Background survey

For producing the cancer cell fraction (CCF) from somatic point mutation calls and the clusters of every mutation with assignment to the second phase, a good way is to use Ccube to analysis the inputs with Ccube format and get the output files to second phase.

The tool Ccube is an R code for clustering the cancer cell fraction. According to the paper by Dr. Ke, It is a probabilistic framework that includes a variational inference method for model fitting, which allows the user to process samples with a large number of variants while quantifying the uncertainty in a Bayesian fashion[1]. From the whole genome sequencing data, the cancer cell fraction and sub-clonal composition of somatic cell point mutations could be inferred. It requires sequencing read analysis of single nucleotide variants (SNV), correcting their copy number changes and purity, and generating cancer cell fraction estimates for all mutations in the sample.

The Ccube choose the suitable number of clusters by variational inference which can get the best evidence lower bound and assign the mutations to the clusters according to the probability. The Ccube pipeline has the switch named “runQC”. With setting it true, the Ccube can remove cluster and re-assign variants to make the result of clustering more accurate.

The summary figures can be known from below. The Ccube find the CCF and VAF has the relationship. The map between them is based on the function:

is the CCF of the mutation, is a uniform sequencing error, and m is the number of mutated chromosomal copy.

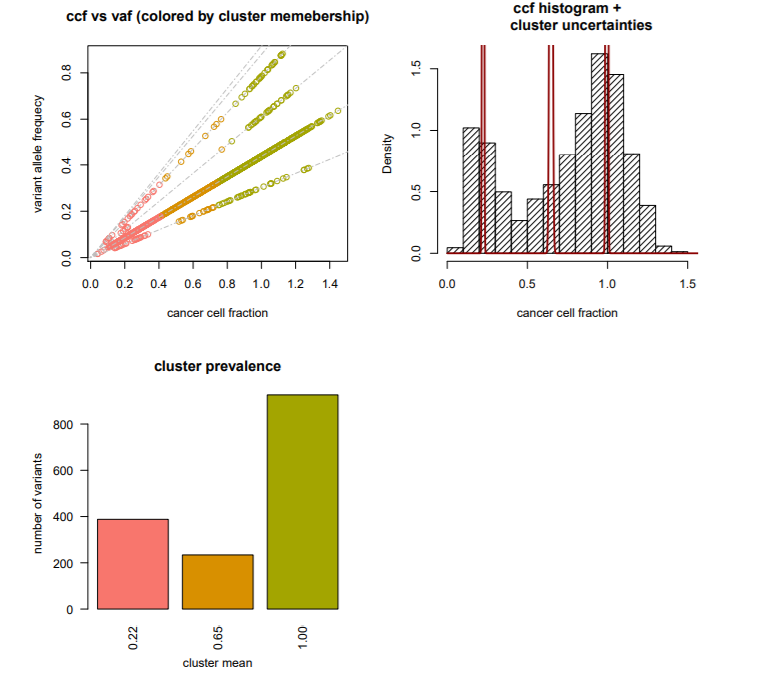


Figure1: the summary of the Ccube results, the relationship between CCF and VAF, and the situation of clusters

Requirements

## For the tool Ccube, the required opearation system is Linux. It can work in the R over version 3.4.0.The core package is “Ccube” with version 0.0.0.9000; it includes the all functions of Ccube to analyze the mutations and get CCF. And it’s also need package “dplyr” to add new variables that are functions of existing variables and the package “doParallel” to run the function with multi-cores.

And before the input entering the Ccube tool, the data should be converted to the format which is suitable for Ccube. The code of this process is based on the python2. It needs the package “os” to read the path from command shell, and the package “numpy” to stack the arrays.

In summary:

* OS: Linux
* Coding language: Python, R
* Main library used: Ccube, doParallel, dplyr
* Other libraries: Numpy, os

## Design and Implementation

## For designing the Ccube pipeline in this software, the designing and working flowchart is like below, the pipeline is used to convert the data by users and get the output files from Ccube R code.

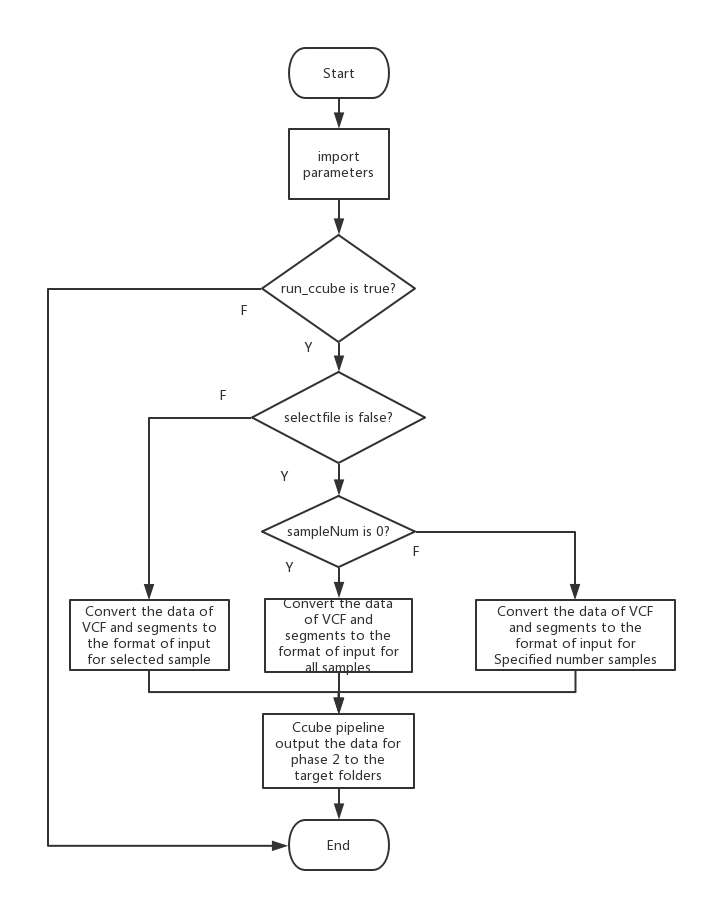


Figure 2, the integrate Flowchart of Ccube Pipeline

The pipeline of Ccube can be divided to three parts. The first part is to pass the parameters to the main functions which will be used in the phase of convert input and analysis. It includes that the values of the seed, the repeat number, the clusters number, the max iterations number, the run cores number, the path and the sample analyzed. After the parameters are passed to the ccube pipeline, the phase of ccube pipeline will work if the flag “run\_ccube” is set true.

The number of samples which will be converted and analyzed is based on the parameter “selectedfile” and “sampleNum”. The second phase is to get the chromosome, position, vaf\_number, ref\_number, the copy number of major and minor from the files of VCF, segment and pp\_table of the samples selected, and write them to the TSV file on the target folders. The format of these data is combined suitable for the Ccube input.

The third phase is that Ccube tool analysis the mutations and output the target files. The parameters passed will decide the accuracy and speed of the Ccube analyzing. According to the default value of the parameters, the average working time for first 10 samples is 54.33 seconds.

The output files are including three text files which are saved on different folders. The subclone structure file has the number of ssms and the CCF of all clusters. The mutation assignments file has the assignment of mutations in the different clusters. The multiplicity file shows that the chromosome, position, tumor copy numbers and multiplicity. They can be read by LDA and NMF in the stage two.

About testing the Ccube pipeline, the command shell will show that the name of samples selected, and the working time of analyzing the sample. The target folders will have the output folder.

Reference

[1] Yuan, K., Macintyre, G., Liu, W. and Markowetz, F. (2018). *Ccube: A fast and robust method for estimating cancer cell fractions*.