Assessing Phylogenetic Posteriors for Single-Cell Sequencing Tumour Data

Yutong Li

Word counts: 4994
Supervisors: Prof. Alexei Drummond, Dr. Jonathan Klawitter, and Dr. Kylie Chen
Keywords: Bayesian analysis, rogue taxa, conditional clade distribution, entropy, single-cell sequencing, variant calling

Abstract

Single-cell sequencing (SCS) can obtain tumour cell sequences, enabling the inference of tumour cell phylogenies, though with errors and biases. We aim to analyse the posterior distributions of Bayesian phylodynamic inference for SCS tumour data. Particularly, we test two methods: one to mitigate and the other to comprehend noise in the data and enable tumour phylogenetic analysis. (1) We use a novel algorithm to identify rogue taxa with methods based on the CCD and entropy. Rogue taxa are the taxa contributing the most phylogenetic entropy to the posterior. We remove them to uncover a "skeleton" phylogeny of well-supported lineage relationships. (2) We will systematically investigate the effect of parameter choice in variant caller software used to pre-process sequence read data. We aim to discover how the pre-processing parameters affect the SCS noise and the phylogenetic signal. Together, these approaches can provide us with the understanding necessary to produce more accurate estimations of tumour age, a more informative summary of the posterior distribution, and a better characterisation of sequencing noise. In summary, our goals are to assess methods to comprehend noise better and to obtain accurate and easy-to-interpret tumour cell phylogenetic inference, which we hope can contribute to better clinical treatment for cancer patients in the future.

Section 1: General background

1.1 Cancer, somatic evolution, and tumour phylogeny

Cancer is a disease that is driven by genetic alterations in the genome. These DNA alterations can include somatic mutations (base substitutions), deletions, rearrangements, and changes in DNA segment copy numbers (Stratton, 2011). The main source of these alterations is genomic instability, especially DNA replication and repair error (Liu, Duenas, Zheng, Reckamp, & Shen, 2022). During DNA repair failure, the somatic mutation rate also increases. Somatic mutations are mutations that occur as somatic cells proliferate (during mitosis), and these can play a role in driver mutations of cancers. The somatic mutations occasionally occur in a set of key genes (i.e., cancer genes) and affect normal cells' proliferation, differentiation, and apoptosis while encouraging tumour cell development. Some number of somatic mutations will occur in cancer genes – this does not specifically tell us how they harm normal cells unless a gene function analysis study is performed (Stratton, 2011). Regardless, the accumulation of somatic mutations can be used to infer the somatic evolutionary trajectory.

One of the primary sources of data for the estimation of tumour phylogeny is genomic sequence data. Samples can either be bulk or single cells, whereas the sequenced samples can be DNA or RNA. Bulk sequencing and single-cell sequencing (SCS) are both common data types. Currently, most genetic analyses of tumours are based on bulk sequencing, which is a high-throughput sequencing technology for analysing DNA. Bulk sequencing is done on a group of cells, so the DNA for multiple cells is mixed in a sample, which limits its usage (Hegenbarth, Lezzoche, De Windt, & Stoll, 2022). A challenge for bulk sequencing analysis is that cancer cells are often heterogeneous - with different genotypes (Gawad, Koh, & Quake, 2016). On the other hand, SCS enables the analysis of the genetic (DNA) or transcriptional profile (RNA) of each individual cell within a heterogeneous tumour population (Malikic, Jahn, Kuipers, Sahinalp, & Beerenwinkel, 2019). There are two kinds of SCS: single-cell RNA sequencing (scRNA-Seq) and single-cell DNA sequencing (scDNA-Seq). The scRNA-Seq is often used to research the expression and functions of the cells, while scDNA-Seq is used to uncover genetic heterogeneity. After isolating the cells and lysing them to obtain DNA, the genome is amplified to increase the number of quantities of DNA available as templates (Andor et al., 2020). In this process, errors can be generated, including loss of coverage, decreased coverage uniformity, allelic imbalance, allelic dropout, and errors during genome amplification (Gawad et al., 2016). Although read coverage is naturally more even for scDNA-Seq, it also exhibits many of the same sources of error. Therefore, much effort has been focused on developing methods to correctly account for these sources of error so that the signal in the data can be accurately extracted.

However, accurate inference of tumour phylogenies remains challenging. In 2014, Jiao's group developed the PhyloSub model, which is a Bayesian model, to obtain explicit single nucleotide variant (SNV) population frequencies and estimate uncertainties in tumour phylogeny with both single and multiple tumour samples. However, the model only works well in high-throughput samples and performs less well when the number of SNVs is low in the tumour sample (Jiao, Vembu, Deshwar, Stein, & Morris, 2014). To overcome this shortcoming, in 2016, Donmez's group developed a CTPsingle model, which is also a Bayesian model, to obtain accurate sequencing results in low-throughput samples, but only for single samples. Nevertheless, CTPsingle is still unsuitable for tumour regions with high instability (Donmez et al., 2016). There is still much research and development of methods to obtain accurate estimates of tumour phylogeny.

