CANCER PHYLOGENETICS USING SINGLE-CELL RNA-SEQ DATA

JIŘÍ C. MORAVEC[1], ROB LANFEAR[2], DAVID L. SPECTOR[3], SARAH D. DIERMEIER[4], AND ALEX GAVRYUSHKIN[1, ✉]

Abstract. Phylogenetic methods are emerging as an useful tool to understand cancer evolutionary dynamics, including tumor structure, heterogeneity, and progression. Most currently used approaches utilize either bulk whole genome sequencing (WGS) or single-cell DNA sequencing (scDNA-seq) and are based on calling copy number alterations and single nucleotide variants (SNVs). Here we explore the potential of single-cell RNA sequencing (scRNA-seq) to reconstruct cancer evolutionary dynamics. scRNA-seq is commonly applied to explore differential gene expression of cancer cells throughout tumor progression. The method exacerbates the single-cell sequencing problem of low yield per cell with uneven expression levels. This accounts for low and uneven sequencing coverage and makes SNV detection and phylogenetic analysis challenging. In this paper, we demonstrate for the first time that scRNA-seq data contains sufficient evolutionary signal and can be utilized in phylogenetic analyses. We explore and compare results of such analyses based on both expression levels and SNVs called from our scRNA-seq data. Both techniques are shown to be useful for reconstructing phylogenetic relationships between cells, reflecting the clonal composition of a tumor. Without an explicit error model, standardized expression values appears to be more powerful and informative than the SNV values at a lower computational cost, due to being a by-product of standard expression analysis. Our results suggest that scRNA-seq can be a competitive alternative or useful addition to conventional scDNA-seq phylogenetic reconstruction. Our results open up a new direction of somatic phylogenetics based on scRNA-seq data. Further research is required to refine and improve these approaches to capture the full picture of somatic evolutionary dynamics in cancer.

Introduction

Phylogenetic analysis is an approach that relies on reconstructing evolutionary relationships between organisms to determine population genetics parameters such as population growth (Kingman 1982; Heled et al. 2008), structure (Müller et al. 2017a) or geographical distribution (Lemey et al. 2009; Lemey et al. 2010). Typically, the reconstructed phylogeny is not the end-goal. Using previously estimated trees, various evolutionary hypotheses can be explored, such as the evolutionary relationship of traits carried by individual taxa (Grafen et al. 1989; Pagel et al. 2004; Freckleton 2012).

Within-organism cancer evolution is increasingly being studied using population genetics approaches, including phylogenetics (Navin et al. 2011; Yuan et al. 2015; Alves et al. 2017; Schwartz et al. 2017; Caravagna et al. 2018; Singer et al. 2018; Alves et al. 2019; Caravagna et al. 2019; Detering et al. 2019; Malikic et al. 2019; Werner et al. 2019; Kuipers et al. 2020), to understand molecular dynamics of cancer cell populatons. These approaches have shown promise to be developed into therapeutic applications in the personalized medicine framework (Gerlinger et al. 2012; Abbosh et al. 2017; Rao et al. 2020). Specifically, the clonal composition of tumors, metastasis initiation, development, and timing can be reconstructed using phylogenetic methods (Yuan et al. 2015;

[1] Department of Computer Science, University of Otago, New Zealand
[2] Ecology, Evolution, and Genetics, The Australian National University, Australia
[3] Cold Spring Harbor Laboratory, New York, United States of America
[4] Department of Biochemistry, University of Otago, New Zealand
E-mail addresses: jiri.moravec@otago.ac.nz, rob.lanfear@anu.edu.au, spector@cshl.edu, sarah.diermeier@otago.ac.nz, ✉alex@biods.org.

Angelova et al. 2018; El-Kebir et al. 2018; Alves et al. 2019). Unlike other evolutionary processes prone to events such as hybridization or horizontal gene transfer, population dynamics of somatic cells is underpinned by a strictly bifurcating clonal process driven by cell division. This is in perfect agreement with theoretical assumptions routinely applied in stochastic phylogenetic models such as coalescent (Kingman 1982; Hudson et al. 1990; Posada 2020) or birth-death processes (Aldous 1996; Aldous 2001; Komarova 2006).

From the methodological perspective, however, cancer is an evolutionary process with unique characteristics which are not modeled in conventional phylogenetic approaches. These include a high level of genomic instability with structural changes (gene losses and duplications) which accumulate along with point mutations during the course of growth and evolution (Beerenwinkel et al. 2015; Posada 2015).

Traditional Whole Genome Sequencing (WGS) methods have been instrumental in understanding cancer mutational profiles and oncogene detection (Mardis et al. 2009; Nakagawa et al. 2018). DNA from a tissue sample is isolated and sequenced "in bulk". This increases the total amount of DNA which improves coverage and reduces amplification errors. To establish the presence or absence of mutations, a variant allele frequency (VAF) is calculated and compared to a threshold, typically $10 - 20\%$ (Strom 2016). This filters out rare mutations present only in a few reads that are likely to be false positives or sequencing errors (Petrackova et al. 2019). More recently, bulk sequencing is used to study cancer evolution using phylogenetic methods (Zhai et al. 2017; Zhao et al. 2016; Ling et al. 2015). However, the usage of bulk sequencing in this context is problematic. Bulk samples contain cells from multiple cell lineages including non-tumor cells, such as immune or blood vessel cells (Racle et al. 2017), and strong evidence also points out a constant migration of metastatic cells between tumors (Aguirre-Ghiso 2010; Cheung et al. 2016; Reiter et al. 2017; Casasent et al. 2018). High VAF thresholds ignore tumor heterogenity, but by lowering the threshold, mutations in non-tumor cells or clonal lineages are retained instead. Sequences or mutational profiles derived from bulk samples thus have a chimeric origin (Alves et al. 2017).

A typical assumption in classical phylogenetics is that the sequences or mutational profiles represent individual taxonomic units, either individuals or populations of closely related individuals. If these methods are used on the data from bulk samples, the reconstructed trees are not phylogenies describing an evolutionary history, but evolutionary meaningless sample similarity trees (Alves et al. 2017). To address this issue, phylogenetic trees are reconstructed by estimating the sequential order of somatic mutations using VAF from one or multiple tumor samples (Deshwar et al. 2015; El-Kebir et al. 2018; Miura et al. 2018). Given the tumor heterogenity and insufficient read depth to reliably estimate VAF, this is not a simple problem and the performance of current methods is limited (Miura et al. 2020).

Single-cell DNA sequencing (scDNA-seq) does not suffer from the chimeric DNA origin of bulk-sequencing as each DNA segment is barcoded to guarantee its known cell of origin. Recent progress in WGS technology made sequencing individual cells cost-efficient (Gawad et al. 2016) and this approach is now regularly used for the phylogenetic reconstruction of metastatic cancer or the subclonal structure of a single tumor (Potter et al. 2013; Roth et al. 2016; Leung et al. 2017; Myers et al. 2019). However, this increased resolution comes with additional complications. Current methods are not sensitive enough to sequence DNA from a single cell and DNA amplification is required (Gawad et al. 2016). This process suffers from a random bias with different parts of the genome amplified in different quantities or not at all (Satas et al. 2018). In addition, polymerase does not replicate DNA without error, this can have a significant impact if the replication errors occur early in the amplification process (Gawad et al. 2016). This does not only increase the error rate for identified SNVs, but a large proportion of SNVs might be simply missing (Hicks et al. 2018). The advantages associated with scDNA-seq led to the development of novel approaches that tackle these challenges using an error model to correct for amplification errors and false-positive SNV calls (Zafar et al. 2016; Zafar et al. 2018; Luquette et al. 2019; Kozlov et al. 2020).

Similar technological development led to proliferation of single-cell RNA sequencing (scRNA-seq) which, compared to traditional bulk RNA sequencing, enabled detection of gene expression profiles for individual cells in the tissue sample (Müller et al. 2017b; Olsen et al. 2018; Jerby-Arnon et al. 2018; González-Silva et al. 2020). This allows to understand tumor heterogenity by identifying different cell populations (Andrews et al. 2018), estimating immune cell content within a tumor (Yu et al. 2019), or even identify individual clones and subclones, as they can differ in their behavior (Fan et al. 2020). However, as the levels of RNA expression vary between genes and cells, the amplification problems of scDNA-seq that cause unequal expression and drop-out effect are more pronounced in scRNA-seq. There is an increased interest for SNV calling on scRNA-seq data using bulk-SNV callers (Chen et al. 2016; Poirion et al. 2018; Liu et al. 2019; Schnepp et al. 2019) as this allows for identification of mutations in actively expressed genes.

In this work, we test if scRNA-seq contains phylogenetic information to reconstruct a population history of cancer. We perform an experiment to guarantee a known population history, and then try to reconstruct this history using computational phylogenetics from both expression values and identified SNVs derived from the same scRNA-seq data set. We then compare phylogenies obtained from these methods against the known population history to evaluate the strength of the phylogenetic signal contained in the scRNA-seq data set.

Methods

Sample preparation and scRNA sequencing. MDA-MB-231-LM2 (GFP+) (Minn et al. 2005) cells were injected into the R4 mammary fat pad of Nu/J mice ($250,000$ cells per mouse, 3 mice), and tumor growth was monitored for 8 weeks. Mice were euthanized when tumor size approached the endpoint ($2\,cm$). Tumors were resected and dissociated into single cells. To extract circulating tumor cells (CTC), up to $1\,ml$ of blood was drawn immediately post euthanasia using cardiac puncture. Red blood cells were removed using RBC lysis buffer. All cells (tumor derived and circulating tumor cells) were stained with DAPI and sorted for DAPI and GFP using a BD FACSAria cell sorter. Libraries were generated using the 10x Chromium single cell gene expression system immediately after cell sorting, and sequenced on an Illumina NextSeq platform.

Mapping and expression analysis. Reads were mapped with the Cellranger v5.0 software to the GRCh38 v15 from the Genome Reference Consortium using the analysis-ready assembly without alternative locus scaffolds (no_alt_analysis_set) and associated GTF annotation file.

The Cellranger software performs mapping, demultiplexing, cell detection, and gene quantification for the 10x Genomics scRNA-seq data.

Postprocessing expression data.

Standardizing expression values. The filtered feature-barcode expression values from Cellranger were processed using the R Seurat v3.2.0 package (Stuart et al. 2019) and according to the Seurat's standard pre-processing workflow. However, low-quality cells, such as cells with small number of unique reads or small number of represented genes, were not removed at this stage and no normalization was performed. The expression values for each gene were centered ($\mu = 0$) and rescaled ($\sigma^2 = 1$).

Discretizing expression values. The rescaled expression values were then categorized into a 5 level ordinal scale ranging from 1 (low level of expression) to 5 (high level of expression). The five-level scale was chosen to capture the data distribution of the rescaled expression values and represent a compromise between introducing data noise with too many levels or artificial similarity with only a few categories.

Interval ranges, according to which the values were categorized, were chosen according to the 60% and 90% High Density Intervals (HDI), the shortest intervals containing 60% or 90% of values respectively. The values inside the 60% HDI were categorized as normal, values inside the 90%

HDI, but outside the 60% as increased/decreased expression and values outside the 90% HDI as a extremely increased/decreased expression.

Genes that contain only a single categorized value for each cell were removed as phylogenetically irrelevant and the discretized values were then transformed into a fasta format.

SNV.

Pre-processing reads for SNV detection. The BAM file from Cellranger was processed using the Broad Institute's Genome Analysis ToolKit (GATK) v4.1.7.0 (Poplin et al. 2018) according to GATK best practices of somatic short variant discovery.

SNV detection and filtering. To obtain SNVs for individual cells of the scRNA-seq data, first a list of SNVs were obtained by running the Mutect2 (Benjamin et al. 2019), treating the data set as a pseudo-bulk sample, and retaining only the SNVs that passed all filters.

SNVs for individual cells were then obtained by individually summarizing reads belonging to each single cell at the positions of the SNVs obtained beforehand using the pysam library, which is built on htslib (Li et al. 2009). The most common base for every cell and every position was retained, base heterogeneity or copy number variation was ignored. This SNV table was then transformed into a fasta format.

Phylogenetic analysis. To reconstruct phylogenetic trees from the categorized expression values and identified SNVs, we have used the IQ-TREE v2.0.3 (Minh et al. 2020) and BEAST2 v2.6.3 (Bouckaert et al. 2019).

For the expression data, IQ-TREE was used with an ordinal model and an ascertainment bias correction (`-m ORDINAL+ASC`). For the SNV values, we used a generalized time reversible model (GTR) (Tavaré 1986) as the most complex substitution model present in both IQ-TREE and BEAST2. Tree support was evaluated using the ultrafast bootstrap (UFBoot2) (Minh et al. 2013; Hoang et al. 2018) with 1000 permutations (`-B 1000`). While the ultrafast bootstrap might not converge given a large amount of missing data, it is faster by several orders of magnitude than a standard parametric bootstrap.

The BEAST2 analysis was performed with a birth-death tree prior (Kingman 1982) and an exponential population growth (Kuhner et al. 1998) as these models most closely mimic the biological conditions of tumor growth. The substitution model of choice for the categorized expression values was the ordinal model available in the Morph-Models package and the GTR model for the SNVs as described above.

Phylogenetic clustering tests. To test if the phylogenetic methods were able to recover the true population history, we employ Mean Pairwise Distance (MPD) (Webb 2000) and Mean Nearest Taxon Distance (MNTD) (Webb 2000) to investigate the relationship between cells from each sample. MPD is calculated as a mean distance between each pair of taxa from the same group, while MNTD is calculated as a mean distance to the nearest taxon from the same group. For each sample and samples isolated from a single individual, MPD and MNTD are calculated and compared to a null distribution obtained by permutating sample labels on a tree and calculating MPD and MNTD for these permutations. P-value is then calculated as a rank of MPD/MNTD in the null distribution normalized by the number of permutations. The MPD and MNTD are calculated using the `ses.mpd` and `ses.mntd` functions implemented in the package `picante` (Kembel et al. 2010) For the Bayesian phylogenies, MPD and MNTD were calculated for a sample of 1000 trees from the posterior distribution and then summarized with mean and 95% confidence interval.

