Using PISCA-box

Data input

At the moment, data must be in the form of a fasta file, which looks something like this:

Absolute copy number sample:

```
>Sample1

222344

>Sample2

223334
```

Where each ‘site’ represents the numeric value of either a chromosome bin or segment, and the sites are ‘aligned’ like in a DNA fasta file where each column is a homologous site.

There isn’t an easy way to look at the alignment like there is DNA, but if you plot out your data you should hopefully see samples have some events in common.

Date information

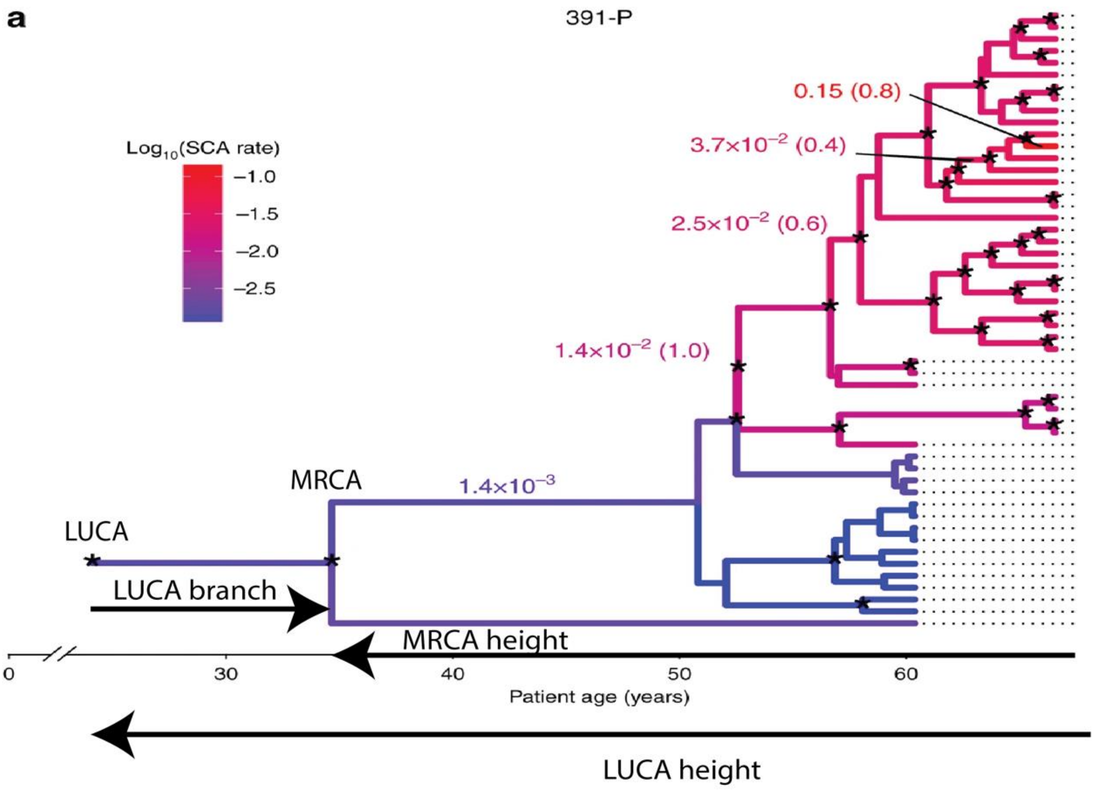
You should include the patient age when each sample was collected (as a decimal). PISCA can then use time-based phylogenetic models. This should be in csv format like this:

```
Sample, age

Sample1, 59.1

Sample2, 63.2
```

Luca_height and Luca_branch specifications

LUCA or the last universal common ancestor, is the last normal cell at the root of a PISCA tree. In a SCA context, we expect this cell to be in a perfectly diploid state. The MRCA or most recent common ancestor, is the cell which gave rise to all the samples. It may have some changes present that all samples have in common, and therefore not be diploid. Unless the tumour is polyclonal or transformation is not SCA-driven and the tumour is very slow-evolving, we expect chromosomal changes between LUCA and MRCA.

To model this we have a degree 1 node (MRCA) connecting LUCA with the rest of the samples. The MRCA is 1 free parameter but with two implementation parameters: luca_height and luca_branch which are two different ways to parameterize this last node of the tree.



LUCA height = LUCA branch + MRCA height

Typically, one of the parameters will be either fixed or estimated, and the other re-calculated internally. Similarly, MRCA operators modifying LUCA should only act on one of the two parameters (LUCA height or LUCA branch).

Implementation:

Use case 1, normal tissue. When studying normal tissue, LUCA time is known (more or less between conception and birth). Here LUCA height would be fixed as the length between zero and the age at latest sample, and the LUCA branch would be between zero and the age of earliest sample. *This would normally be the case for phyfum or biallelic binary.*

- Luca height is fixed to age at last sample
- Luca branch is bounded (from 0 to age at first sample)
- No need to operate on luca branch or height
- No need to put a prior on luca height

Use case 2, tissues without known age (e.g., premalignant conditions or cancer): In this case, we need to estimate LUCAs position, which can float between the MRCA and birth. *This would normally be the case for BE or IBD data.*

The easiest way of doing this is making the branch length rule and adding an operator for the branch length (if we operated on the height, a lot of times we would be suggesting invalid heights, depending on changes on the rest of the tree).

In this case, we will have a prior on luca_height (because it is biologically meaningful, the branch not so much) but operate on luca_branch for the practical reason explained above.

- Luca height is bounded (maximum is patient age and minimum is (age at last sample – age at first sample)
- Luca branch is bounded (from 0 to age at first sample)
- we operate on Luca branch
- We put a prior on luca height (maximum is patient age and minimum is (age at last sample – age at first sample).