Another problem for accurate phylogeny inference is tumour heterogeneity, which is one of the characteristics of malignant tumours leading to noisy posterior trees (Guo et al., 2023). There are three kinds of tumour heterogeneity: intra-tumour heterogeneity, inter-tumour heterogeneity, and inter-site heterogeneity (Gupta, Chandan, & Sarwat, 2022). We mainly talk about intra-tumour heterogeneity (ITH) in this proposed study.

ITH refers to different genotypes and phenotypes with diverse biological behaviours in the subpopulations of cells within a primary tumour or its metastases (Fisher, Pusztai, & Swanton, 2013). During evolution, somatic cells acquire random mutations that change their phenotypes, and the accumulation of the mutations is one of the processes of cancer evolution. There are two mechanisms that lead to intra-tumour heterogeneity: cell-autonomous and non-cell-autonomous. For cell-autonomous mechanisms, the little errors during DNA replication last for a long time and accumulate into genetic diversity in the giant tumour. At the same time, replication stress and mutated phenotypes cause unstable inheritance. Cell-autonomous mechanisms can cause some serious errors during DNA replication, forming obvious genotype changes. The non-cell autonomous mechanism is related to the tumour and microenvironment interactions. All these mechanisms contribute to intra-tumour heterogeneity (Fisher et al., 2013).

1.2 Bayesian phylogenetic inference

A primary goal of Bayesian phylogenetic inference is to produce accurate estimates of phylogenetic relationships and characterise the uncertainty in those estimates. In Bayesian phylogenetic inference, the Metropolis-Hastings (MH) algorithm is the most employed method, belonging to the Markov chain Monte Carlo (MCMC) method, producing a sample from the posterior distribution under complex phylodynamic models (Andrieu, de Freitas, Doucet, & Jordan, 2003; Gay, Baker, & Graham, 2016; Ramachandran & Tsokos, 2021). A well-known challenge in Bayesian phylogenetic inference is the nontrivial problem of determining the convergence of the MCMC chain. It is generally difficult to determine whether the MCMC chain has been run long enough to sample from the posterior distribution accurately (Cowles & Carlin, 1996). Furthermore, practical considerations mean that the number of effectively independent trees sampled from the posterior distribution is usually relatively small for non-trivial phylogenetic inference problems, and the degree of autocorrelation in the resulting sample is difficult to determine.

Standard Bayesian statistical tools cannot be directly applied to tree topologies (i.e. tree space). Topologies are discrete, categorical, and non-nested (Cranston & Rannala, 2007), while including divergence times produces a non-Euclidean continuous tree space with metric properties that depend on the parameterization (Gavryushkin & Drummond, 2016). These features mean that Bayesian phylogenetic inference falls outside the score of standard Bayesian statistical tools such as Stan (Gelman, Lee, & Guo, 2015), JAGS (Plummer, 2003) and BUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) while complicating certain kinds of hypothesis testing (Cranston & Rannala, 2007). As a result, specialised software has been developed to produce posterior samples from phylogenetic tree space using MCMC, including

BEAST1 (Suchard et al., 2018), BEAST2 (R. Bouckaert et al., 2019), MrBayes (Ronquist et al., 2012) or RevBayes (Höhna et al., 2016). Averaging over the tree space is done using MCMC, and each tree is sampled with a probability proportional to its posterior probability (Suchard et al., 2018). This allows us to quantify the uncertainty of the estimated phylogenetic trees (R. Bouckaert et al., 2019).

We will use a Bayesian inference pipeline to infer and analyse tumour phylogeny to obtain the posterior tree distribution (Fig.1). We first collected cells with different genotypes from a tumour region, and then scDNA-Seq produces read data. This data is then analysed by a variant caller to call the single-nucleotide variants (SNVs) as a pre-processing step. The aligned sequences of SNVs are then used as data to perform inference using BEAST2 with a substitution and population growth model to get output as posterior tree distribution.

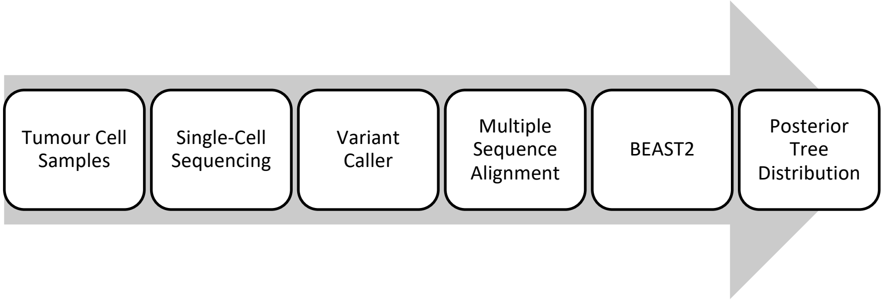

Figure 1. The original Bayesian inference pipeline to obtain posterior tree distribution consists of six sequential steps.

To visualise and interpret the posterior, we can use DensiTree, a software in the BEAST2 ecosystem that visualises all trees within the posterior tree distribution and calculates consensus trees (R. R. Bouckaert, 2010). With our existing workflow (Fig.1), we can only obtain a noisy phylogeny without clear relationships among taxa in tumour cell data (Fig.2A). In this proposed research, we aim to investigate two pathways to improving the phylogenetic signal and reducing the sequencing error noise (Fig.2B).