Code and data availability. Code required to replicate the data processing steps is available at `https://github.com/bioDS/phyloRNAanalysis`.

To aid in creating pipelines for phylogenetic analysis of scRNA-seq data, we have integrated a number of common tools in the R `phyloRNA` package, which is available at `https://github.com/bioDS/phyloRNA`.

All data will become available in the NCBI GEO under the accession number GSE163210 upon acceptance of this paper by a peer-reviewed journal.

Results

Phylogenetic expectation derived from experimental design. To test if scRNA-seq contains sufficient phylogenetic information to reconstruct a population history of cancer, immunosuppressed mice were injected into the mammary fat pad with human breast cancer cells. The tumors that develop are derived from the same population and thus share a common ancestor, but evolved independently in each mouse and should form separate clades on reconstructed phylogenetic trees when analysed together. As each tumor was seeded by a population of cancer cells, number of smaller sample-specific clades representing subclonal diversity of the population sample should be observed. We would thus expect clustering of each tumor and CTC sample as well as clustering of tumor and CTC samples isolated from single individual. Due to the lack of a specialized scRNA-seq caller or error model to account for the uncertainty in the data, some intermixing is possible, but heavy intermixing would demonstrate an insufficiency of scRNA-seq for phylogenetic analyses.

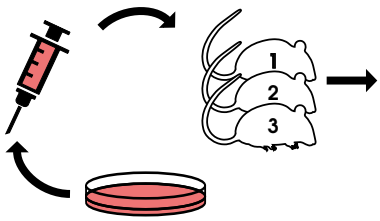
Sample overview. In total, five samples were used in this analysis, three tumor samples (T1, T2, T3) and two CTC samples (CTC1, CTC2). The number of cells isolated from the CTC3 sample was too small for scRNA sequencing and the sample was removed from the study. The number of detected cells in the tumor samples was generally smaller than in the CTC samples, but the reverse was true for the total number of detected unique molecular identifiers (UMIs) – the number of unique mRNA transcripts (see Table 1). In the T2 sample, a large number of cells but a small number of UMIs were detected in a similar pattern to the CTC samples.

Compared to the fluorescent-activated cell sorting (FASC), Cellranger detected fewer cells for tumor samples, but more cells for the CTC samples. Cellranger classifies barcodes as cells based on the amount of UMI detected to distinguish real cells from a background noise (Lun et al. 2019). The large number of detected cells in the CTC samples is likely a result of lysed cells or cell-free RNA (Fleming et al. 2019). In all cases, the number of detected genes across data sets was relatively low, with the best sample T3 amounting to about 3% of expressed genes when summarized across all cells.

Recording unexpressed genes as unknown data. The amount of coverage in a standard bulk RNA-seq expression analysis is usually sufficient to conclude that genes for which no molecule was detected are not expressed (Lähnemann et al. 2020). In scRNA-seq however, the sequencing coverage is very small, drop out effect is likely, and thus this assumption does not hold. According to the standard expression pipeline, these values are commonly treated as real zeros, i.e., no expression and have a significant impact on the data distribution during the normalization and rescaling steps (Hicks et al. 2018). Without an explicit model of drop out effect to account for technical zeros, these values are more accurately represented as unknown values rather than true biological zeros (Van den Berge et al. 2018). We have modified the Seurat code to treat these values as unknown values (`NA` in R) and included modified functions in the `phyloRNA` package.

We will further use *data density* to describe the number of unknown values in both expression and SNV datasets, with 100% representing dataset without unknown values. For example, the T3 sample has 1% data density after recoding zeros as unknown values.

TABLE 1. An overview of data set used in this work. In total, five samples were isolated from three individuals (Table 1a): 3 tumor samples (T1, T2, T3) and 2 circulating tumor cell samples (CTC1, CTC2). For each sample, the number of cells from fluorescent-activated cell sorting (FASC), the number of identified by Cellranger, the number of detected genes, the number of unique molecular identifiers (UMIs), UMI/Cell ratio, and the data density are reported (Table 1b).



	Tumor	Circulating tumor cells
mouse 1	T1	CTC1
mouse 2	T2	CTC2
mouse 3	T3	N/A

(a) Experiment overview

Sample	Cells (FASC)	Cells (Cellranger)	Genes	UMI	UMI/Cell	data density
T1	11,258	701	3,428k	17k	4518	2.97 %
T2	20,233	2,794	69k	5k	25	0.04 %
T3	13,865	806	2,876k	18k	3569	2.57 %
CTC1	605	3,125	129k	8k	41	0.06 %
CTC2	415	3,161	155k	9k	49	0.06 %
total:	46,376	10,587	6.4M	20k	604	0.44 %

(b) Sample overview

SNV identification and data density. To identify SNVs in scRNA-seq data, we first identified a list of SNVs by treating the single-cell reads as a pseudo-bulk sample. The total of 120,310 SNVs that passed all quality filters were identified this way. When these SNVs were called for each individual cell, the resulting data set had data density of less than 0.09%. The expression data is expected to have higher data density than SNV because for expression quantification a presence or absence of a molecule is sufficient while for SNV, knowledge of each position is required. This expectation is confirmed in Table 1, where data density of the expression data is summarized. About 16% of the 10,587 cells represented in this data set did not contain any known SNV (data density 0%), these were relatively equally distributed among the T2 (651), CTC1 (571) and CTC2 (475) samples. This represents a challenge from a data analysis perspective given the large sample size and its small data density.

Finding the most well-represented subset of data. When treating the potentially unexpressed genes as unknown values, only a small proportion of the expression count values was known, with the data set derived from SNV suffering from the same problem due to the low number of reads for each cell.

While model-based phylogenetic methods can process missing data by treating the missing data as phylogenetically neutral, this significantly flattens the likelihood space which can cause artefacts, convergence problem or increase computational time (Wiens 2006; Jiang et al. 2014; Xi et al. 2016). It is however not only the proportion of the unknown values, which is about 99% for both data sets, but also the sheer size of the data set that is problematic. With over $10,000$ cells and more than $54,000$ genes or over $120,000$ SNVs, the unfiltered data set would require substantial computational resources.

In comparison, for the published phylogenetic tools designed for single-cell DNA, data sets ranged from 47 cells and 40 SNVs (Jahn et al. 2016) to 370 cells and 50 SNVs (Singer et al. 2018) or in an extreme case 18 cells and $50,000$ SNVs (Singer et al. 2018) with at most 58% of missing data across these data sets.

To help alleviate these issues, we have employed a stepwise filtering algorithm to find the densest subset of a data set. By iteratively cutting out cells and genes/SNVs with the smallest number of known values, we increase the data density until a local maximum or desired density is reached. The advantage of this method is that a desired density can be reached with the least amount of data removed.

To test the effect of unknown data, the categorized expression values were filtered using the above method to get categorized expression data sets containing 20%, 50%, and 90% of known values.

Filtering scRNA-seq expression and SNV data sets. The expression data set filtered to 20% density contained 1,627 cells and 16,187 genes. Cells were mainly represented by T1 and T3 samples which form over 92% of the data set. In contrast, other samples (T2, CTC1, CTC2) were significantly underrepresented despite their larger amount of cells in an unfiltered data set. In filtering to 50% density, the numbers decreased to 1,454 cells and 1,634 genes. The sample diversity also decreased, with T2 dropping out entirely. When filtered to 90% density, the data set was reduced to 593 cells and 528 genes. The data diversity further decreased to T1, T3 and CTC2, although the CTC2 sample was reduced to 2 cells.

When the SNV data set was filtered to 20% density, the numbers decreased significantly – to 1,498 cells and 1,297 SNVs. The sample composition was similar to the expression data set, with the vast majority of retained cells belonging to the T1 and T3 samples. However, the T2 sample was already missing. In subsequent filtering to both 50% and 90% data density, the CTC1 sample has vanished.

Despite the difference in starting dimension between expression and SNV data sets, the dimension of data set filtered to the same data density is similar, with the expression data having higher density than SNVs. This suggests that in both cases, known values are concentrated at the same cell subsets and the filtering algorithm can successfully localize them.

Maximum likelihood phylogeny from expression data. We inferred Maximum likelihood trees of the expression data filtered to 20%, 50%, and 90% data density (see Supplementary materials). In the reconstructed phylogeny at the 20% density filtering (Figure 1), individual tumor samples did not form three separate clades, but a large number of smaller, well-supported clades. These clades are distributed along the central spine of the unrooted maximum likelihood tree and have little internal structure. The T2, CTC1, and CTC2 samples form relatively compact clades, while the more represented T1 and T3 clades are generally intermixed. The MNT and MNTD test confirm this (Table 2), with T2, CTC1, and CTC2 showing significant clustering signal.

When the data is filtered to 50% and 90% data density (see Supplementary materials), the intermixing between T1 and T3 is reduced, with many clades showing a sample-specific pattern in their internal structure, and statistically significant MPD and MNTD tests support this for T3, but not T1 sample. As data are filtered, CTC1 still forms a compact cluster, but the CTC2 cluster disappears as the cells are being removed. The phylogenetic position of the remaining CTC2 cells is however stable as sister cells to the T2 sample. Due to a small number of cells for T2, CTC1, and CTC2, the phylogenetic clustering tests for these samples are no longer significant.

Bootstrap branch support scores are estimates of topology uncertainty for each branch of a tree, with 0 signifying no support while 100 high support. Given a large number of branches in our trees, displaying them directly on a tree is challenging. For this reason, we show a mean and a standard deviation of bootstrap scores for each reconstructed Maximum likelihood tree (Table 3).

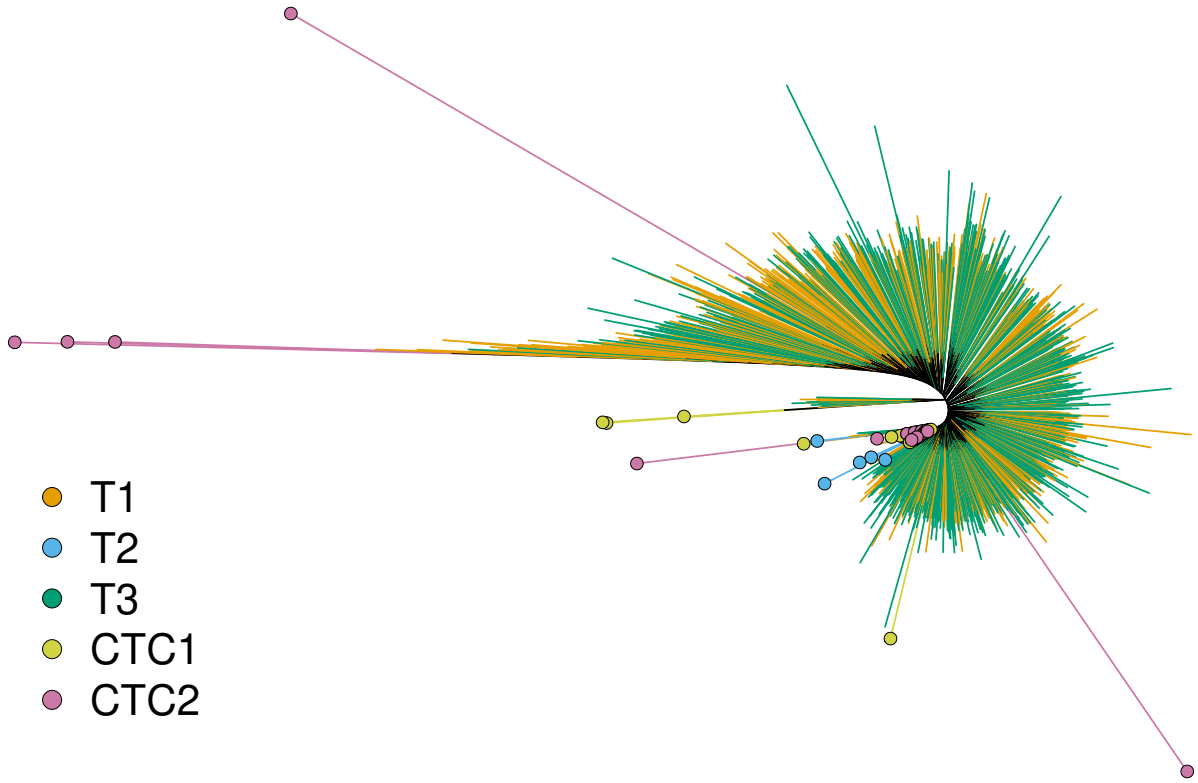


FIGURE 1. Maximum likelihood tree from the expression data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). For additional clarity, less represented samples are marked with colored circles.

TABLE 2. Test of phylogenetic clustering on the Maximum Likelihood tree from the expression data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated on the Maximum Likelihood trees from the expression data for 20%, 50%, and 90% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2) and expected clustering for cells isolated from a single individual (T1 with CTC1, and T2 with CTC2). Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

	20% data density			50% data density			90% data density		
Groups	Cells	MPD	MNTD	Cells	MPD	MNTD	Cells	MPD	MNTD
T1	701	1.000	1.000	688	1.000	1.000	329	0.987	0.152
T2	11	*0.001	*0.001	0	–	–	0	–	–
T3	806	0.157	1.000	758	*0.001	*0.001	262	*0.003	*0.007
CTC1	58	*0.001	*0.001	3	0.126	0.088	0	–	–
CTC2	51	*0.001	*0.001	5	1.000	1.000	2	0.998	1.000
T1 & CTC1	759	0.995	0.999	691	1.000	1.000	329	0.987	0.152
T2 & CTC2	62	*0.001	*0.001	5	1.000	1.000	2	0.998	1.000

* significant support

TABLE 3. Mean and standard deviation for the ultrafast bootstrap support values for the Maximum likelihood trees from the expression and SNVs data for 20%, 50%, and 90% data density.

