# pyclone

## Background survey

Although development of measuring allele prevalence by using deep sequencing has grown successfully, there was not statistical method for clustering deep digital sequencing of mutations into biological groupings.

Aiming at classifying sets of deeply sequenced somatic mutations into presumptive clonal structure, a hierarchical Bayes statistical model for assumption of clonal population clusters was developed by Andrew Roth and his workmates [1]. Except from classification, the model calculates clusters’ cellular prevalence (which refers to cancer cell fraction (CCF) in this dissertation) as well as makes an explanation of allelic imbalances.

The paper highlights that, during clonal phylogeny, the clusters of mutations which locate at the same point are sharing cellular prevalence. Taking advantage of this property, the model considers it as markers of clonal populations. However, allelic prevalence of a mutation contributes by several features, which means it is not relevant directly to cellular prevalence. it will be aggravated especially single sample estimation is taken.

Based on the property and Bayes theory, a Bayes statistical software tool, PyClone, was developed to overcome above difficulty. Pyclone consists of four innovation modelling processes, including beta-binomial emission densities, estimation of flexible prior probability for possible mutational genotypes, Bayesian nonparametric clustering and joint analysis of multiple samples.

The software tool was tested in idealized datasets which extracted from four 1000 Genomes project samples and mutational profiles of multiple samples provided from a high-grade serous ovarian cancer (HGSOC) [2][3]. By comparing two results, Pyclone obtain the similar conclusions among two datasets, which proves its robustness. The experiments illustrate Pyclone provides a reliable statistic inference method to support researches in progression of cancer. Additionally, according to cluster accuracy, paper reveals that Pyclone with beta-binomial emission densities with parental and total copy-number priors surpass all other methods in the process of emission densities [1].

## Requirement

For the Pyclone, the operation system is Unix system only (like Linux and MacOs) and the version of python must be 2.7.

Additionally, it should be mentioned that Pyclone is the main library, and Pyclone is an open-source code and has been released at [https://bitbucket.org/ aroth85/pyclone](https://bitbucket.org/%20aroth85/pyclone). According to its documents, the main dependencies are PyDp (higher or equal than 0.2.3), PyYAML (higher or equal than 3.10). Those two libraries are used for analysing. Besides, Matplotlib (higher or equal than 1.2.0), Numpy (higher or equal than 1.6.2), Pandas (higher or equal than 0.11), Scipy (higher or equal than 0.11) and seaborn (higher or equal than 0.6.0) are required for plotting result and clustering. Pyclone and all the required libraries can be installed by “conda install pyclone -c aroth85”

Additionally, R scripts are used to pre-process input files and generate output files. The required libraries for R scripts are “dplyr”, “mcclust”.

The final requirement is source data. According to the documents of pyclone, the input is a set of deeply sequenced mutations from one or more samples extracted from a single patient as well as allele-specific copy number at each mutation trajectory in each sample [1]. In our case, there are 500 samples from different patients.

In summary:

* OS: Unix System (Linux and MacOs)
* Coding language: Python and R
* Main library used: Pyclone
* Other libraries:

Python: PyDp, PyYAML, Matplotlib, Numpy, Pandas, Scipy, Seaborn

R: dplyr, mcclust

* 500 samples data

## Design and Implement

The main framework contains pre-processing, analysing and postprocess.

In pre-processing, 500 simulated tumour samples are inputted as source data. There are constructed as following:

However, those data cannot be analysed directly. The required field of Pyclone is [4]

* unique ID (mutation\_id)
* the number of reads covering the mutation which contain the reference allele (ref\_count)
* the number of reads covering the mutation which contain the variant allele (var\_count)
* the copy number of the cells in the normal population (normal\_cn)
* the minor copy number of the cancer cell (minor\_cn)
* major copy number of the cancer cell (major\_cn)

Thus, a convert program is needed to extract useful information from given files and generate target file contains required field. Additionally, a configure file is also needed for analysing. In configure file, working directory, density method, number of iterations and sample information are needed to be specified.

The next step is analysis, it can be done by calling function “run\_analysis” in pyclone. This stage will generate a folder “trace” to store the clustering process.

In the final stage, all the result produced by Pyclone are postprocessed by R script. The output files of this process are sub-clonal structure containing cancer cell fraction (CCF) for each putative cluster, mutation assignment for each mutation and multiplicity.

The purpose of this project is to create a pipeline for three tools and two analysis methods. Thus, the code of Pyclone exports interfaces for final code integration. There are:

* Prefix: the prefix name for inputted samples.
* burnIn: the burnIn number N will drop out first N row of output data for analysing.
* Num\_iter: the number of iterations.
* Purity: purity of input sample.

References

[1] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-côté, A. and Shah, S.P., 2014. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, *11*(4), p.396.

[2] 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), p.1061.

[3] Harismendy, O., Schwab, R.B., Bao, L., Olson, J., Rozenzhak, S., Kotsopoulos, S.K., Pond, S., Crain, B., Chee, M.S., Messer, K. and Link, D.R., 2011. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. *Genome biology*, *12*(12), p. R124.

[4] https://bitbucket.org/aroth85/pyclone/wiki/Usage