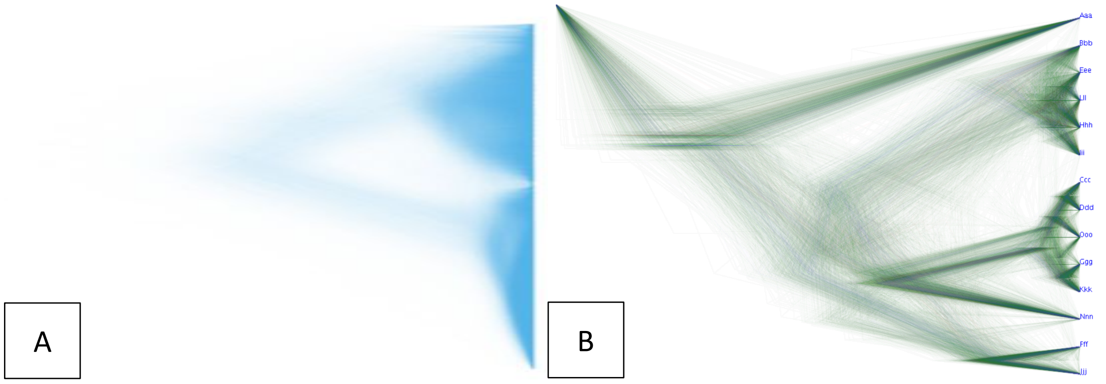

Figure 2. Phylogenies generated by DensiTree. **A.** A noisy phylogeny generated by DensiTree with a colorectal cancer dataset (Chen, Welch, & Drummond, 2021). **B.** An example of a phylogenetic posterior with lower phylogenetic entropy generated by DensiTree (R. R. Bouckaert, 2010).

1.3 Rogue detection

One source of phylogenetic uncertainty can be caused by so-called “rogue” taxa. Rogue taxa are taxa that have greater uncertainty in their phylogenetic placement than most other taxa in the analysis. We hypothesise that some rogues are harmless or even friendly since they can still contribute to improved accuracy in other parameters and subtree estimates by their inclusion. However, an arguably more likely scenario is the ‘mischievous rogue’ that its inclusion will cause a reduction in the signal-to-noise ratio of parameters of interest (Westover, Rusinko, Hoin, & Neal, 2013). We anticipate the presence of rogue taxa in tumour posterior tree distributions and that they contribute to uncertainty in estimates of parameters of interest (Aberer, Krompass, & Stamatakis, 2012). Hence, if we can successfully identify rogue taxa and remove them, we can get more informative posterior tree distributions.

An early attempt at developing a principled approach to the identification of rogue taxa was developed by Cranston and Rannala (2007). Through subtree pruning, multiple posterior tree distributions generated were analysed and compared after pruning specific taxa from the trees. The criteria used was to maximise the probability of the most probably phylogenetic tree in the remaining taxa subset. Summary data is often reported to explain the tree distribution and the topology with the highest posterior probability is called the maximum a posteriori or MAP tree. Cranston and Rannala aimed to maximise the probability of the MAP tree by pruning and labelling the pruned taxa rogues. At the same time, the credible set would become smaller to offer us more precise information about the relationships of the remaining taxa. As the number of pruned “rogue” taxa increases, a highly supported “skeleton” tree emerges.

According to previous research, the pruned taxa are so-called rogue taxa, contributing to tree topology differences but without contributing to any relationship among the tips (Wilkinson, Thorley, & Upchurch, 2000). However, this approach is highly dependent on posterior trees and only worked well in a small sample, in which simple estimates of the MAP topology provided an effective optimality criterion. Since the posterior tree distribution is only represented by a small sample, many unsampled trees will generally be in the posterior. Thus, for non-trivial posterior distributions, simple estimates of the MAP topology are ineffective, and the method of Cranston and Rannala can’t be used.

1.4 Conditional clade distribution – a tractable tree distribution

The conditional clade probability (CCP) was introduced in 2012 and brought a new way to estimate the posterior probabilities of individual tree topologies (including unsampled trees), even when only a small proportion of the topologies have been sampled by MCMC. A CCP score for a possible tree is multiplied by each clade's conditional frequency, which is the probability that the bi-partition pattern appears in all possible trees (Höhna & Drummond, 2011). One year later, the whole set of CCP was named the conditional clade distribution (CCD) by Larget (2013). The methods were improved with more experiments to demonstrate that CCD can produce a more accurate and faster approximation of the posterior distribution on tree topologies, including the nontrivial trees for rooted trees and splits for unrooted trees that do not exist in sampling. Larget introduced the principle of conditional independence into the method, which can simplify the calculation (Larget, 2013).