	Data density		
	20%	50%	90%
Expression	73 ± 23	69 ± 24	78 ± 20
SNV	71 ± 22	69 ± 23	69 ± 20

Most branches (77%) were not statistically supported (UFBoot2 < 95) with 20% branches had very weak score (UFBoot2 < 50). Only 22% of branches were statistically supported by the ultrafast bootstrap (UFBoot2 ≤ 95). These were usually branches close to the tips of the tree.

However, the ultrafast bootstrap failed to converge for each filtered dataset.

Maximum likelihood phylogeny from the identified SNVs. Similarly to the phylogeny reconstructed from the expression data, the tree reconstructed from the SNV data (Figure 2) consisted of a large number of well-supported clades. In contrast to the expression phylogeny, the clades in the SNV trees showed significantly more internal structure.

While T1 and T3 samples were still intermixed, sample-specific clades are present and this clustering tendency was confirmed by phylogenetic clustering tests (Table 4). The clustering pattern for T1 is different from the clustering pattern of T3 as only MNTD was significant for T1, and only MPD was significant for T3. The CTC1 cells formed a compact clade but placed on a very long branch, suggesting a long independent evolution of the CTC1 cells. CTC2 samples did not form a clade but were placed as sister cells to the T1 sample. The general structure of the phylogeny did not change when the data was further filtered to 50% and 90% data density (see Supplementary materials), except for the clustering pattern of T1, which was no longer significant.

The branch support values for the SNV tree shared a similar pattern to the tree from the expression data (Table 3), with a small drop in an overall support as only 19% branches were supported (UFBoot2 ≤ 95).

Only the smallest data set, SNV filtered to 90% data density, converged in the ultrafast bootstrap analysis. However, simply increasing the number of iterations is not an efficient solution as the analysis already took a substantial amount of computational resources (over 100 CPU hours). An alternative data reduction strategy that would preserve sample diversity is required.

Alternative filtering approach. The filtering used above was designed to remove the least amount of data possible. While this is generally a positive behavior, it might be problematic if the amount of missing data varies between individual samples as we are dependent on the presence of all samples for the assessment of the performance of the phylogenetic reconstruction.

To ameliorate this, we employ an alternative filtering strategy and select cells with the least amount of missing data as the best representation of each data set. In this reduced data set, genes that were not present in any of the cells or present only in a single cell, are removed. Data set is then filtered to a desired data density using the method described above, but to retain the same sample size, only genes were removed this way. Given the smaller size, the full data set, and data sets filtered to 50% and 90% of data density were then analyzed using Maximum Likelihood and Bayesian method to further explore the topological uncertainty.

Total of 58 cells were retained: 20 cells for T1 and T3 samples and six cells for T2, CTC1, and CTC2 samples.

The full expression data set contained 30% of known data distributed across 7,520 genes. This was further reduced to 3,261 and 219 genes when filtered to 50% and 90% data density.

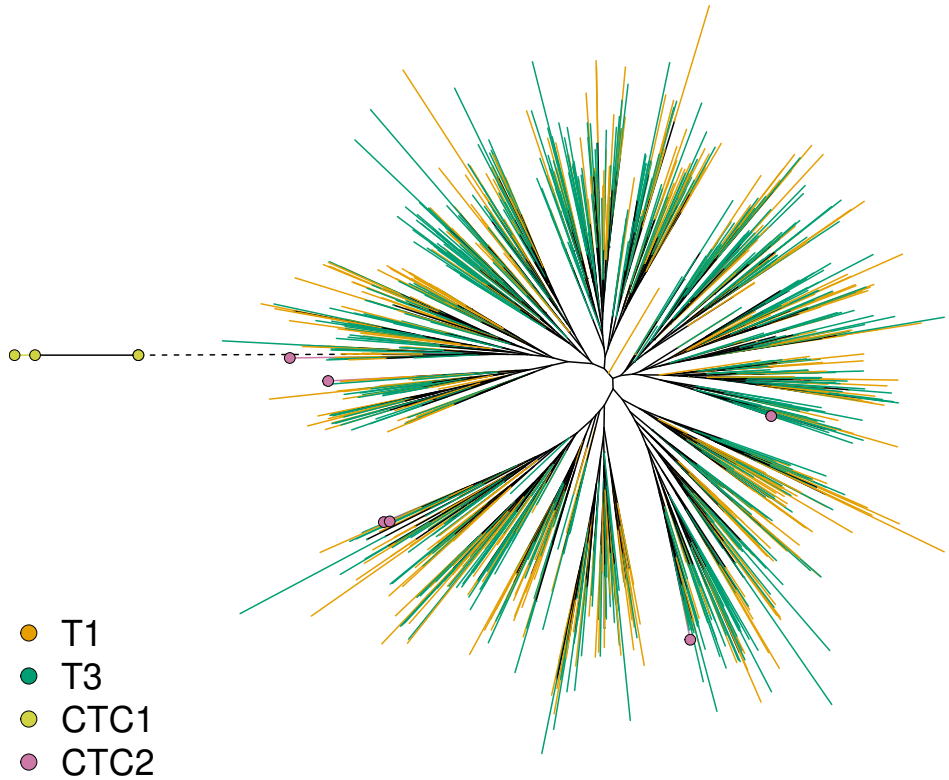


FIGURE 2. Maximum likelihood tree from the SNV data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Long branch with CTC1 cells is shortened and drawn as a dashed line. Less represented samples are marked with colored circles.

TABLE 4. Test of phylogenetic clustering on the Maximum Likelihood tree from the SNV data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated on the Maximum Likelihood trees from the SNV data for 20%, 50%, and 90% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2) and expected clustering for cells isolated from a single individual (T1 with CTC1, and T2 with CTC2). Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

	20% data density			50% data density			90% data density		
Groups	Cells	MPD	MNTD	Cells	MPD	MNTD	Cells	MPD	MNTD
T1	693	0.054	*0.003	343	0.997	0.759	40	0.630	0.087
T2	0	–	–	0	–	–	0	–	–
T3	795	*0.007	0.450	490	*0.001	0.133	150	0.304	0.818
CTC1	4	*0.001	*0.001	0	–	–	0	–	–
CTC2	6	0.960	0.601	4	0.940	0.899	1	–	–
T1 & CTC1	697	0.989	*0.002	343	0.997	0.759	40	0.630	0.087
T2 & CTC2	6	0.960	0.601	4	0.940	0.899	1	–	–

* significant support

The full SNV data set contained 12% of known data distributed across 3,980 SNVs. When further filtered, this decreased to 433 SNVs at 50% data density and to 29 SNVs at 90% data density.

Phylogenetic reconstruction from expression data. The Maximum likelihood tree reconstructed from the reduced expression data set showed significant clustering of all samples (Figure 3). This is confirmed by the phylogenetic clustering tests where all but CTC2 cells had a significant MPD p-value (Table 5).

Four out of six CTC2 cells clustered together, but on the opposite side of the tree with phylogenetic proximity to the T1 cells. This close phylogenetic relationship suggests that T1 and CTC1 were isolated from a single individual. This pattern is further reinforced as T2 cells clustered in a single compact clade with phylogenetic proximity to the CTC1 sample. A similar pattern has been observed already in the density filtering method used above, although this was complicated as majority of cells that did not belong to T1 or T3 samples were filtered out due to their low data density. Given the strong signal in the data, the simplest explanation is that samples were mislabeled and T2 and CTC1 come from the same individual (although this explanation cannot be ruled out, we were unable to isolate an event in the experiment when the mislabeling could have happen). When this relationship was tested with phylogenetic clustering methods, both MPD and MNTD confirmed the strong clustering signal between T2 and CTC1. The same tests were not significant for the T1-CTC2 grouping likely due to the presence of two non-clustering CTC2 cells. This phylogenetic structure remains stable for subsequent filtering to 50% and 90% data density (see Supplementary materials).

The phylogenies reconstructed from the same data using the Bayesian inference show similar pattern of clustering (Figure 4, Table 6), although neither CTC1 nor CTC2 formed a compact cluster. With posterior clade support for major bifurcations being weak, only the phylogenetic relationship between T2 and CTC1 was well supported across expression trees build from different data density (see Supplementary materials).

Phylogenetic reconstruction from the SNV data. The Maximum likelihood tree reconstructed from the reduced SNV data set (Figure 5) displayed many similar patterns to the previously analyzed SNV tree (Figure 2). As T2 cells were not filtered, they are now placed together with the CTC1 cells on a long branch suggesting a long shared evolutionary history. Alternatively, this could be a methodological artifact called Long Branch Attraction (Felsenstein 1978; Huelsenbeck 1997), where unrelated taxa with a large amount of accumulated changes are grouped into a single clade, although given the evidence from the expression data, this is unlikely. The phylogenetic clustering tests confirm the grouping of all samples except the CTC1 sample, which is dispersed around the tree. Clustering of samples from a single individual is rejected. Instead, the alternative hypothesis about possible mislabeling between CTC1 and CTC2 samples is supported, with a strong signal for clustering of T1 with CTC2, and T2 with CTC1.

A similar, but substantially weaker pattern of sample clustering can be observed on the Bayesian phylogeny reconstructed from the same data (Figure 6). T2 and CTC1 do not form a well-supported clusters (Table 8), but there is a strong support for a combined T2-CTC1 cluster. Additionally, while many clustering patterns are still supported by the MPD, they are rejected by MNTD. This is likely a case of the higher sensitivity of MNTD to one or two taxa that break the pattern, especially when the number of taxa is low.

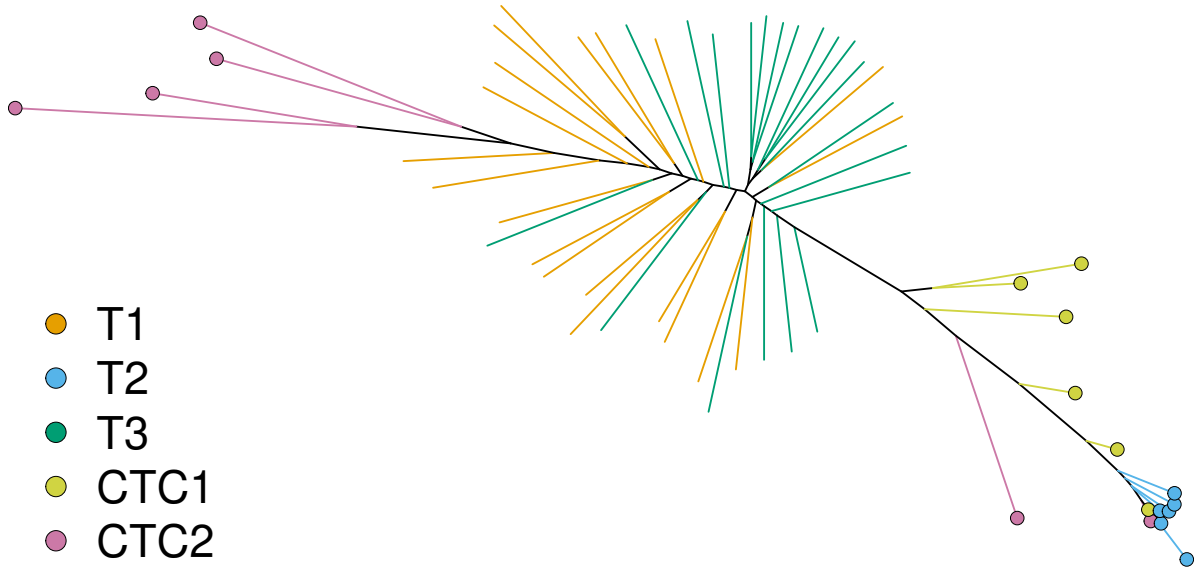


FIGURE 3. Maximum likelihood tree from the expression data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Less represented samples are marked with colored circles.

TABLE 5. Test of phylogenetic clustering on the Maximum Likelihood tree from the reduced expression data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated on the Maximum Likelihood trees from the reduced expression data for 20%, 50%, and 90% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2) and expected clustering for cells isolated from a single individual (T1 with CTC1, and T2 with CTC2). Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

		full dataset		50% data density		90% data density	
Groups	Cells	MPD	MNTD	MPD	MNTD	MPD	MNTD
T1	20	*0.001	0.745	*0.001	0.848	*0.001	*0.002
T2	6	*0.001	*0.001	*0.001	*0.001	*0.001	*0.001
T3	20	*0.001	0.406	*0.001	0.575	*0.002	0.573
CTC1	6	*0.003	*0.005	*0.001	*0.001	0.153	0.096
CTC2	6	1.000	0.977	1.000	0.997	0.995	0.985
T1 & CTC1	26	0.036	0.393	0.107	0.361	0.070	*0.001
T2 & CTC2	12	0.997	0.037	0.984	*0.024	0.995	0.504
T1 & CTC2	26	0.611	1.000	0.285	1.000	*0.007	0.912
T2 & CTC1	12	*0.001	*0.001	*0.001	*0.001	*0.018	*0.007

* significant support

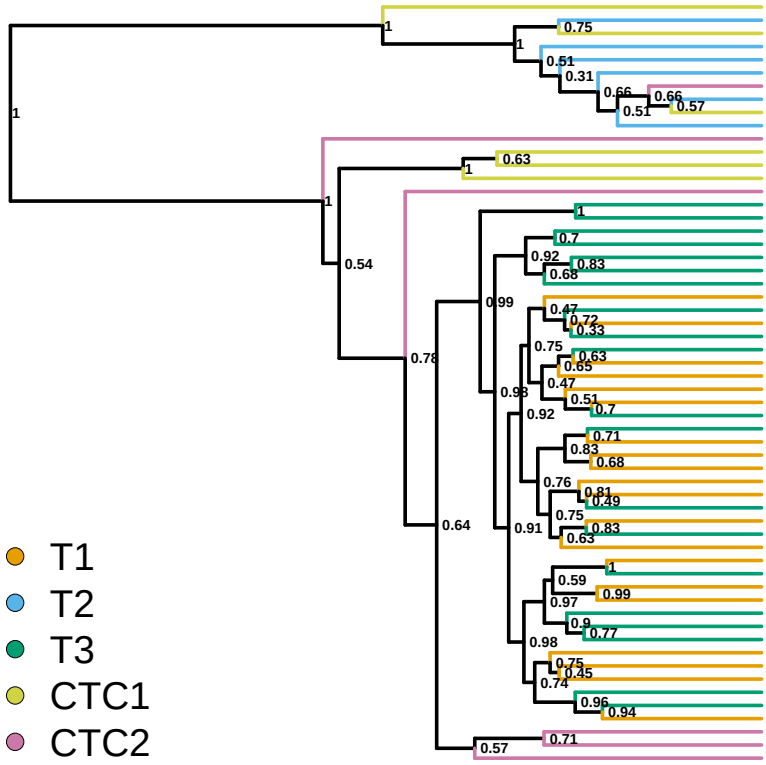


FIGURE 4. Bayesian tree from the expression data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Bayesian posterior values show the topology uncertainty.

