platform: `Mapping250K_Nsp.cdf` and `Mapping250K_Sty.cdf`. After selecting the CDF files for the platform, click the button "Start Analysis" to initiate the analysis. The processing steps will show up in the white text box at the bottom. The analysis results will be saved in the folder "results" under the BACOM installation path.

4 Analysis Results

In the folder "results", there are four files: `alpha.txt`, which contains the estimated fraction of normal tissue contamination, `BACOM_locations.txt`, which contains the Bayesian analysis results of each deletion segments and the start and end positions of the deletion segments are the physical locations on the respective chromosomes, `BACOMresults.txt`, which also contains the Bayesian analysis results of each deletion segments and the start and end positions of the deletion segments are indices, and `data.csv`. `data.csv` is a comma-separated values file, which contains five columns: the first column is the probe set IDs, the second column is the chromosome IDs, the third column is the physical locations of the markers on the respective chromosomes, the fourth column is the observed copy number signals, and the fifth column is the estimated segment means.

5 (Optional) Visualization

We also provide an R script to plot the DNA copy number alteration in the tumor sample. R is required to be installed for the visualization, which can be downloaded at `http://www.r-project.org`. After performing the analysis with BACOM, run `VisualBACOM.sh` (Linux) or `VisualBACOM-Windows.bat` (Windows). The file `VisualBACOM-Windows.bat` should be manually edited to include R's installation path in the file before its first run.