Though many efforts have been made, something still needs to be discovered within the posterior distribution because we mostly have some unsampled posterior trees. To quantify the uncertainties in the posterior trees, a measurement was represented by Lewis et al. (2016), which is entropy. In general, entropy represents how much unknown information is in a system and can tell us the difference between them. Lewis group applied entropy in CCD to say the difference between the prior and posterior distributions. However, the entropy they defined could not deal with low-density data and ignored them. Therefore, using the entropy to directly measure the information in a phylogenetic system needs to be more accurate.

1.5 Variant calling

Two kinds of variants can be called in scDNA-Seq: SNV and CNV. SNV is what we are mainly calling in this proposed study, with two strategies to call. One uses bulk samples as references to reduce the false-positive rate, and the other uses single-cell data only to compare the reads. In the process, the allelic imbalance correction exists to incorporate error correction by setting the threshold to filter technical noise or molecular barcoding (Gawad et al., 2016). Nevertheless, all the existing tools are not optimal for the scDNA-Seq variant calling because some of them cannot deal with technical biases well during cell processing. In contrast, others are specialised and non-generic (Valecha & Posada, 2022).

Commonly used tools for calling somatic mutations include MuTect2 (David et al., 2019), VarScan (Koboldt et al., 2009), and SCCaller (Dong et al., 2017), considered (Koboldt, 2020). VarScan can be set to detect low-frequency variants. As a first step, we will explore the VarScan variant caller as it is more user-friendly. VarScan uses position-sorted input of scDNA-Seq results of healthy and tumour cells by comparing with the reference sequence to call the variants. The algorithm can call different categories for sites, such as reference, germline, somatic, LOH, and indel filters can be told. We can set the parameters to filter the variants (Koboldt et al., 2009). MuTect2 aims to detect somatic mutations, including SNVs and indels, via a local assembly of haplotypes based on the Bayesian somatic genotyping model (David et al., 2019). At the same time, the single-cell variant caller (SCCaller) is specifically designed for SCS data to correct local allelic amplification bias in SNV calling (Dong et al., 2017). Due to previous research, VarScan has a lower false-positive rate but low sensitivity, which means it can be more cautious in providing results so that it may ignore some real variants. GATK, to which MuTect2 belong, shows highly balanced accuracy but requires high computing time and physical memory (Huang, Mullikin, & Hansen, 2015). Compared to the previous two callers, SCCaller is specifically designed for SCS data and accounts for amplification errors and has similar sensitivity to the previous ones (Dong et al., 2017), which makes it the most suitable for scDNA-Seq data. We will also explore the use of SCCaller in addition to VarScan.

Section 2: Research Question, Hypothesis, Aims and Objectives

2.1 Research Questions and Hypothesis of Proposed Study

In this proposed study, we will use two methods to deal with the noise in tumour phylogeny and get clearer posterior trees. The research questions asked by this study are:

(1) How can rogue detecting and removal methods be used to improve the interpretation of tumour phylogenetic inference?

(2) How do variant calling parameters affect sequencing error and phylogeny ifrom single-cell genome sequences of tumours?

We will use Entropy-based methods and change variant calling parameters in scDNA-Seq tumour data and test how much the methods can improve our posterior trees. We hope that the Entropy-based method could highly reduce the noise in tumour phylogeny, and 'minimum total reads' and 'minimum variant allele frequency' in VarScan can reduce the noise in the input sequences well.

2.2 Research Aim and Objectives

(1) Test the Entropy-based method to obtain clearer posterior trees.

- The Entropy-based method will be used after obtaining the initial posterior tree distribution.
- The clarity of posterior trees will be assessed with a visual inspection of DensiTree cladograms and the entropy value as a mathematical comparison.

(2) Evaluate the effect of variant calling parameters on the trees and error parameters.

- Some specific variant calling parameters will be adjusted to manipulate the input sequences.
- The branches, model parameters, and topologies of posterior trees would be observed while changing the variant calling parameters.

We are using these two methods above to understand noise better and test how much we can reduce the noise. The noise can be quantified in the Entropy-based method as entropy difference and can also be visualised in both methods as observing the DensiTree cladogram becomes clear.

Section 3: Methods, Experimental Approach, requirements, and expected findings

3.1 Entropy-based Methods

An Entropy-based method can be applied to deal with the posterior tree distribution to find rogue taxa. We will first introduce how we use CCD and CCP in the dataset.

When we obtain a posterior tree distribution {T1,…,Tk}, we assume that a parent clade C can be partitioned into many child clades. We define ρ(C) as all occurring partitions of C, and P is one of the partitions among ρ(C). We use f(C) and f(P) for clades and partitions to represent their frequencies. The CCP is:

$$\mathrm{Pr(P) = f(P) / f(C)} \qquad \text{(Eq.1)}$$

Note: Always f(P) ⩽ f(C), $\sum_{P\in\rho(C)} \mathrm{f(P)} = f(C)$, $\sum_{P\in\rho(C)} \Pr\,(\mathrm{P}) = 1$.