TABLE 6. Test of phylogenetic clustering on the Bayesian tree from the reduced expression data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated on the Bayesian tree reconstructed from the reduced expression data for 20% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2), expected clustering for cells isolated from a single individual (T1 with CTC1, and T2 with CTC2) and to test a possible mislabeling between CTC1 and CTC2 samples (T1 with CTC2, and T2 with CTC1). P-values were calculated for the sample of 1,000 trees and this distribution of p-values is summarized with mean and 95% confidence interval. Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

		mean (95% HPD)	
Groups	Cells	MPD	MNTD
T1	20	*0.002 (0.001–0.004)	0.210 (0.001–0.931)
T2	6	*0.001 (0.001–0.001)	*0.003 (0.001–0.008)
T3	20	*0.001 (0.001–0.002)	0.155 (0.010–0.328)
CTC1	6	0.188 (0.106–0.362)	0.361 (0.225–0.464)
CTC2	6	0.908 (0.854–0.998)	0.799 (0.560–1.000)
T1 & CTC1	26	0.451 (0.339–0.584)	0.381 (0.028–0.803)
T2 & CTC2	12	0.803 (0.677–0.893)	0.488 (0.259–0.759)
T1 & CTC2	26	*0.012 (0.003–0.028)	0.811 (0.461–1.000)
T2 & CTC1	12	0.033 (0.009–0.110)	0.153 (0.017–0.339)

* significant support

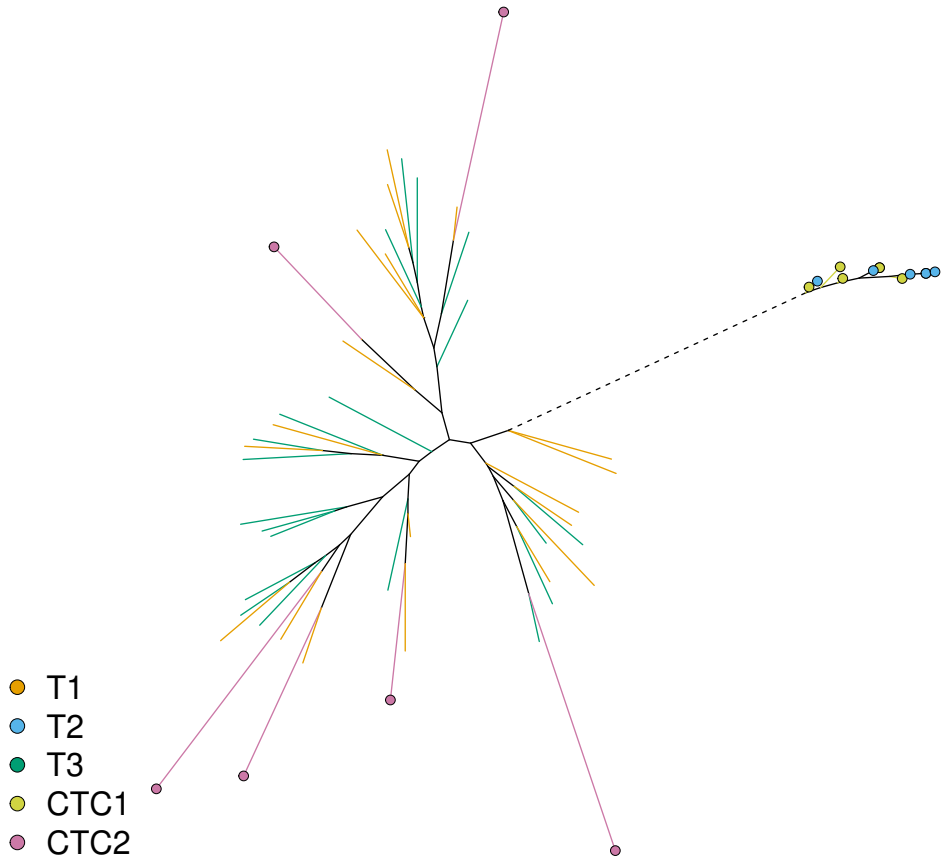


FIGURE 5. Maximum likelihood tree from the SNV data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Long branch with T2 and CTC1 cells is shortened and drawn as a dashed line. Less represented samples are marked with colored circles.

TABLE 7. Test of phylogenetic clustering on the Maximum Likelihood tree from the SNV data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated on the Maximum Likelihood trees from the reduced SNV data for 20%, 50% and 90% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2) and expected clustering for cells isolated from a single individual (T1 with CTC1 and T2 with CTC2). Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

		full dataset		50% data density		90% data density	
Groups	Cells	MPD	MNTD	MPD	MNTD	MPD	MNTD
T1	20	*0.001	0.727	*0.002	0.969	*0.001	0.288
T2	6	*0.001	*0.001	*0.001	*0.001	0.039	0.222
T3	20	*0.001	0.708	*0.001	0.884	*0.001	0.307
CTC1	6	*0.001	*0.001	*0.001	*0.001	0.034	0.235
CTC2	6	0.207	0.568	0.200	0.557	0.246	0.542
T1 & CTC1	26	0.542	0.156	0.554	0.647	0.569	0.685
T2 & CTC2	12	1.000	0.792	1.000	0.038	1.000	0.807
T1 & CTC2	26	*0.001	0.985	*0.001	0.944	*0.001	0.545
T2 & CTC1	12	*0.001	*0.001	*0.001	*0.001	*0.001	*0.002

* significant support

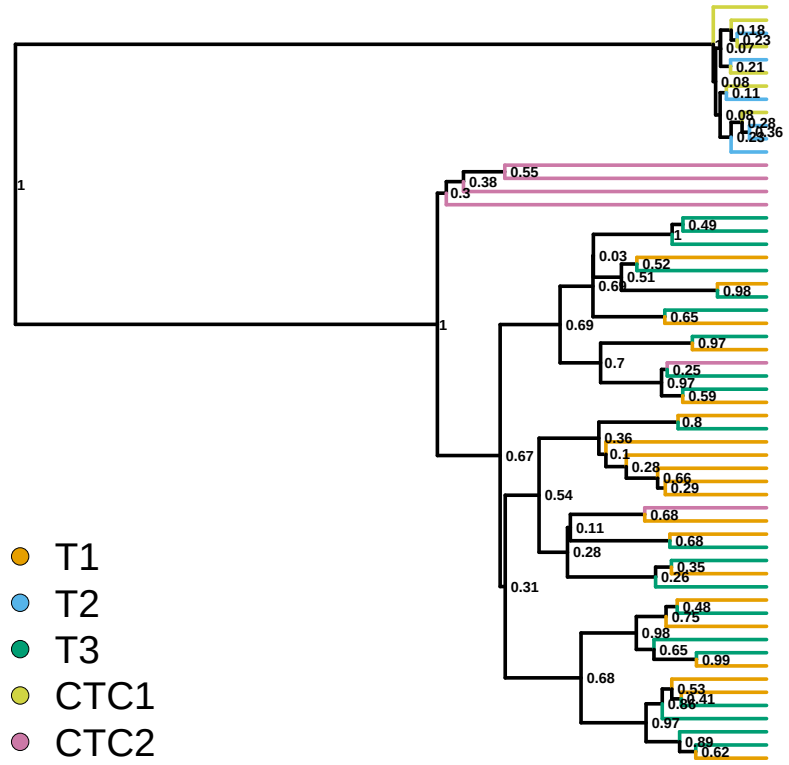


FIGURE 6. Bayesian tree from the SNV data for 20% data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Bayesian posterior values show the topology uncertainty.

TABLE 8. Test of phylogenetic clustering on the Bayesian tree from the expression data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated for the Bayesian tree reconstructed from the reduced SNV data for 20% data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2), expected clustering for cells isolated from a single individual (T1 with CTC1, and T2 with CTC2) and to test a possible mislabeling between CTC1 and CTC2 samples (T1 with CTC2, and T2 with CTC1). P-values were calculated for the sample of 1000 trees and this distribution of p-values is summarized with mean and 95% confidence interval. Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

		mean (95% HPD)	
Groups	Cells	MPD	MNTD
T1	20	*0.003 (0.001–0.006)	0.274 (0.001–0.683)
T2	6	0.060 (0.001–0.239)	0.050 (0.001–0.269)
T3	20	*0.002 (0.001–0.005)	0.298 (0.001–0.693)
CTC1	6	0.063 (0.001–0.236)	0.050 (0.001–0.245)
CTC2	6	0.164 (0.061–0.242)	0.309 (0.097–0.581)
T1 & CTC1	26	0.655 (0.541–0.756)	0.554 (0.108–0.992)
T2 & CTC2	12	0.996 (0.989–1.000)	0.427 (0.040–0.755)
T1 & CTC2	26	*0.001 (0.001–0.002)	0.250 (0.001–0.687)
T2 & CTC1	12	*0.011 (0.001–0.048)	*0.011 (0.001–0.038)

* significant support

Discussion

Phylogenetic methods using scDNA-seq data are becoming increasingly common in tumor evolution studies. scRNA-seq is currently used for studying expression profiles of cancer cells and their behavior. However, while clustering approaches to identify cells with similar expression profiles are common and frequently used, scRNA-seq data are yet to be used in phylogenetic analyses to reconstruct the population history of somatic cells. To test if the scRNA-seq contains a phylogenetic signal to reliably reconstruct the population history of cancer, we have performed an experiment to produce a known history by infecting immunosuppressed mice with human cancer cells derived from the same population. Then using two different forms of scRNA-seq data, expression values and SNVs, we have reconstructed phylogenies using Maximum likelihood and Bayesian phylogenetic methods. By comparing the reconstructed trees to the known population history, we have been able to confirm that scRNA-seq contains sufficient phylogenetic signal to reconstruct the population history of cancer. Without an explicit error model to account for an increased uncertainty in the data (Hicks et al. 2018), the phylogeny from the expression values describes the expected population history better than the one reconstructed from SNV, despite requiring lower computational costs to reconstruct.

Our results hence are a new step in cancer research as the dual-purpose of scRNA-seq can be utilized to explore both the physiological behavior of cancer cells and their population history using a single source of data.

Without any specialized phylogenetic or error models for the scRNA-seq data, conventional methods and software tools developed for systematic biology are able to reconstruct population history from this data at low computational cost. This implies that more accurate inference will be possible when and if specialized models and software is developed, and serious computational resources are employed. For example, computationally more intensive standard parametric bootstrap or Bayesian methods on the unfiltered data sets are certainly within the reach of modern computing clusters. This is a future direction for research.

In this work, we tested for phylogenetic signal on a single data set consisting of 5 tumor samples seeded using a population sample. Due to the nature of the experiment and the amount of uncertainty in the scRNA-seq data, this barred us from a more detailed exploration of the tree topology as only broad patterns, the phylogenetic clustering of cells according to sample and individual of origin, could be considered. While our clustering analysis implies that the inference results are significantly non-random, analyses of other data sets will allow to further validate our conclusions.

The degree to which low and uneven gene expression plays a role in scRNA-seq requires special attention as this causes not only a large proportion of missing data, but also burdens the known values with a significant error rate. Research should aim at trying to quantify this expression-specific error rate and build specialized models to include the uncertainty about the observed data in the phylogenetic reconstruction itself. This could potentially include removing a large proportion of low-coverage data in favor of robust analysis and proper uncertainty estimation of the inferred topology.

The estimation of the topological uncertainty, be it the Bootstrap branch support or the Bayesian posterior clade probabilities, is a staple for phylogenetic analyses. Currently existing methods for the phylogenetic analysis of scDNA-seq, such as SCITE (Jahn et al. 2016), SiFit (Zafar et al. 2017), or SCIΦ (Singer et al. 2018), do not provide this uncertainty estimate. This makes interpretation of the estimated topology difficult because a single topology can only be marginally more accurate than a number of alternative topologies. Out of package we are aware of only CellPhy, through its integration in the phylogenetic software RAxML-NG (Kozlov et al. 2019), provides an estimate of topological uncertainty through the bootstrap method. Bayesian methods could be a solution as they provide an uncertainty estimate through the posterior distribution. However, they are significantly

more computationally intensive than Maximum likelihood methods. Instead, as the size of single-cell data sets will only increase, bootstrap approximations optimized for a large amount of missing data need to be developed to provide a fast and accurate estimate of topological uncertainty.

