ScisTree: A Program to Maximum Likelihood Cell Tree Inference and Genotype Calling from Noisy Single Cell Data

User Manual

Version 1.1.0

March 20, 2019

Yufeng Wu
CSE Department, University of Connecticut Storrs, CT 06269, U.S.A.
Email: yufeng.wu@uconn.edu

Yufeng Wu, Accurate and Efficient Cell Lineage Tree Inference from Noisy Single Cell Data: the Maximum Likelihood Perfect Phylogeny Approach, manuscript, 2019.

1 Getting Started with ScisTree

1.1 Program availability

ScisTree is written in C++. Executables for popular platforms such as Linux 32 bits or 64 bits and MacOS are downloadable from GitHub:
`https://github.com/yufengwudcs/ScisTree`. Files can be downloaded using "Save Link/Target As..." After downloading the softwares, you may need to change file access permissions (e.g. chmod u+x scistree-linux). In case that you want to compile the code yourself, source code is also available for download at the above URL. To compile the code, first put the gzip file in the directory youd like and unzip it: use gunzip and tar commands such as:
▷ gunzip ⟨scistree-src.tar.gz⟩
▷ tar -xvf ⟨scistree-src.tar⟩

Then type:
▷ make at the prompt. This creates an executable called stells, which can be run by typing
▷ scistree at the prompt. You will need to specify some input options - see below.

1.2 What is ScisTree?

First, where does the name ScisTree come from? It stands for {S}ingle {c}ell {i}nfinite {s}ites {Tree}.

ScisTree is designed to infer cell tree (the evolutionary tree of single cells) and call genotypes from noisy single cell data. ScisTree takes uncertain genotypes in the form of genotype probability. This is because genotypes called from single cell sequence data tend to be very noisy. It is usually difficult to call a fixed genotype at a position. Different from several existing methods for cell tree inference, ScisTree works with uncertain genotypes with **individualized** genotype probabilities. That is, each genotype (at a site and a cell) can have its own probability, which specifies how likely this genotype has a particular genotype state. This can better utilize information contained in single cell data: often some genotypes in the single cell data can be almost fully determined while others have much larger uncertainty.

ScisTree infers cell tree and call genotypes simultaneously. It can deal with single cell technological noises such as doublets. One key advantage of ScisTree over some existing methods is that ScisTree is very efficient: it works for data with hundreds of cells and thousands of single nucleotide variants (SNVs). My tests show that ScisTree can be 100 times or more faster than existing methods such as SCITE.

1.3 How does ScisTree work?

ScisTree assumes the infinite sites model. This allows a very simple algorithm for finding the maximum likelihood estimate (MLE) of the cell tree and genotypes that maximize the likelihood of the data under the infinite sites model. Refer to the paper for more details on the methodology of ScisTree.

2 Functionalities and Usage of ScisTree

2.1 Preparing inputs

To run ScisTree, the user needs to provide the genotype probabilities for the SNVs of the cells under study. The genotype probability is specified in a matrix. Genotypes can be either binary or ternary. Here is a simple example of the input.

```
# This is an example.
HAPLOID 5 4
0.8    0.02    0.8    0.8
0.02    0.02    0.02    0.8
0.8    0.02    0.02    0.8
0.02    0.8    0.8    0.8
```

0.8 0.02 0.8 0.02

ScisTree ignores lines starting with #, which are considered to be comments. The first (non-comments) line should have: $< \mathsf{FORMAT} >< \mathtt{num-sites} >< \mathtt{num-cells} >$. Here, "FORMAT" can be either "Haploid" or "Ternary". Haploid format refers to binary genotypes, while ternary format refers to ternary genotypes. The user needs to specify the number of (SNV) sites and the number of cells. There is a line for the probabilities for genotypes of each site. For each genotype at a site, one specifies the genotype probability sequentially: for binary genotype, use a single value to specify the probability of genotype 0; for ternary genotype, use two values to specify the probability of genotype 0 and genotype 1. In the above example, at the first site, genotype probabilities given mean that the four cells have probability of 0.8, 0.02, 0.8 and 0.8 of being genotype 0. Note that the probability of being genotype 1 is not specified since the probabilities of being 0 and 1 add up to 1.

2.2 Usage

To run ScisTree, you must provide an input file with the single cell genotype probability (using the format as specified above).

▷ ./scistree-mac <genotype-probability-file>

By default, ScisTree outputs the inferred cell tree (in the Newick format) only. In order to output the called genotypes, you should specify the "-v" option:

▷ ./scistree-mac -v <genotype-probability-file>

ScisTree outputs the imputed genotypes from genotype probability. For reference, it first outputs the maximal probability genotypes (taking the most probable genotype at each position in the matrix). Then ScisTree shows the genotypes that are changed from the maximal probability genotypes. The called genotypes are shown below "Imputed genotypes:".

By default, ScisTree infers a cell tree without branch length. To infer a tree with branch length, use the "-l" option.

▷ ./scistree-mac -l <genotype-probability-file>

ScisTree has also implemented several additional functionalities. These include doublet imputation. See the following for more details.

2.3 Command line options

For ease of reference, I now provide the list of (optional) command line options.

1. -d ⟨threshold⟩: only use genotypes with genotype probability clearly favoring a single genotype by a clear margin (specified by the threshold value). This option works for binary genotypes at the moment. In this case, if the genotype probability of 0 is 0.98 and the threshold is 0.9, this genotype is considered to be reliable: the difference between 0.98 and 0.02 (the probability of the alternative state) is 0.96, which is larger than the threshold value (0.9). If the genotype probability of 0 is 0.9 instead then

this genotype will be discarded when constructing initial trees. This is because the difference between probabilities of alternative genotypes is 0.8, which is smaller than the threshold value. Default setting is that ScisTree will use all genotypes for initial tree construction. My simulation shows that when data contains significant noise, choosing a relatively high threshold (e.g. 0.9) can improve the inference accuracy.

2. -v: output more information about the results by ScisTree. It outputs the called genotypes, along with genotypes called by simple single site maximal probability genotypes, difference between the two set of genotypes, and some additional information.

3. -l. Infer cell tree with branch lengths. Note that this uses a different likelihood model. So don't compare likelihood values for the case of with branch lengths with the case without branch lengths.

4. -n: only output the cell tree constructed by simple neighbor joining. In this case, neighbor joining is run with the maximal probable genotypes from single positions. This can be useful when one wants to find a quick cell tree for very large data.

5. -e: output a tree that is called perfect phylogeny. Briefly, this tree may not be binary. The perfect phylogeny is implied by the imputed genotypes. The main difference is that by default, ScisTree infers a binary tree. When the data size is small, some branches may not be well supported (e.g. there is no mutations along a branch). In this case, turn on "-e" option may help to give a tree that has significant support from the data (i.e. avoid false positives).

3 Revision History

1. 03/20/2019: Release of v.1.1.0. Added the option to discard low quality genotypes when constructing initial trees. My simulation shows that this can be important to use when dealing with data with significant noise.

2. 10/01/2018: Release of v.1.0.0. Include basic functionality of cell tree inference and genotype calling from uncertain genotypes.