The probability Pr(T) of a tree T can be calculated by multiplying the CCP of each partition clade. To represent the nontrivial clades, we want C to belong to all clades $\mathrm{C}(T)$, including the nontrivial clades, of T, and let P(C, T) represent the partition of C in T. Therefore, we have the probability of a tree T:

$$\mathrm{Pr(T)} \approx \prod_{C\in\mathrm{C}(T)} \Pr\,(P(C,T)) \qquad \text{(Eq.2)}$$

Here is an example to show how they are calculated. We assume that there are two trees in our posterior sample, with each sampled once (Fig.3). The observation is that the root clade ABCDE is partitioned in two ways, ABCD|E and ABC|DE, while the root clade ABC is also partitioned in two ways A|BC and AB|C. The CCPs of trivial clades are Pr({ABCD|E}) = 1/2, Pr({ABC|DE}) = 1/2, Pr({ABCD|E}) = 1, Pr({A|BC}) = 1/2, Pr({AB|C}) = 1/2, Pr({B|C}) = 1, Pr({A|B}) = 1, and Pr({D|E}) = 1.

The resulting CCD provides us with probabilities of each tree calculated by Eq.2. In our posterior sample (Fig.3), the probabilities of each tree are both Pr(T) = 1/4. However, the sum of probabilities of all trees in posterior tree distribution should be 1. Therefore, there are two unsampled trees with a probability of 1/4 each (Fig.4).

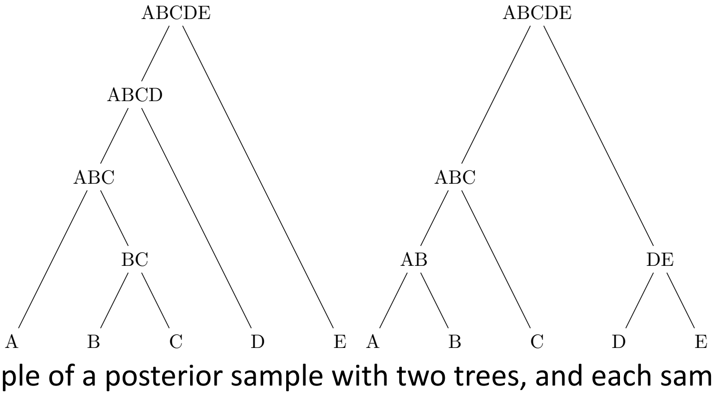

Figure 3. The example of a posterior sample with two trees, and each sampled once.

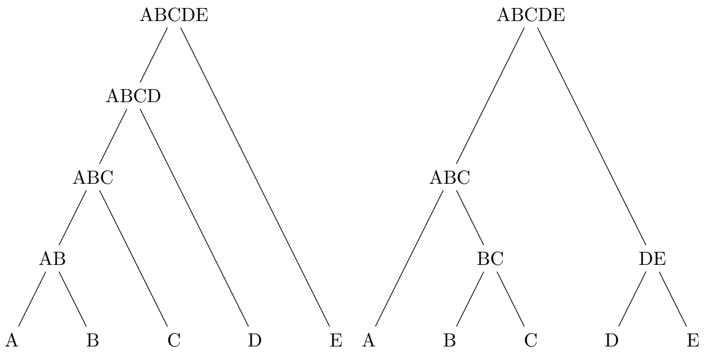

Figure 4. The unsampled topologies from the example posterior sample of Fig.3.

With CCD, we can obtain more information about the posterior tree distribution through dynamic programming, such as the number of different topologies, entropies, and the topology with the highest probability. To learn the difference between two CCDs, we can use entropy as a measurement of uncertainty. For each clade C, we define H*(C) as the entropy of this clade, and then we can obtain the entropy by Lewis et al. (2016):

$$\mathrm{H}^*(\mathrm{C}) = \sum_{P\in\rho(C)} -\Pr(\mathrm{P})(\log\Pr(\mathrm{P}) - H*(C1) - H*(C2)) \qquad \text{(Eq.3)}$$

When we ask C = C ρ, we can obtain the entropy of the CCD. When there is only one tree in the posterior distribution, the entropy would be 0, representing that there is no uncertainty in the posterior distribution. Vice versa.

To detect rogue taxa, CCD can be an option to estimate the whole posterior tree distribution. We first obtain the initial CCD using Eq.2 and the entropy with Eq.3. Then, entropies of CCDs without one taxon i are calculated and recorded, and we will finally get a set of entropies. We can rank the differences among the decreasing entropies by comparing this set of entropies with the reference entropy. The removed taxa that led to the biggest decrease in entropy will be considered rogue taxa, which we can then remove in the posterior tree distributions or the input sequences alignment to obtain skeleton trees as clearer posterior trees.