Multiomic approaches are increasingly popular as they integrate information from multiple biological layers (Bock et al. 2016; Hasin et al. 2017; Nam et al. 2020). While copy number variations were ignored in this paper, it is possible to detect large-scale copy number variations from scRNA-seq data (Müller et al. 2018; Harmanci et al. 2020). Combined with the SNVs and expression data as analyzed in this paper, this enables a multiomic approach using just a single scRNA-seq data source, without the additional cost of DNA sequencing.

Acknowledgement

We thank Dr. Jon Preall and the Genomics Technology Development Core (CSHL) for scRNA-seq library preparation, and Pamela Moody and the Flow Cytometry Facility (CSHL) for support with single-cell sorting. We acknowledge Suzanne Russo for technical assistance with animal experiments.

AG and JCM acknowledge support from the Royal Society te Apārangi through a Rutherford Discovery Fellowship (RDF-U001702), AG, RL, and SDD acknowledge support of from an Endeavour Smart Ideas grant (U00X1912), AG acknowledges support from a Data Science Programmes grant (UOAX1932), SDD acknowledges support from a Rutherford Discovery Fellowship (RDF-U001802) and the NHI/NCI grant (1K99CA215362-01), and DLS acknowledges support from the NCI grant (SP01CA013106-Project 3). We would also like to acknowledge the CSNL Next-Gen Sequencing Core (NCI-2P30CA45508).

References

[1] Christopher Abbosh, Nicolai J Birkbak, Gareth A Wilson, Mariam Jamal-Hanjani, Tudor Constantin, Raheleh Salari, John Le Quesne, David A Moore, Selvaraju Veeriah, et al. "Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution". *Nature* 545.7655 (Apr. 2017), pp. 446–451.

[2] Julio A Aguirre-Ghiso. "On the theory of tumor self-seeding: implications for metastasis progression in humans". *Breast Cancer Res.* 12.2 (Apr. 2010), p. 304.

[3] David Aldous. "Probability Distributions on Cladograms". *Random Discrete Structures.* Springer New York, 1996, pp. 1–18.

[4] David J Aldous. "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today". *Stat. Sci.* 16.1 (Feb. 2001), pp. 23–34.

[5] João M Alves, Sonia Prado-Lopez, Jose Manuel Cameselle-Teijeiro, and David Posada. "Rapid evolution and biogeographic spread in a colorectal cancer". May 2019.

[6] João M Alves, Tamara Prieto, and David Posada. "Multiregional Tumor Trees Are Not Phylogenies". *Trends Cancer Res.* 3.8 (Aug. 2017), pp. 546–550.

[7] Tallulah S Andrews and Martin Hemberg. "Identifying cell populations with scRNASeq". *Mol. Aspects Med.* 59 (Feb. 2018), pp. 114–122.

[8] Mihaela Angelova, Bernhard Mlecnik, Angela Vasaturo, Gabriela Bindea, Tessa Fredriksen, Lucie Lafontaine, Bénédicte Buttard, Erwan Morgand, Daniela Bruni, et al. "Evolution of Metastases in Space and Time under Immune Selection". *Cell* (Oct. 2018).

[9] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. "Cancer evolution: mathematical models and computational inference". *Syst. Biol.* 64.1 (Jan. 2015), e1–25.

[10] David Benjamin, Takuto Sato, Kristian Cibulskis, Gad Getz, Chip Stewart, and Lee Lichtenstein. "Calling Somatic SNVs and Indels with Mutect2". *bioRxiv* (2019). DOI: 10.1101/861054. eprint: https://www.biorxiv.org/content/early/2019/12/02/861054.full.pdf. URL: https://www.biorxiv.org/content/early/2019/12/02/861054.

[11] Y Benjamini and Y Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *J. R. Stat. Soc.* (1995).

[12] Christoph Bock, Matthias Farlik, and Nathan C Sheffield. "Multi-Omics of Single Cells: Strategies and Applications". *Trends Biotechnol.* 34.8 (Aug. 2016), pp. 605–608.

[13] Remco Bouckaert, Timothy G Vaughan, Joelle Barido-Sottani, Sebastian Duchene, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kuhnert, et al. "BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis". *PLoS Comput. Biol.* 15.4 (Apr. 2019), e1006650.

[14] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A Graham, Guido Sanguinetti, and Andrea Sottoriva. "Detecting repeated cancer evolution from multi-region tumor sequencing data". *Nat. Methods* 15.9 (Sept. 2018), pp. 707–714.

[15] Giulio Caravagna, Timon Heide, Marc Williams, Luis Zapata, Daniel Nichol, Ketevan Chkhaidze, William Cross, George D Cresswell, Benjamin Werner, et al. "Model-based tumor subclonal reconstruction". Mar. 2019.

[16] Anna K Casasent, Aislyn Schalck, Ruli Gao, Emi Sei, Annalyssa Long, William Pangburn, Tod Casasent, Funda Meric-Bernstam, Mary E Edgerton, et al. "Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing". *Cell* 172.1-2 (Jan. 2018), 205–217.e12.

[17] Jiahuan Chen, Qian Zhou, Yangfan Wang, and Kang Ning. "Single-cell SNP analyses and interpretations based on RNA-Seq data for colon cancer research". *Sci. Rep.* 6.1 (Sept. 2016), p. 34420.

[18] Kevin J Cheung and Andrew J Ewald. "A collective route to metastasis: Seeding by tumor cell clusters". *Science* 352.6282 (Apr. 2016), pp. 167–169.

[19] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. "PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors". *Genome Biol.* 16 (Feb. 2015), p. 35.

[20] Harald Detering, Laura Tomás, Tamara Prieto, and David Posada. "Accuracy of somatic variant detection in multiregional tumor sequencing data". May 2019.

[21] Jean Fan, Kamil Slowikowski, and Fan Zhang. "Single-cell transcriptomics in cancer: computational challenges and opportunities". *Exp. Mol. Med.* 52.9 (Sept. 2020), pp. 1452–1465.

[22] Joseph Felsenstein. "Cases in which Parsimony or Compatibility Methods will be Positively Misleading". *Syst. Biol.* 27.4 (Dec. 1978), pp. 401–410.

[23] Joseph Felsenstein. "Confidence Limits on Phylogenies: An Approach Using the Bootstrap". *Evolution* 39.4 (1985), pp. 783–791.

[24] Stephen J Fleming, John C Marioni, and Mehrtash Babadi. "CellBender remove-background: a deep generative model for unsupervised removal of background noise from scRNA-seq datasets". Oct. 2019.

[25] Robert P. Freckleton. "Fast likelihood calculations for comparative analyses". *Methods in Ecology and Evolution* 3.5 (2012), pp. 940–947. DOI: `10.1111/j.2041-210X.2012.00220.x`. eprint: `https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210X.2012.00220.x`. URL: `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2012.00220.x`.

[26] Charles Gawad, Winston Koh, and Stephen R Quake. "Single-cell genome sequencing: current state of the science". *Nat. Rev. Genet.* 17.3 (Mar. 2016), pp. 175–188.

[27] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, M Math, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing". *N. Engl. J. Med.* 366.10 (Mar. 2012), pp. 883–892.

[28] Laura González-Silva, Laura Quevedo, and Ignacio Varela. "Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies". *Trends Cancer Res.* 6.1 (Jan. 2020), pp. 13–19.

[29] Alan Grafen and William Donald Hamilton. "The phylogenetic regression". *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 326.1233 (1989), pp. 119–157. DOI: 10.1098/rstb.1989.0106. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.1989.0106. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1989.0106.

[30] Akdes Serin Harmanci, Arif O Harmanci, and Xiaobo Zhou. "Inference of Clonal Copy Number Alterations from RNASequencing Data". *Journal of Cancer Immunology* 2.3 (2020).

[31] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". *Genome Biol.* 18.1 (May 2017), p. 83.

[32] Joseph Heled and Alexei J Drummond. "Bayesian inference of population size history from multiple loci". *BMC Evol. Biol.* 8 (Oct. 2008), p. 289.

[33] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. "Missing data and technical variability in single-cell RNA-sequencing experiments". *Biostatistics* 19.4 (Oct. 2018), pp. 562–578.

[34] Diep Thi Hoang, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. "UFBoot2: Improving the Ultrafast Bootstrap Approximation". *Mol. Biol. Evol.* 35.2 (Feb. 2018), pp. 518–522.

[35] Richard R Hudson et al. "Gene genealogies and the coalescent process". *Oxford surveys in evolutionary biology* 7.1 (1990), p. 44.

[36] J P Huelsenbeck. "Is the Felsenstein zone a fly trap?" *Syst. Biol.* 46.1 (Mar. 1997), pp. 69–74.

[37] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. "Tree inference for single-cell data". *Genome Biol.* 17 (May 2016), p. 86.

[38] Livnat Jerby-Arnon, Parin Shah, Michael S Cuoco, Christopher Rodman, Mei-Ju Su, Johannes C Melms, Rachel Leeson, Abhay Kanodia, Shaolin Mei, et al. "A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade". *Cell* 175.4 (Nov. 2018), 984–997.e24.

[39] Wei Jiang, Si-Yun Chen, Hong Wang, De-Zhu Li, and John J Wiens. "Should genes with missing data be excluded from phylogenetic analyses?" *Mol. Phylogenet. Evol.* 80 (Nov. 2014), pp. 308–318.

[40] Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. "Inferring parsimonious migration histories for metastatic cancers". *Nat. Genet.* 50.5 (May 2018), pp. 718–726.

[41] S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. "Picante: R tools for integrating phylogenies and ecology". *Bioinformatics* 26 (2010), pp. 1463–1464.

[42] J F C Kingman. "The coalescent". *Stochastic Process. Appl.* 13.3 (Sept. 1982), pp. 235–248.

[43] Natalia L Komarova. "Spatial stochastic models for cancer initiation and progression". *Bull. Math. Biol.* 68.7 (Oct. 2006), pp. 1573–1599.

[44] Alexey M Kozlov, Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. "RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference". *Bioinformatics* 35.21 (Nov. 2019), pp. 4453–4455.

[45] Alexey Kozlov, João Alves, Alexandros Stamatakis, and David Posada. "CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data". Aug. 2020.

[46] M K Kuhner, J Yamato, and J Felsenstein. "Maximum likelihood estimation of population growth rates based on the coalescent". *Genetics* 149.1 (May 1998), pp. 429–434.

[47] Jack Kuipers, Mustafa Anıl Tuncel, Pedro Ferreira, Katharina Jahn, and Niko Beerenwinkel. "Single-cell copy number calling and event history reconstruction". Apr. 2020.

[48] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, et al. "Eleven grand challenges in single-cell data science". *Genome Biol.* 21.1 (Dec. 2020), p. 1.

[49] Philippe Lemey, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. "Bayesian phylogeography finds its roots". *PLoS Comput. Biol.* 5.9 (Sept. 2009), e1000520.

[50] Philippe Lemey, Andrew Rambaut, John J Welch, and Marc A Suchard. "Phylogeography takes a relaxed random walk in continuous space and time". *Mol. Biol. Evol.* 27.8 (Aug. 2010), pp. 1877–1885.

[51] Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, et al. "Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer". *Genome Res.* 27.8 (Aug. 2017), pp. 1287–1299.

[52] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16 (June 2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. eprint: https://academic.oup.com/bioinformatics/article-pdf/25/16/2078/531810/btp352.pdf. URL: https://doi.org/10.1093/bioinformatics/btp352.

[53] Shaoping Ling, Zheng Hu, Zuyu Yang, Fang Yang, Yawei Li, Pei Lin, Ke Chen, Lili Dong, Lihua Cao, et al. "Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution". *Proc. Natl. Acad. Sci. U. S. A.* 112.47 (Nov. 2015), E6496–E6505.

[54] Fenglin Liu, Yuanyuan Zhang, Lei Zhang, Ziyi Li, Qiao Fang, Ranran Gao, and Zemin Zhang. "Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data". *Genome Biol.* 20.1 (Nov. 2019), p. 242.

[55] Aaron T L Lun, Samantha Riesenfeld, Tallulah Andrews, The Phuong Dao, Tomas Gomes, participants in the 1st Human Cell Atlas Jamboree, and John C Marioni. "EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data". *Genome Biol.* 20.1 (Mar. 2019), p. 63.

[56] Lovelace J Luquette, Craig L Bohrson, Max A Sherman, and Peter J Park. "Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance". *Nat. Commun.* 10.1 (Aug. 2019), p. 3908.

[57] Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. "Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data". *Nat. Commun.* 10.1 (June 2019), p. 2750.

[58] Elaine R Mardis and Richard K Wilson. "Cancer genome sequencing: a review". *Hum. Mol. Genet.* 18.R2 (Oct. 2009), R163–8.

[59] Bui Quang Minh, Minh Anh Thi Nguyen, and Arndt von Haeseler. "Ultrafast approximation for phylogenetic bootstrap". *Mol. Biol. Evol.* 30.5 (May 2013), pp. 1188–1195.

[60] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era". *Molecular Biology and Evolution* 37.5 (Feb. 2020), pp. 1530–1534. ISSN: 0737-4038. DOI: 10.1093/molbev/msaa015. eprint: https://academic.oup.com/mbe/article-pdf/37/5/1530/33386032/msaa015.pdf. URL: https://doi.org/10.1093/molbev/msaa015.

[61] Andy J Minn, Gaorav P Gupta, Peter M Siegel, Paula D Bos, Weiping Shu, Dilip D Giri, Agnes Viale, Adam B Olshen, William L Gerald, et al. "Genes that mediate breast cancer metastasis to lung". *Nature* 436.7050 (July 2005), pp. 518–524.

