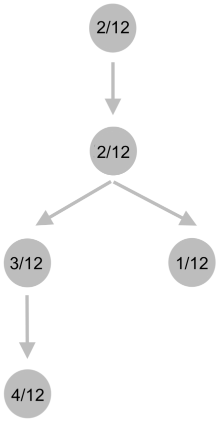
Assignment 2
CCBI Cancer evolution
Deadline: Friday April 5th, 2019

A. Getting to know your tools We start with the following evolutionary tree, where the node labels represent clone sizes π.



1. Compute and report the CCFs τ for each mutation set. For the following questions you will be using these cancer cell fractions (CCFs) τ.
2. PyClone[1] and Ccube[2] are programs that infer the CCFs given VAFs and CNAs using the infinite mixture modeling approach we discussed in the lectures.

   Run both programs and report their runtime for two datasets:

   - Files `inputpyclone_noCN.tsv` and `inputccube_noCN.RDS` contain the number of reads for the reference and variant alleles at each locus, in a situation in which there is no copy number change.
   - Files `inputpyclone_0_7CN.tsv` and `inputccube_0_7CN.RDS` contain the number of reads for the reference and variant alleles at each locus, in a situation in which there are copy number changes in 70% of loci.

   In PyClone, you can use the same parameters as in the tutorial.
3. For each case, what is the inferred number of clusters of CCFs? Plot the true CCFs vs the estimated ones. How well did the reconstructions work?
4. Files `inputccube_0CN_20.RDS` and `inputccube_0CN_100.RDS`, and `inputccube_0_7CN_20.RDS` and `inputccube_0_7CN_100.RDS`, are downsampled versions of `inputccube_0_7CN.RDS` and `inputccube_0_7CN.RDS` respectively. How many CCFs does Ccube find now?
5. Comment on the methodological similarities and differences between Ccube and PyClone. In particular, you will have observed that one method runs faster than the other one – why?

B. Simulate your own data In the first part of the assignment, the data were given. In particular, you learned what data format you need to run PyClone and Ccube. Now you will simulate different data sets yourself.

6. In the lectures, you learned three scenarios of how copy number changes and mutations can be combined.

   Given the clonal frequencies of the tree in Assignment #1, simulate copy number states for all loci (explain how you do it) and show how VAF changes under these different scenarios by plotting VAF vs CCF for all loci.

[1] `https://github.com/aroth85/PyClone`, which can be installed with `conda install -c aroth85 PyClone` for Python 2.7.
[2] `https://github.com/keyuan/Ccube`; Description `https://www.biorxiv.org/node/144692.full`

7. Now run PyClone and Ccube using the VAFs you have simulated. How do the results change compared to the example in (A)? Can you explain this observation?

8. Starting again from the tree in Assignment #1 (= ‘the primary tumour’), extend it by two ’metastatic trees’ which are examples of (i) early and (i) late spread. Plot the CCFs of the metastases versus the primary. Explain how the difference between early and late spread shows in these plots.