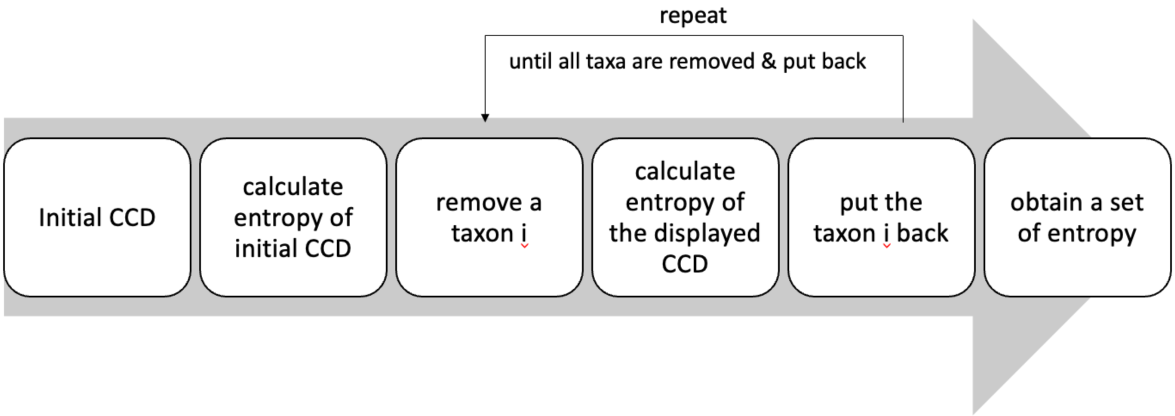
repeat
until all taxa are removed & put back
Initial CCD
calculate entropy of initial CCD
remove a taxon i
calculate entropy of the displayed CCD
put the taxon i back
obtain a set of entropy

Figure 5. The process of the entropy-based method.

However, the simple version (Fig.5) removes only one taxon in each iteration, which might not be optimal if we remove several taxa overall. A way to improve this method is also to consider whole clades as potential rogues. For example, if we know that there is a pair of adjacent tips, A and B, that always appear together and can be anywhere on a tree in the posterior tree distributions. When we then remove the tip A, the entropy would stay the same. We always know that tip B is there in the tree, so the uncertainties do not increase or decrease. However, when we remove A and B together, the entropy would decrease due to eliminating the whole set of uncertainties about this clade. Therefore, besides targeting only single taxa, considering also clades as rogues in each iteration, we can find a sequence of rogues that, for the same number of taxa removed, gives a higher decrease in entropy.

3.2 Effect of variant calling parameters

We also want to explore the effect of noise by modifying the variant calling parameters. However, in real single-cell datasets, we do not have ground truth, and the true evolutionary tree is unknown. Hence, we propose to evaluate the effects of the noise from the data and variant callers on the inferred phylogeny. Therefore, this method can only help us learn about how the parameters can change the branch lengths, topologies, and parameters of the tree right now and can help scientists develop more methods in the future.

The variant caller we focus on is VarScan, which is using a heuristic approach to call the variant tumour genotypes compared to a reference and healthy sequence (Koboldt et al., 2009). Varscan can detect different types of variants, including germline, somatic, LOH, etc. In this proposal, we will outline how Varscan can be used to evaluate the effects of variant caller parameters on phylogenetic inference. However, we note that other callers like SCCaller and Mutect2 can also be used.

For our analyses, we will use both public single-cell colorectal cancer datasets (Wu et al., 2017), and datasets of seven patients produced by our collaborator David Posada's lab. These datasets have between 24 to 79 cells and between 198 and 55,000 SNVs. For the input, we need to prepare a reference sequence, normal cell sequences, and tumour cell sequences. The normal cell sequences should come from the same individual. Within the VarScan somatic mutation calling, comparisons would be made only if both normal and tumour sequences have sufficient coverage at each position. The VarScan would first call germline variants, SNPs and indels by a heuristic and statistical comparing method. Then, the genotypes are compared with the reference sequences by the algorithm shown in Figure 6. At a position, if the tumour and the normal read do not match, VarScan did Fisher's Exact Test to calculate the significance of allele frequency difference and use the p-value to represent this significance. If the difference is significant, we call this read somatic (Koboldt et al., 2009). For our study, we will focus on somatic variants.

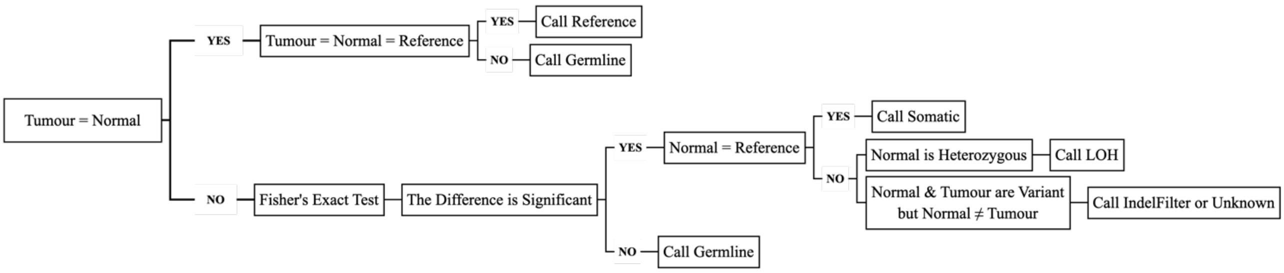

Figure 6. The algorithm when VarScan compares genotypes in somatic mutation calling.