[62] Sayaka Miura, Karen Gomez, Oscar Murillo, Louise A Huuki, Tracy Vu, Tiffany Buturla, and Sudhir Kumar. "Predicting clone genotypes from tumor bulk sequencing of multiple samples". *Bioinformatics* 34.23 (June 2018), pp. 4017–4026.

[63] Sayaka Miura, Tracy Vu, Jiamin Deng, Tiffany Buturla, Olumide Oladeinde, Jiyeong Choi, and Sudhir Kumar. "Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data". *Sci. Rep.* 10.1 (Feb. 2020), p. 3498.

[64] Nicola F Müller, David A Rasmussen, and Tanja Stadler. "The Structured Coalescent and Its Approximations". *Mol. Biol. Evol.* 34.11 (Nov. 2017), pp. 2970–2981.

[65] Sören Müller, Ara Cho, Siyuan J Liu, Daniel A Lim, and Aaron Diaz. "CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones". *Bioinformatics* 34.18 (Sept. 2018), pp. 3217–3219.

[66] Sören Müller, Gary Kohanbash, S John Liu, Beatriz Alvarado, Diego Carrera, Aparna Bhaduri, Payal B Watchmaker, Garima Yagnik, Elizabeth Di Lullo, et al. "Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment". *Genome Biol.* 18.1 (Dec. 2017), p. 234.

[67] Matthew A Myers, Gryte Satas, and Benjamin J Raphael. "CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples". *Cell Syst* 8.6 (June 2019), 514–522.e5.

[68] Hidewaki Nakagawa and Masashi Fujita. "Whole genome sequencing analysis for cancer genomics and precision medicine". *Cancer Sci.* 109.3 (Mar. 2018), pp. 513–522.

[69] Anna S Nam, Ronan Chaligne, and Dan A Landau. "Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics". *Nat. Rev. Genet.* (Aug. 2020).

[70] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, et al. "Tumour evolution inferred by single-cell sequencing". *Nature* 472.7341 (Apr. 2011), pp. 90–94.

[71] Thale Kristin Olsen and Ninib Baryawno. "Introduction to Single-Cell RNA Sequencing". *Current Protocols in Molecular Biology* 122.1 (2018), e57. DOI: 10.1002/cpmb.57. eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpmb.57. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpmb.57.

[72] Mark Pagel, Andrew Meade, and Daniel Barker. "Bayesian Estimation of Ancestral Character States on Phylogenies". *Systematic Biology* 53.5 (Oct. 2004), pp. 673–684. ISSN: 1063-5157. DOI: 10.1080/10635150490522232. eprint: https://academic.oup.com/sysbio/article-pdf/53/5/673/24197159/53-5-673.pdf. URL: https://doi.org/10.1080/10635150490522232.

[73] Anna Petrackova, Michal Vasinek, Lenka Sedlarikova, Tereza Dyskova, Petra Schneiderova, Tomas Novosad, Tomas Papajik, and Eva Kriegova. "Standardization of Sequencing Coverage Depth in NGS: Recommendation for Detection of Clonal and Subclonal Mutations in Cancer Diagnostics". *Front. Oncol.* 9 (Sept. 2019), p. 851.

[74] Olivier Poirion, Xun Zhu, Travers Ching, and Lana X Garmire. "Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage". *Nat. Commun.* 9.1 (Nov. 2018), p. 4892.

[75] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, et al. "Scaling accurate genetic variant discovery to tens of thousands of samples". *bioRxiv* (2018). DOI: 10.1101/201178. eprint: https://www.biorxiv.org/content/early/2018/07/24/201178.full.pdf. URL: https://www.biorxiv.org/content/early/2018/07/24/201178.

[76] David Posada. "Cancer Molecular Evolution". *J. Mol. Evol.* 81.3-4 (Oct. 2015), pp. 81–83.

[77] David Posada. "CellCoal: Coalescent Simulation of Single-Cell Sequencing Samples". *Mol. Biol. Evol.* 37.5 (May 2020), pp. 1535–1542.

[78] Nicola E Potter, Luca Ermini, Elli Papaemmanuil, Giovanni Cazzaniga, Gowri Vijayaraghavan, Ian Titley, Anthony Ford, Peter Campbell, Lyndal Kearney, et al. "Single-cell mutational profiling and clonal phylogeny in cancer". *Genome Res.* 23.12 (Dec. 2013), pp. 2115–2125.

[79] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. "Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data". *eLife* 6 (Nov. 2017).

[80] Sneha R Rao, Jason A Somarelli, Erdem Altunel, Laura E Selmic, Mark Byrum, Maya U Sheth, Serene Cheng, Kathryn E Ware, So Young Kim, et al. "From the Clinic to the Bench and Back Again in One Dog Year: How a Cross-Species Pipeline to Identify New Treatments for Sarcoma Illuminates the Path Forward in Precision Medicine". *Front. Oncol.* 10 (Feb. 2020), p. 117.

[81] Johannes G Reiter, Alvin P Makohon-Moore, Jeffrey M Gerold, Ivana Bozic, Krishnendu Chatterjee, Christine A Iacobuzio-Donahue, Bert Vogelstein, and Martin A Nowak. "Reconstructing metastatic seeding patterns of human cancers". *Nat. Commun.* 8 (Jan. 2017), p. 14114.

[82] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A Smith, Cydney B Nielsen, Jessica N McAlpine, et al. "Clonal genotype and population structure inference from single-cell tumor sequencing". *Nat. Methods* 13.7 (July 2016), pp. 573–576.

[83] Gryte Satas and Benjamin J Raphael. "Haplotype phasing in single-cell DNA-sequencing data". *Bioinformatics* 34.13 (July 2018), pp. i211–i217.

[84] Patricia M Schnepp, Mengjie Chen, Evan T Keller, and Xiang Zhou. "SNV identification from single-cell RNA sequencing data". *Hum. Mol. Genet.* 28.21 (Nov. 2019), pp. 3569–3583.

[85] Russell Schwartz and Alejandro A. Schäffer. "The evolution of tumour phylogenetics: principles and practice". *Nature Reviews Genetics* 18.4 (Apr. 2017), pp. 213–229. ISSN: 1471-0064. DOI: `10.1038/nrg.2016.170`. URL: `https://doi.org/10.1038/nrg.2016.170`.

[86] Jochen Singer, Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. "Single-cell mutation identification via phylogenetic inference". *Nat. Commun.* 9.1 (Dec. 2018), p. 5144.

[87] Samuel P Strom. "Current practices and guidelines for clinical next-generation sequencing oncology testing". *Cancer Biol Med* 13.1 (Mar. 2016), pp. 3–11.

[88] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, et al. "Comprehensive Integration of Single-Cell Data". *Cell* 177 (2019), pp. 1888–1902. DOI: `10.1016/j.cell.2019.05.031`. URL: `https://doi.org/10.1016/j.cell.2019.05.031`.

[89] Simon Tavaré. "Some probabilistic and statistical problems in the analysis of DNA sequences". *Lectures on mathematics in the life sciences* 17.2 (1986), pp. 57–86.

[90] Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Dudoit, and Lieven Clement. "Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications". *Genome Biol.* 19.1 (Feb. 2018), p. 24.

[91] Campbell O Webb. "Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees". *Am. Nat.* 156.2 (Aug. 2000), pp. 145–155.

[92] Benjamin Werner, Jack Case, Marc J Williams, Kate Chkhaidze, Daniel Temko, Javier Fernandez-Mateos, George D Cresswell, Daniel Nichol, William Cross, et al. "Measuring single cell divisions in human cancers from multi-region sequencing data". Feb. 2019.

[93] John J Wiens. "Missing data and the design of phylogenetic analyses". *J. Biomed. Inform.* 39.1 (Feb. 2006), pp. 34–42.

[94] Zhenxiang Xi, Liang Liu, and Charles C Davis. "The Impact of Missing Data on Species Tree Estimation". *Mol. Biol. Evol.* 33.3 (Mar. 2016), pp. 838–860.

[95] Xiaoqing Yu, Y Ann Chen, Jose R Conejo-Garcia, Christine H Chung, and Xuefeng Wang. "Estimation of immune cell content in tumor using single-cell RNA-seq reference data". *BMC Cancer* 19.1 (July 2019), p. 715.

[96] Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. "BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies". *Genome Biol.* 16.1 (Feb. 2015), p. 36.

[97] H Zafar, N Navin, K Chen, and L Nakhleh. "SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data". *bioRxiv* (2018).

[98] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. "SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models". *Genome Biol.* 18.1 (Sept. 2017), p. 178.

[99] Hamim Zafar, Yong Wang, Luay Nakhleh, Nicholas Navin, and Ken Chen. "Monovar: single-nucleotide variant detection in single cells". *Nat. Methods* 13.6 (June 2016), pp. 505–507.

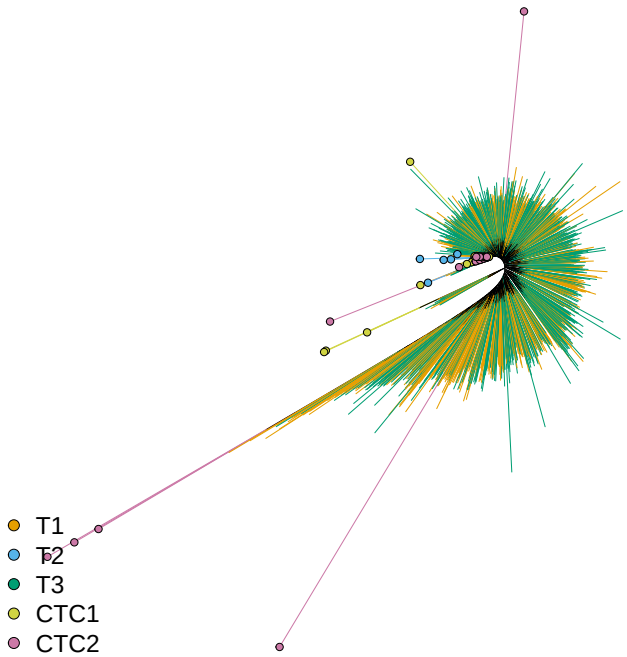
[100] Weiwei Zhai, Tony Kiat-Hon Lim, Tong Zhang, Su-Ting Phang, Zenia Tiang, Peiyong Guan, Ming-Hwee Ng, Jia Qi Lim, Fei Yao, et al. "The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma". *Nat. Commun.* 8 (Feb. 2017), p. 4565.

[101] Zi-Ming Zhao, Bixiao Zhao, Yalai Bai, Atila Iamarino, Stephen G Gaffney, Joseph Schlessinger, Richard P Lifton, David L Rimm, and Jeffrey P Townsend. "Early and multiple origins of metastatic lineages within primary tumors". *Proc. Natl. Acad. Sci. U. S. A.* 113.8 (Feb. 2016), pp. 2140–2145.

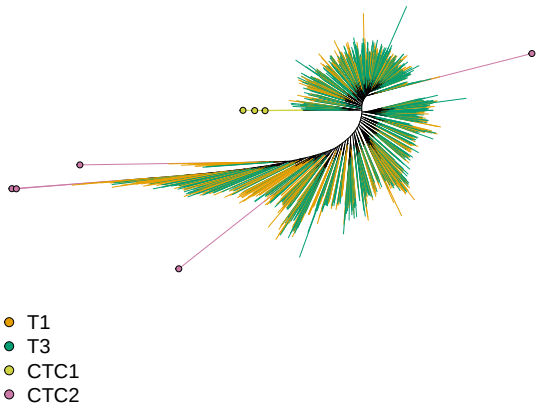
Supplementary materials

.1. **Ultrafast bootstrap convergence.** Ultrafast bootstrap (Minh et al. 2013; Hoang et al. 2018) is an approximation of the standard parametric bootstrap (Felsenstein 1985). Compared to the standard parametric bootstrap, the ultrafast bootstrap is significantly faster and in some cases more accurate (Minh et al. 2013). However, the ultrafast bootstrap failed to converge for on the non-reduced data sets. This might affect the final topology as well as bootstrap support scores. According to the authors, one case when the ultrafast bootstrap fails to converge is on data sets with a low phylogenetic signal (Minh et al. 2013). Alternatively, this might be caused by a large proportion of missing data in our analyses and significantly longer runs would be required. Instead, we have decided to reduce the data set to 58 best performing cells. On these reduced data sets, the convergence was no longer an issue.

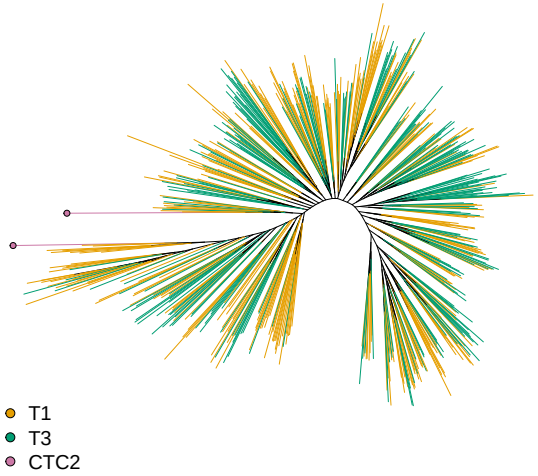
Tree figures.



