Producing inferences of cancer cell phylogenies

Joseph Grace (58610162)
supervised by Sasha Gavryushkina, Alex Gavryushkin

January 2024

1 Introduction

The purpose of this project was to analyse genetic sequencing data of human cancer cells provided by German Cancer Research Center (DKFZ). We use two software packages, BEAST2 and CellPhy, to attempt to establish a phylogeny of various samples of these cells. We aim to establish confidence in these models, which may be used in future to infer differences between two populations of these cells.

2 Data Background

The sequencing data had been obtained from human glioblastoma cells grown in mice. The Wnt antagonist SFRP1 was introduced into some glioblastomas. The researchers found SFRP1 affected development of the tumors. [1]. There were 5 batches of cells: SS3_1, SS3_2, SFRP1_1, SFRP1_2, and SFRP1_3, where the SFRP1 batches were obtained from mice 'treated' with SFRP1. We consider the SS3 batches a control.

We received the data in the form of a genotype matrix containing Single Nucleotide Variant (SNV) data for 1431 (362 SS3_1, 312 SS3_2, 263 SFRP1_1, 245 SFRP1_2, 249 SFRP1_3) cells across 89484 sites (comprising 189564 variants). The matrix was quite sparse.

We also received metadata on the cells - which we used to label cells with their cycling state (cycling, non-cycling, or unknown) and stage (quiescent, activated, differentiated, or unknown) - and metadata on sites, which we used to log which sites were used in the analyses.

3 Software

3.1 BEAST 2

BEAST 2 is a tool for performing Bayesian analysis, in particular using the Markov-Chain Monte Carlo method, on rooted phylogenetic trees [2]. The results given here were produced using BEAST v 2.7.6 with the 'beast-phylonco' package [3].

Tracer v1.7.2 [4] was used to monitor the parameters of the MCMC runs in order to measure convergence.

TreeAnnotator, a program that comes bundled with BEAST 2, was used to summarise the trees produced by each MCMC run.

3.2 CellPhy

CellPhy [5] uses maximum likelihood algorithms to perform tree inference, and includes models for dealing with sequencing errors. In contrast to BEAST 2, CellPhy produces an unrooted tree.

3.3 FigTree

FigTree v1.4.4 [6] was used to display and export trees produced by both BEAST 2 and CellPhy.

4 Method

4.1 Sampling and Filtering

Most data manipulation was carried out using the pandas Python library (`https://pandas.pydata.org/`).

Firstly we combined the SNV data for each site to yield between zero and two characters per site. Sites containing more than two characters were ignored. This resulted in sequences with an unphased diploid genotype.

Next cells were chosen from subsets of the batches, based on what proportion of sites had data for any particular cell (site density).

Here we consider three groups of cells:

- **control-20**: The top 20 cells from the SS3 batches by site density.
- **sfrp1-20**: The top 20 cells from the SFRP1 batches by site density.
- **sfrp1-split-21**: The top 7 cells from each SFRP1 batch by site density combined.

  The sfrp1-split-21 group was created as sfrp1-20 had an over-representation of SFRP1_1 cells, and thus might not be reliable to assess whether clades form from the batches.

For BEAST 2 analyses, the sites were filtered so that the proportion of cells with data for any particular site (density) met a certain threshold. Any sites that had a constant value (excluding unknowns) were also filtered out, as they hold no information on differences between cells.

- control-20 was filtered to a threshold of 0.7 density, yielding 674 sites.
- sfrp1-20 was filtered to a threshold of 0.8 density, yielding 597 sites.
- sfrp1-split-21 was filtered to a threshold of 0.7 density, yielding 590 sites.

For CellPhy runs, all non-constant sites were used.