The parameters of interest are ‘minimum total reads’ and ‘minimum variant allele frequency’ in VarScan. Within the VarScan, the parameters control the strictness of filters and the level of noise (false positives, false negatives). The output sequences will be aligned and sent as the input of BEAST2 to do Bayesian analysis and generate the posterior tree distribution. Changing the variant calling parameters, we can eventually obtain different posterior tree distributions and visually inspect them through DensiTree by generating a cladogram made of all sets of trees and consensus trees. Then, we will observe how the tree branches, model parameters (including error and mutation rate parameters), and tree topologies change as we vary variant caller parameters. At last, we will have more understanding towards the variants and noise in tumour phylogeny.

3.3 Expected Findings

1) Obtain a lower entropy.

- Find the potential rogue taxa and remove them to obtain a lower entropy.
- Find out the lowest entropy with the removed rogue taxa.
- Quantify how much noise we have removed.

2) Get a clearer tumour phylogeny.

- Remove rogue taxa and obtain a clearer cladogram, including consensus trees and all sets of trees.
- We are able to obtain more effective signals from the cladogram.

3) Learn more about the noise.

- From the rogue taxa detection and removal, the role of rogue taxa generating uncertainties in tumour phylogeny will be observed and summarised.

4) Discover how the targeted parameters contribute to the trees.

- After the test, changing the variant caller parameters will systematically evaluate and summarise how they affect the branch length of trees, parameters of models, and topologies of trees.

Section 4: Limitations and Implications of the proposed research

4.1 Potential limitations of the proposed research

In the process of obtaining clearer posterior trees, the Entropy-based methods are novel. Höhna and Drummond (2011) have discussed CCP, while Larget (2013) has shown the strong power of CCD, we can learn that it is a useful tool for learning the posterior distribution better. Papers also discussed detecting rogue taxa in the way of generating best-supported subtrees (Cranston & Rannala, 2007), and others used entropy to represent the difference between two sets of information (Lewis et al., 2016). However, there is no existing research using combined all these methods to obtain a clear phylogeny. Therefore, it is novel and challenging.

For the variant caller, it is hard to balance the power of filtering and the clarity of output sequences. When we set the parameters with a strict condition, a lot of positions would be filtered out, but this set of data can generate a clear posterior tree distribution. We have no ground truth for the trees in real datasets, so the output data cannot say anything but only our guesses. At the same time, though it is a useful tool to call somatic mutations with low frequencies, it also reduces the overall false-positive rate (Koboldt, 2020), and leaves many false in the output. Therefore, everything should be verified through experiments in the future if we want to say something through the output of the variant caller.

This proposed study is complex. We need to test and use a fitted growth model in BEAST2 while testing the parameters in the variant caller, which is time-consuming. Furthermore, a high demand for calculation appears in the Entropy-based methods. At the same time, the overall study has many complicated steps.

4.2 Implications of the proposed research

With the Entropy-based methods, we improve our existing Bayesian inference pipeline by removing rogue taxa in the posterior tree distributions or in the input sequences. Uncertainties can be reduced, and we will have more understanding of them. At the same time, we are testing the Entropy-based methods to provide a novel method with real data and will offer a critical review of the new method for scientists in future research.

Through this proposed study, we mainly focus on the noise in tumour phylogeny. With the methods, we detect and remove the rogue taxa and play with the variant calling parameters. In this process, we learn the noise better. At the same time, we aim to find efficient ways to reduce noise, from which we can figure out some uncertainties in tumour phylogeny. At last, a clear tumour phylogeny allows us to have a better understanding of tumour in general and thus, in the long run, be helpful for clinical treatments.

References

Aberer, A. J., Krompass, D., & Stamatakis, A. (2012). Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology, 62*(1), 162-166. doi:10.1093/sysbio/sys078

Andor, N., Lau, B. T., Catalanotti, C., Kumar, V., Sathe, A., Belhocine, K., . . . Ji, H. P. (2020). Joint single cell DNA-Seq and RNA-Seq of cancer reveals subclonal signatures of genomic instability and gene expression. *bioRxiv*, 445932. doi:10.1101/445932

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An Introduction to MCMC for Machine Learning. *Machine Learning, 50*(1), 5-43. doi:10.1023/A:1020281327116

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., . . . Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology, 15*(4), e1006650-e1006650. doi:10.1371/journal.pcbi.1006650

Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics, 26*(10), 1372-1373. doi:10.1093/bioinformatics/btq110

Chen, K., Welch, D., & Drummond, A. J. (2021). Ignoring errors causes inaccurate timing of single-cell phylogenies. *bioRxiv*, 2021.2003.2017.435906. doi:10.1101/2021.03.17.435906

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association, 91*(434), 883-904. doi:10.2307/2291683

Cranston, K. A., & Rannala, B. (2007). Summarizing a Posterior Distribution of Trees Using Agreement Subtrees. *Systematic Biology, 56*(4), 578-590. doi:10.1080/10635150701485091

David, B., Takuto, S., Kristian, C., Gad, G., Chip, S., & Lee, L. (2019). Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*, 861054. doi:10.1101/861054

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., & Vijg, J. (2017). Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods, 14*(5), 491-493. doi:10.1038/nmeth.4227

Donmez, N., Malikic, S., Wyatt, A. W., Gleave, M. E., Collins, C. C., & Sahinalp, S. C. (2016). *Clonality Inference from Single Tumor Samples Using Low Coverage Sequence Data*, Cham.