(a) 20% data density

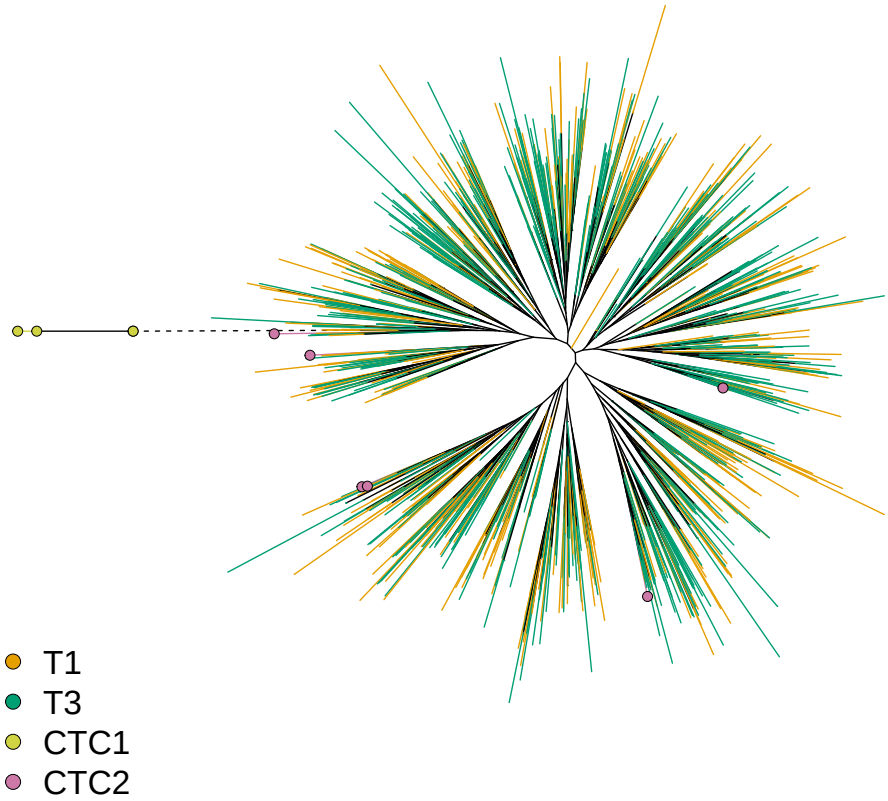


(b) 50% data density

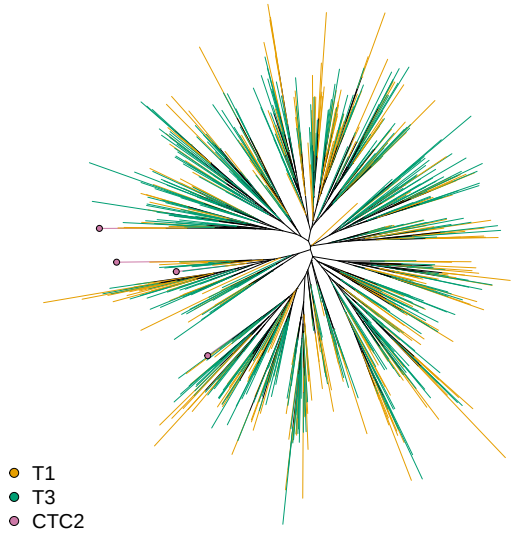


(c) 90% data density

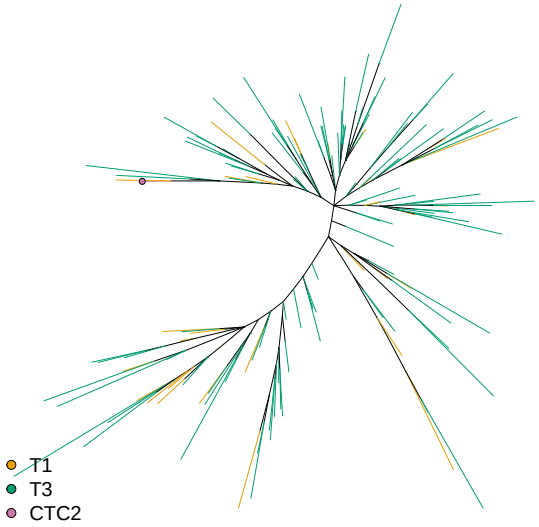
FIGURE 7. Maximum likelihood trees from the expression data for 20% (Figure 7a), 50% (Figure 7b), and 90% (Figure 7c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Less represented samples are marked with colored circles.



(a) 20% data density

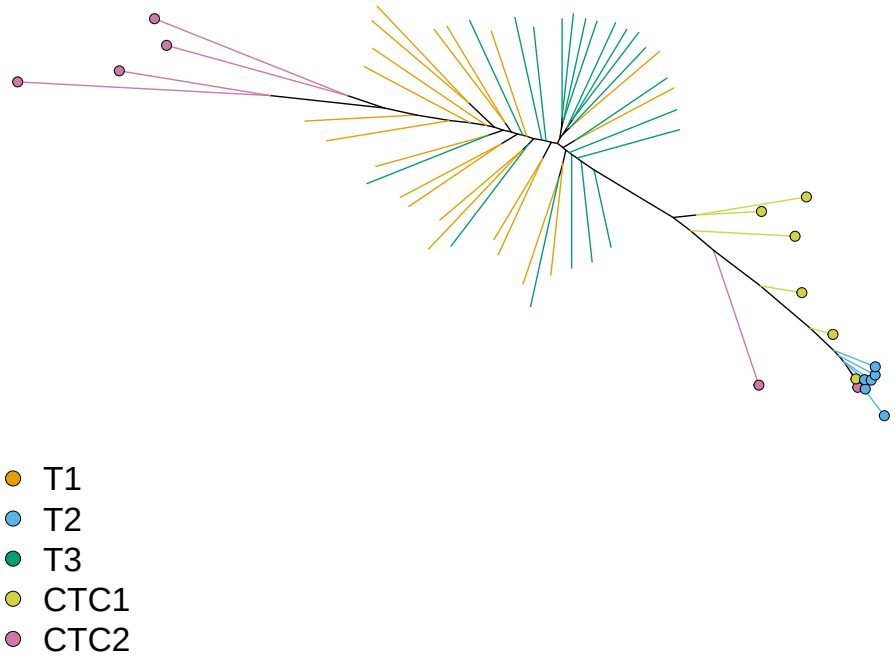


(b) 50% data density

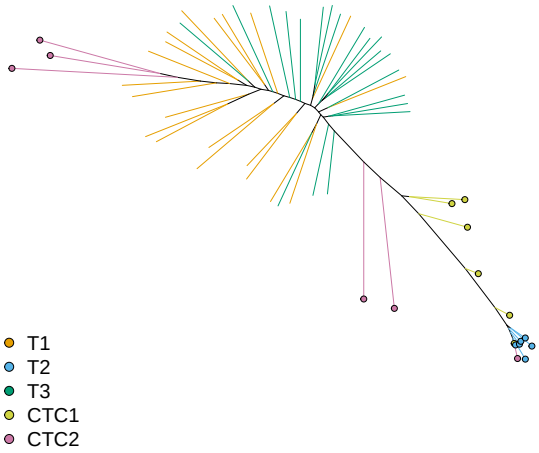


(c) 90% data density

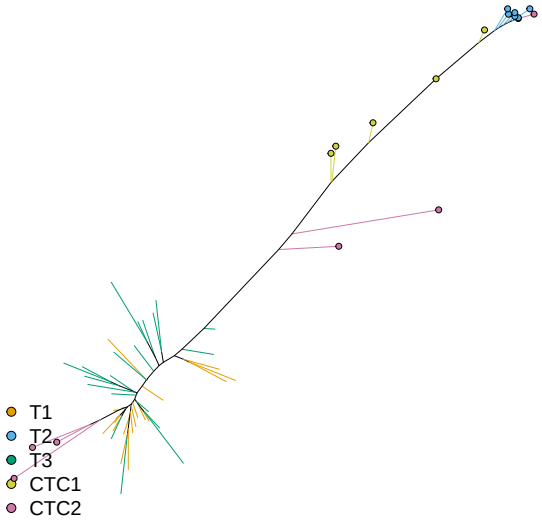
FIGURE 8. Maximum likelihood trees from the SNV data for 20% (Figure 8a), 50% (Figure 8b), and 90% (Figure 8c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Less represented samples are marked with colored circles.



(a) 20% data density

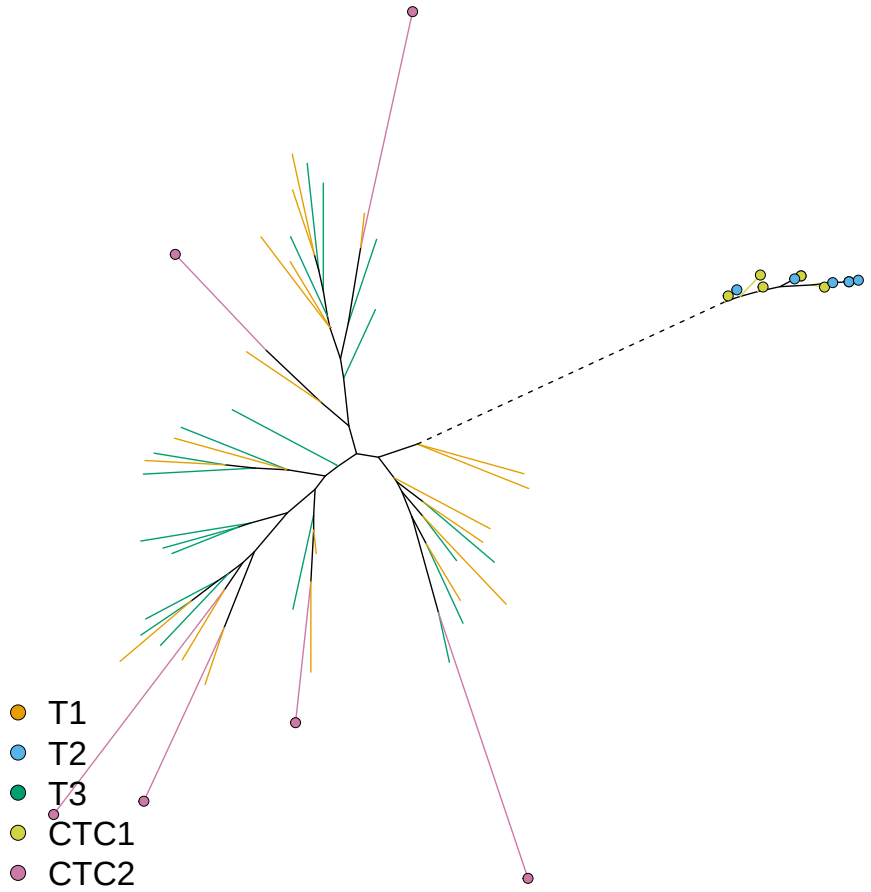


(b) 50% data density

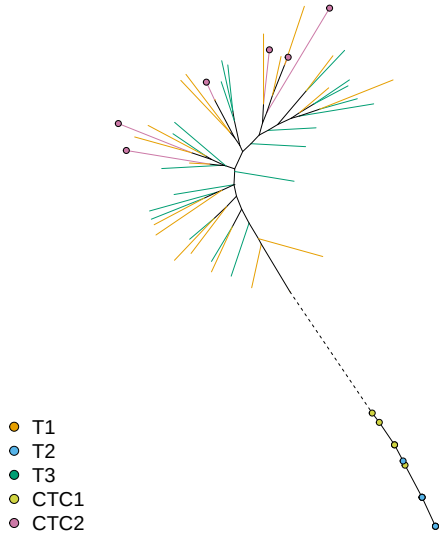


(c) 90% data density

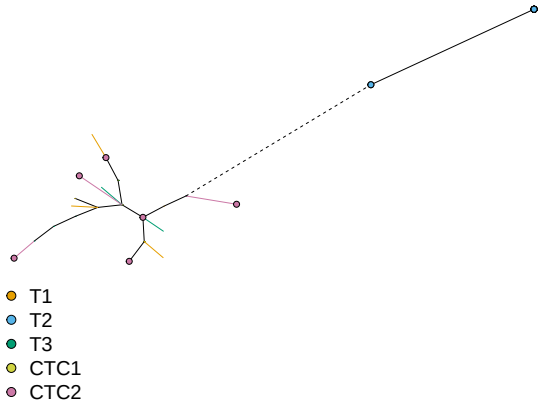
FIGURE 9. Maximum likelihood trees from the reduced expression data for 20% (Figure 9a), 50% (Figure 9b), and 90% (Figure 9c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Less represented samples are marked with colored circles.