4.2 MCMC Model

Two MCMC models were used for each group:

4.2.1 Error Model

Using the example XML file for the error model given in the documentation of beast-phylonco [3] (beast-phylonco/examples/test_GT16_error.xml in the beast-phylonco repository) as a template, we ran BEAST 2.7 on the datasets for 10,000,000 iterations, or until the ESS (Effective Sample Size) of all logged parameters (tree likelihood, mutation rates, error delta and epsilon, and tree height) was above the standard value [7] of 200. The runs for control-20 and sfrp1-split-21 reached this threshold, whilst sfrp1-20 did not. (Hence the tree for sfrp1-20 is not shown here).

4.2.2 Error Model + Birth-Death Prior

Here a Birth-Death model was used to inform the prior distribution for the MCMC. Again, we ran BEAST 2.7 until the ESS of all logged parameters was above 200. For this model control-20 and sfrp1-split-21 reached this threshold, however sfrp1-20 failed to reach it after 12,000,000 iterations. However most parameters did reach the threshold, so the tree is included here.

4.3 CellPhy

The data was converted to FASTA format and analysed by running the provided 'cellphy.sh' script. This produces a best tree with bootstrap support values.

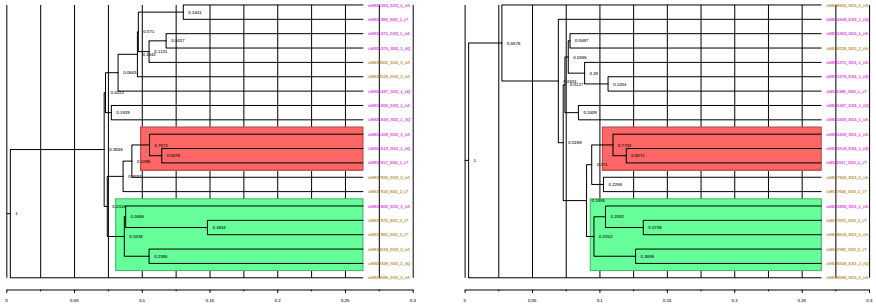
5 Results

The following figures show the trees produced as above, displayed using FigTree. Some examples of similar/identical clades between runs are highlighted in identical colours. In trees produced by BEAST, internal nodes are labelled with posterior support values for the node's clade. In trees produced by CellPhy, internal nodes are labeled by bootstrap support values.

Cell names are coloured by batch and are of the format cell{number}_{batch}_[nc?][QAD?] where 'n' or 'c' represents non-cycling or cycling state repectively, 'Q', 'A' or 'D' represents the stage, and '?' represents missing metadata.

6 Discussion

Since each batch of cells was sampled from independently growing tumors, we would expect cells from each batch to form distinct clades. However this is not seen in any of the produced trees. As an example of this, both BEAST 2 analyses on sfrp1-split-21 produce a clade with high support, which contains one cell from each batch (highlighted purple in Figure 5). Furthermore this mixing of batches in the produced trees is frequent. This indicates the models used are not approximating the true phylogeny accurately.



(a) No Birth-Death Prior (b) Birth-Death Prior

Figure 1: Trees produced for control-20 group by BEAST 2.

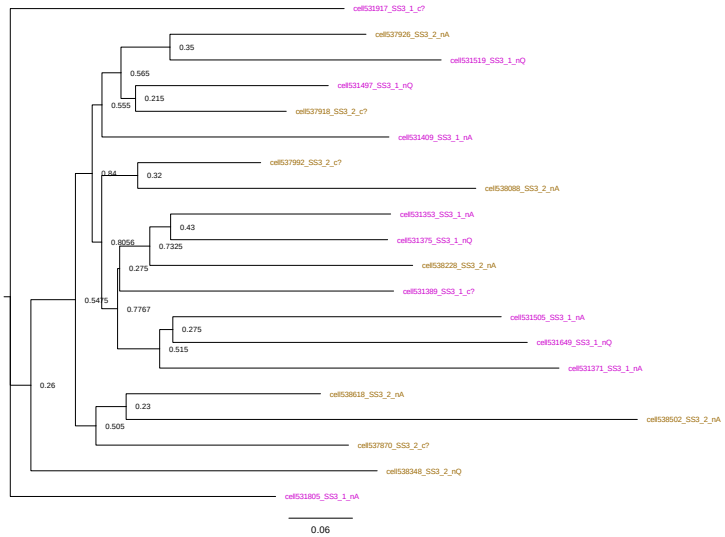


Figure 2: Tree produced for control-20 group by CellPhy.