Fisher, R., Pusztai, L., & Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer, 108*(3), 479-485. doi:10.1038/bjc.2012.581

Gavryushkin, A., & Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *Journal of theoretical biology, 403*, 197-208. doi:10.1016/j.jtbi.2016.05.001

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics, 17*(3), 175-188. doi:10.1038/nrg.2015.16

Gay, L., Baker, A. M., & Graham, T. A. (2016). Tumour Cell Heterogeneity. *F1000Res, 5*. doi:10.12688/f1000research.7210.1

Gelman, A., Lee, D., & Guo, J. (2015). Stan:A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics, 40*(5), 530-543. doi:10.3102/1076998615606113

Guo, S., Zhu, X., Huang, Z., Wei, C., Yu, J., Zhang, L., . . . Li, Z. (2023). Genomic instability drives tumorigenesis and metastasis and its implications for cancer therapy. *Biomedicine & Pharmacotherapy, 157*, 114036. doi:https://doi.org/10.1016/j.biopha.2022.114036

Gupta, M., Chandan, K., & Sarwat, M. (2022). Natural products and their derivatives as immune check point inhibitors: Targeting cytokine/chemokine signalling in cancer. *Seminars in Cancer Biology, 86*, 214-232. doi:https://doi.org/10.1016/j.semcancer.2022.06.009

Hegenbarth, J.-C., Lezzoche, G., De Windt, L. J., & Stoll, M. (2022). Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. *Frontiers in Molecular Medicine, 2*. doi:10.3389/fmmed.2022.839338

Höhna, S., & Drummond, A. J. (2011). Guided Tree Topology Proposals for Bayesian Phylogenetic Inference. *Systematic Biology, 61*(1), 1-11. doi:10.1093/sysbio/syr074

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., . . . Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology, 65*(4), 726-736. doi:10.1093/sysbio/syw021

Huang, H. W., Mullikin, J. C., & Hansen, N. F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics, 16*, 235. doi:10.1186/s12859-015-0624-y

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., & Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics, 15*(1), 35. doi:10.1186/1471-2105-15-35

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine, 12*(1), 91. doi:10.1186/s13073-020-00791-w

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., . . . Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics, 25*(17), 2283-2285. doi:10.1093/bioinformatics/btp373

Larget, B. (2013). The Estimation of Tree Posterior Probabilities Using Conditional Clade Probability Distributions. *Systematic Biology, 62*(4), 501-511. doi:10.1093/sysbio/syt014

Lewis, P. O., Chen, M.-H., Kuo, L., Lewis, L. A., Fučíková, K., Neupane, S., . . . Shi, D. (2016). Estimating Bayesian Phylogenetic Information Content. *Systematic Biology, 65*(6), 1009-1023. doi:10.1093/sysbio/syw042

Liu, B., Duenas, D., Zheng, L., Reckamp, K., & Shen, B. (2022). Genomic instability as a major mechanism for acquired resistance to EGFR tyrosine kinase inhibitors in cancer. *Protein & Cell, 13*(2), 82-89. doi:10.1007/s13238-021-00855-6

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi:10.1023/A:1008929526011

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C., & Beerenwinkel, N. (2019). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature Communications, 10*(1), 2750. doi:10.1038/s41467-019-10737-5

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. *3rd International Workshop on Distributed Statistical Computing (DSC 2003); Vienna, Austria, 124*.

Ramachandran, K. M., & Tsokos, C. P. (2021). Chapter 13 - Empirical methods. In K. M. Ramachandran & C. P. Tsokos (Eds.), *Mathematical Statistics with Applications in R (Third Edition)* (pp. 531-568): Academic Press.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., . . . Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol, 61*(3), 539-542. doi:10.1093/sysbio/sys029

Stratton, M. R. (2011). Exploring the Genomes of Cancer Cells: Progress and Promise. *Science, 331*(6024), 1553-1558. doi:doi:10.1126/science.1204040

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution, 4*(1). doi:10.1093/ve/vey016

Valecha, M., & Posada, D. (2022). Somatic variant calling from single-cell DNA sequencing data. *Comput Struct Biotechnol J, 20*, 2978-2985. doi:10.1016/j.csbj.2022.06.013

Westover, K. M., Rusinko, J. P., Hoin, J., & Neal, M. (2013). Rogue taxa phenomenon: A biological companion to simulation analysis. *Molecular Phylogenetics and Evolution, 69*(1), 1-3. doi:https://doi.org/10.1016/j.ympev.2013.05.010

Wilkinson, M., Thorley, J. L., & Upchurch, P. (2000). A chain is no stronger than its weakest link: double decay analysis of phylogenetic hypotheses. *Syst Biol, 49*(4), 754-776. doi:10.1080/106351500750049815

Wu, H., Zhang, X. Y., Hu, Z., Hou, Q., Zhang, H., Li, Y., . . . Wu, S. (2017). Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. *Oncogene, 36*(20), 2857-2867. doi:10.1038/onc.2016.438