(a) 20% data density

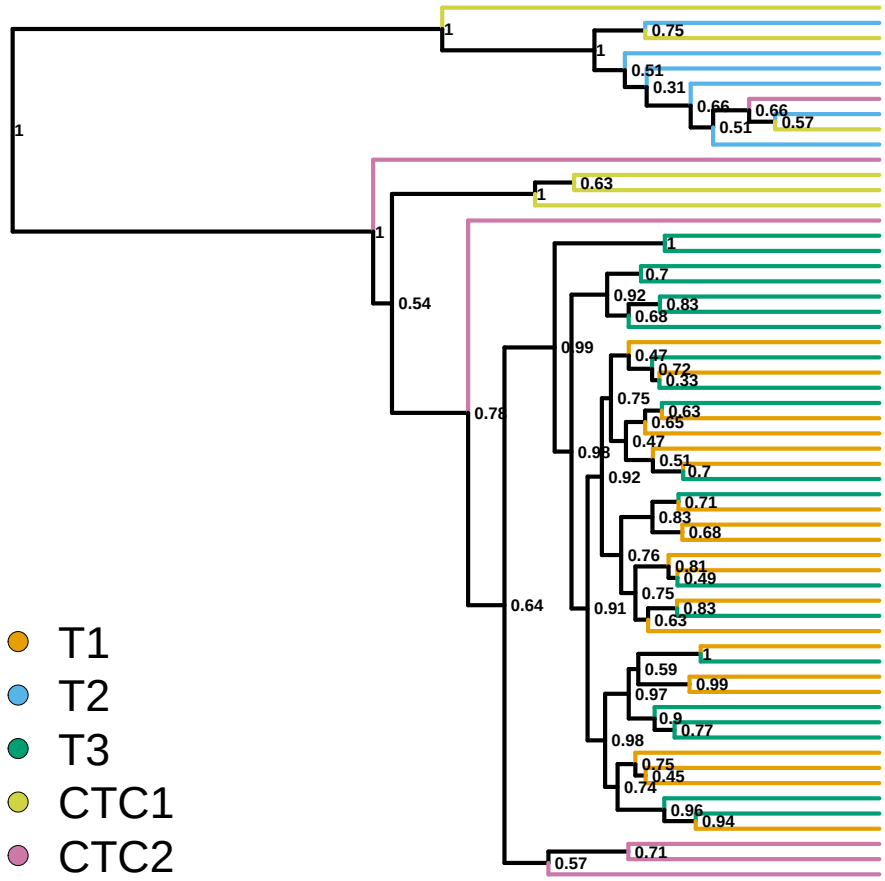


(b) 50% data density

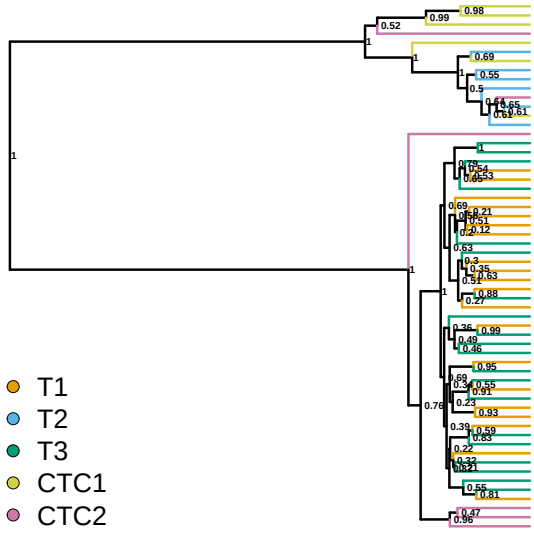


(c) 90% data density

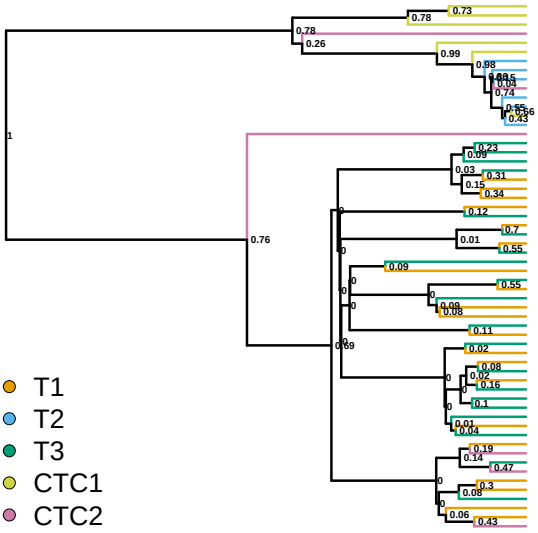
FIGURE 10. Maximum likelihood trees from the reduced SNV data for 20% (Figure 10a), 50% (Figure 10b), and 90% (Figure 10c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Long branch with T2 and CTC1 cells is shortened and drawn as a dashed line. Less represented samples are marked with colored circles.



(a) 20% data density

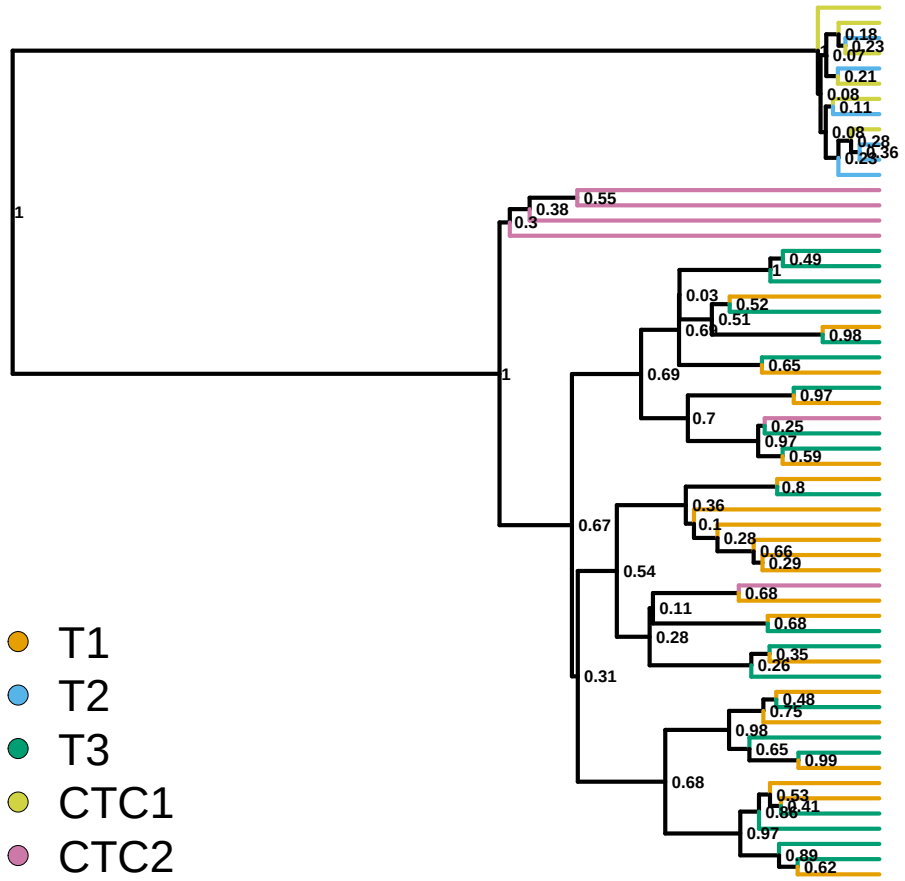


(b) 50% data density

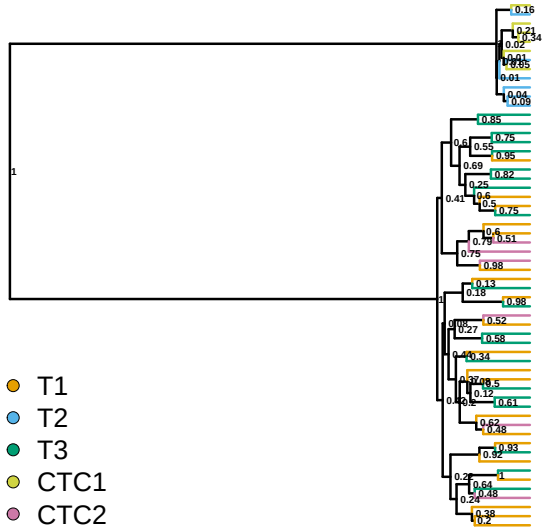


(c) 90% data density

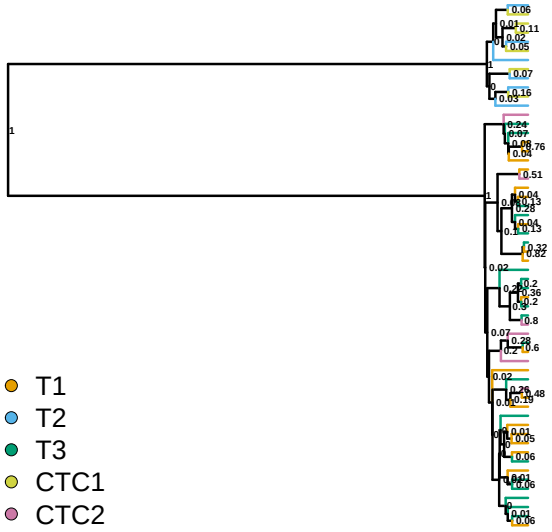
FIGURE 11. Bayesian trees from the reduced expression data for 20% (Figure 11a), 50% (Figure 11b), and 90% (Figure 11c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Bayesian posterior values show the topology uncertainty.



(a) 20% data density



(b) 50% data density



(c) 90% data density

FIGURE 12. Bayesian trees from the reduced SNV data for 20% (Figure 12a), 50% (Figure 12b), and 90% (Figure 12c) data density. Terminal branches are colored according to cell's sample of origin (T1, T2, T3, CTC1, CTC2). Bayesian posterior values show the topology uncertainty.

Phylogenetic clustering tests for Bayesian trees.

TABLE 9. Test of phylogenetic clustering on the Bayesian tree from the expression data and SNV data. Mean Pairwise Distance (MPD) and Mean Nearest Taxon Distance (MNTD) calculated for the Bayesian trees reconstructed from the reduced expression (Table 9a) and reduced SNV data (Table 9b) for 20%, 50% and 90% data density data density. P-values for MPD and MNTD were calculated for each sample (T1, T2, T3, CTC1, CTC2), expected clustering for cells isolated from a single individual (T1 with CTC1 and T2 with CTC2) and to test a possible mislabeling between CTC1 and CTC2 samples (T1 with CTC2, and T2 with CTC1). P-values were calculated for the sample of 1000 trees and this distribution of p-values is summarized with mean and 95% confidence interval. Significant p-values at $\alpha = 0.05$ after correcting for multiple comparisons using the False Discovery Rate method (Benjamini et al. 1995) are marked with an asterisk.

(a) Expression data

		full dataset		50% data density		90% data density	
Groups	Cells	MPD	MNTD	MPD	MNTD	MPD	MNTD
T1	20	*0.001 (0.001–0.001)	*0.014 (0.001–0.040)	*0.001 (0.001–0.001)	*0.043 (0.004–0.100)	*0.002 (0.001–0.004)	0.210 (0.001–0.931)
T2	6	*0.002 (0.001–0.003)	*0.004 (0.001–0.014)	*0.001 (0.001–0.001)	*0.002 (0.001–0.004)	*0.001 (0.001–0.001)	*0.003 (0.001–0.008)
T3	20	*0.001 (0.001–0.002)	*0.048 (0.006–0.104)	*0.001 (0.001–0.002)	0.097 (0.011–0.203)	*0.001 (0.001–0.002)	0.155 (0.010–0.328)
CTC1	6	0.989 (0.974–0.999)	0.472 (0.297–0.616)	0.176 (0.151–0.197)	0.311 (0.151–0.437)	0.188 (0.106–0.362)	0.361 (0.225–0.464)
CTC2	6	0.697 (0.635–0.745)	0.926 (0.758–1.000)	0.834 (0.748–0.888)	0.554 (0.456–0.627)	0.908 (0.854–0.998)	0.799 (0.560–1.000)
T1 & CTC1	26	0.152 (0.117–0.185)	0.382 (0.136–0.645)	0.342 (0.312–0.376)	0.397 (0.136–0.663)	0.451 (0.339–0.584)	0.381 (0.028–0.803)
T2 & CTC2	12	0.946 (0.907–0.975)	0.113 (0.001–0.278)	0.718 (0.665–0.783)	0.108 (0.004–0.259)	0.803 (0.677–0.893)	0.488 (0.259–0.759)
T1 & CTC2	26	*0.005 (0.001–0.009)	0.919 (0.783–0.994)	*0.009 (0.003–0.014)	0.892 (0.747–0.996)	*0.012 (0.003–0.028)	0.811 (0.461–1.000)
T2 & CTC1	12	0.650 (0.541–0.747)	0.051 (0.001–0.156)	*0.024 (0.016–0.033)	*0.011 (0.001–0.046)	*0.033 (0.009–0.110)	0.153 (0.017–0.339)

* significant support

(b) SNV data

		full dataset		50% data density		90% data density	
Groups	Cells	MPD	MNTD	MPD	MNTD	MPD	MNTD
T1	20	*0.001 (0.001–0.002)	0.419 (0.205–0.663)	*0.002 (0.001–0.005)	0.870 (0.643–0.979)	*0.003 (0.001–0.006)	0.274 (0.001–0.683)
T2	6	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	0.060 (0.001–0.239)	0.050 (0.001–0.269)
T3	20	*0.001 (0.001–0.002)	0.412 (0.192–0.626)	*0.001 (0.001–0.003)	0.669 (0.345–0.960)	*0.002 (0.001–0.005)	0.298 (0.001–0.693)
CTC1	6	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	0.063 (0.001–0.236)	0.050 (0.001–0.245)
CTC2	6	0.233 (0.206–0.259)	0.698 (0.560–0.853)	0.124 (0.003–0.205)	0.460 (0.254–0.597)	0.164 (0.061–0.242)	0.309 (0.097–0.581)
T1 & CTC1	26	0.516 (0.436–0.571)	0.132 (0.023–0.266)	0.630 (0.541–0.741)	0.455 (0.099–0.814)	0.655 (0.541–0.756)	0.554 (0.108–0.992)
T2 & CTC2	12	0.993 (0.979–1.000)	0.611 (0.342–0.789)	0.989 (0.958–1.000)	0.058 (0.001–0.167)	0.996 (0.989–1.000)	0.427 (0.040–0.755)
T1 & CTC2	26	*0.001 (0.001–0.002)	0.985 (0.954–1.000)	*0.001 (0.001–0.002)	0.811 (0.422–0.999)	*0.001 (0.001–0.002)	0.250 (0.001–0.687)
T2 & CTC1	12	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.001 (0.001–0.001)	*0.011 (0.001–0.048)	*0.011 (0.001–0.038)

* significant support

Overview of phylogenetic support for reduced datasets.

Table 10. Overview of tree support from reduced data sets. Mean and standard deviation of ultrafast bootstrap scores and posterior probabilities for the Maximum Likelihood and Bayesian trees respectively.

	Data density		
	full	50%	90%
Bootstrap			
Expression	77 ± 22	71 ± 25	78 ± 19
SNV	54 ± 23	56 ± 21	43 ± 21
Posterior			
Expression	75 ± 19	62 ± 27	28 ± 32
SNV	53 ± 30	48 ± 32	20 ± 28