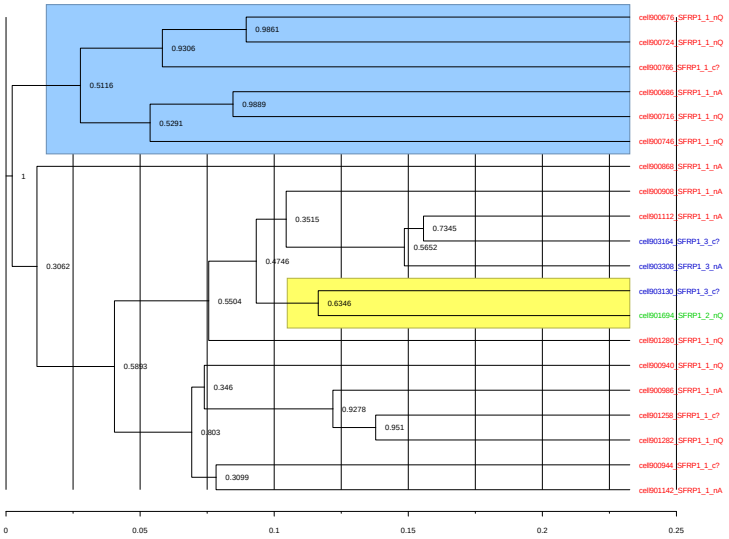


Figure 3: Tree produced for control-20 group by BEAST 2 (With Birth-Death prior).

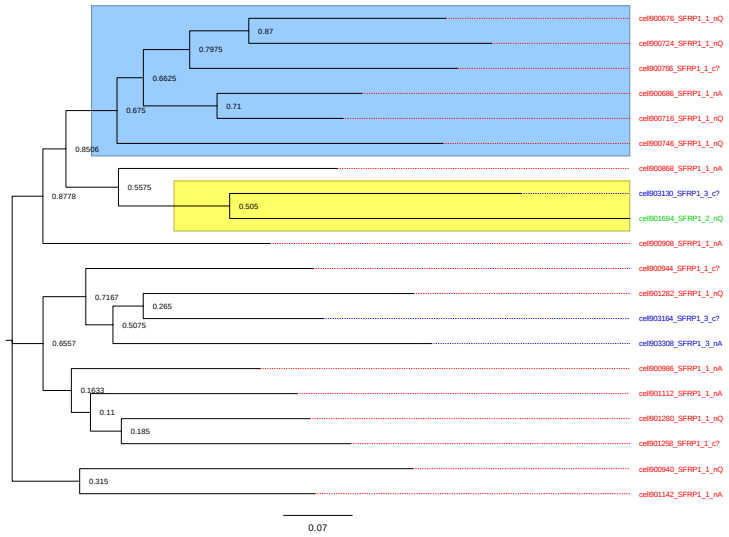
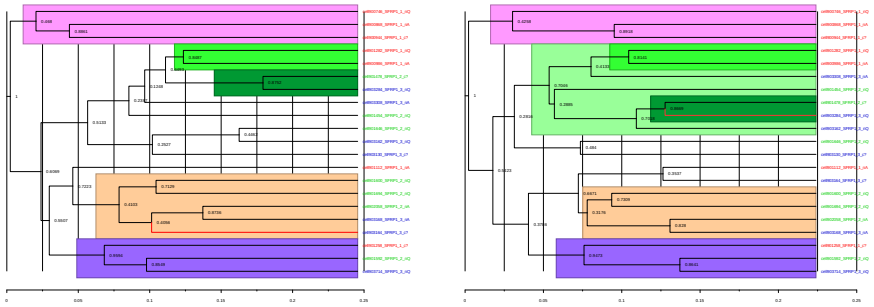


Figure 4: Tree produced for sfrp1-20 group by CellPhy.



(a) No Birth-Death Prior (b) Birth-Death Prior

Figure 5: Trees produced for sfrp1-split-21 group by BEAST 2.

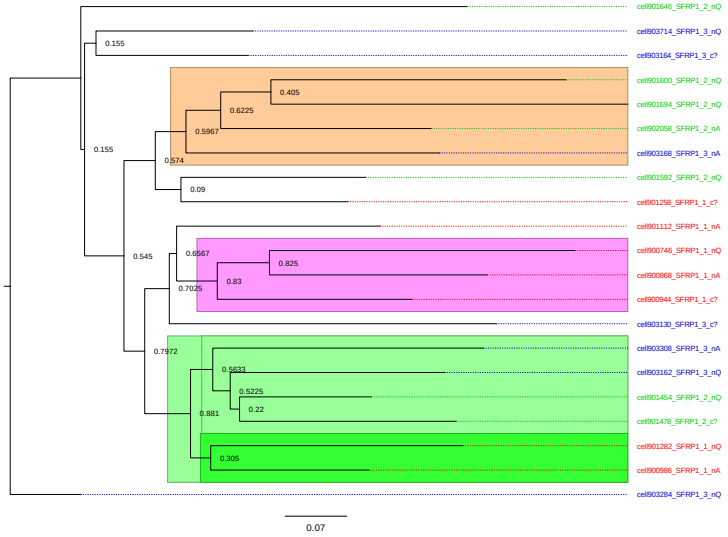


Figure 6: Tree produced for sfrp1-split-21 group by CellPhy.

Another problem is that the support values for internal nodes are often low (less than 0.5), especially for the trees produced by BEAST 2 for control-20. The sparseness of the data may be a contributing factor here.

7 Conclusions

Although some progress was made in modelling the phylogeny of selected cells, we failed to produce a feasible tree using the above datasets and models.

Further work is needed, both to ensure quality of the sampled data, and to find models capable of producing feasible trees with the sampled data.

One potential avenue of exploration could be to filter sites on more sophisticated criteria than just data density, for instance taking variation across cells into account.

References

[1] L. C. Foerster, O. Kaya, V. Wüst, M. Bekavac, K. C. Ziegler, V. Akcay, N. Stinchcombe, N. G. Perez, X. Ma, A. Sadik, P. U. Le, K. Petrecca, C. Opitz, H. Liu, C. R. Wirtz, S. Anders, A. Goncalves, and A. Martin-Villalba, "Identification of astrocyte-driven pseudolineages reveals clinical stratification and therapeutic targets in glioblastoma," *bioRxiv*, 2023.

[2] C. for Computational Evolution, "Beast 2." `https://www.beast2.org`. Accessed 09/02/2024.

[3] "beast-phylonco." `https://github.com/bioDS/beast-phylonco`. Accessed 09/02/2024 (commit 0b03608).

[4] A. Rambaut, A. J. Drummond, D. Xie, G. Baele, and M. A. Suchard, "Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7," *Systematic Biology*, vol. 67, pp. 901–904, 04 2018.

[5] A. Kozlov, J. M. Alves, A. Stamatakis, and D. Posada, "Cellphy: accurate and fast probabilistic inference of single-cell phylogenies from scdna-seq data," *Genome Biology*, vol. 23, 2022.

[6] "Figtree." http://tree.bio.ed.ac.uk/software/figtree/. Accessed 09/02/2024.

[7] R. Lanfear, X. Hua, and D. L. Warren, "Estimating the Effective Sample Size of Tree Topologies from Bayesian Phylogenetic Analyses," *Genome Biology and Evolution*, vol. 8, pp. 2319–2332, 07 2016.