BMC Genomics

RESEARCH Open Access

A common methodological phylogenomics framework for intra-patient heteroplasmies to infer SARS-CoV-2 sublineages and tumor clones

Filippo Utro*, Chaya Levovitz, Kahn Rhrissorrakrai and Laxmi Parida*

From The 19th Asia Pacific Bioinformatics Conference (APBC 2021)
Tainan, Taiwan. 3-5 February 2021

Abstract

Background: All diseases containing genetic material undergo genetic evolution and give rise to heterogeneity including cancer and infection. Although these illnesses are biologically very different, the ability for phylogenetic retrodiction based on the genomic reads is common between them and thus tree-based principles and assumptions are shared. Just as the different frequencies of tumor genomic variants presupposes the existence of multiple tumor clones and provides a handle to computationally infer them, we postulate that the different variant frequencies in viral reads offers the means to infer multiple co-infecting sublineages.

Results: We present a common methodological framework to infer the phylogenomics from genomic data, be it reads of SARS-CoV-2 of multiple COVID-19 patients or bulk DNAseq of the tumor of a cancer patient. We describe the Concerti computational framework for inferring phylogenies in each of the two scenarios.To demonstrate the accuracy of the method, we reproduce some known results in both scenarios. We also make some additional discoveries.

Conclusions: Concerti successfully extracts and integrates information from multi-point samples, enabling the discovery of clinically plausible phylogenetic trees that capture the heterogeneity known to exist both spatially and temporally. These models can have direct therapeutic implications by highlighting "birth" of clones that may harbor resistance mechanisms to treatment, "death" of subclones with drug targets, and acquisition of functionally pertinent mutations in clones that may have seemed clinically irrelevant. Specifically in this paper we uncover new potential parallel mutations in the evolution of the SARS-CoV-2 virus. In the context of cancer, we identify new clones harboring resistant mutations to therapy.

Keywords: Tumor evolution, Clonal evolution, Phylogeny, COVID-19

*Correspondence: futro@us.ibm.com; parida@us.ibm.com
IBM Research, T.J. Watson Research Center, Yorktown Heights, USA

Background

Deep sequencing genomic datasets contain intricate details that can be mined to reveal intra-patient heterogeneity present in disease states. The classic example that has been explored is the heterogeneity present in cancer, whether it be within a single tumor, across a patient's metastatic sites, or a tumor's evolution in response to treatment over the course of a disease. Interestingly, these same principles of heterogeneity can be explored in other scenarios that have similar sequencing data demonstrating different variant frequencies, including SARS-CoV-2 virus causing the COVID-19 infection. Evidence in several studies have highlighted the intra-host genomic diversity of SARS-CoV-2 [1–5]. As in cancer, the presence of different, heterogenic reads in a COVID-19 patient assumes the existence of multiple sublineages, or subclones, rather than the occurrence of recombination. The genetic evolution giving rise to heterogeneity is a common characteristic of all aspects of disease that contain genetic material, including cancer and infection. This common thread of increased heterogeneity involves many of the same processes. Once these assumptions are established, the same tools and methodologies that are used to analyze tumor heterogeneity can be applied with a level of confidence to SARS-CoV-2 datasets.

Implications of viral heteroplasmy in COVID-19 patients. The novel SARS-CoV-2 coronavirus that appeared in the city of Wuhan, China, in late 2019 has caused a large scale COVID-19 pandemic, spreading to more than 70 countries. Broad sequencing efforts have been made in an effort to understand the natural evolution of this virus. Several studies published with SARS-CoV-2 sequencing data reveal different viral allele frequencies in the same patient. The most likely explanation for the presence of intra-patient heterogenic viral reads is the existence of different viral strains rather than recombination since the probability of a fully functional single stranded virus emerging after entering a cell and its subsequent disassembly and reassembly into a virion with a different sequence is low [6]. Multiple viral strains infecting the same host has enormous clinical implications in terms of treatment, epidemiology, and the potential to overcome the pandemic and thus needs to be considered and analyzed. Variations in viral strains can harbor different resistance mechanisms, levels of transmissibility, response to therapy, and explain the large variation of symptomology. Even more important, treatment and vaccine success would rely on targeting the collection of strains present and not simply targeting one. It is for these reasons it is imperative that the research community consider the likely scenario that patients are coinfected with multiple strains.

Implications of heterogeneity in tumors of cancer patients. The presence of multiple tumor clones in the same patient has significant treatment implications. Multiple mechanisms of resistance can exist in separate clones [7]. Drug targets can 'disappear' or develop over time [8, 9]. Alternate pathways can be inhibited by the introduction of new alterations [10]. Even gross phenotypes can change due to underlying genomic changes [11]. Thus, it is imperative that we continue to monitor patient tumor evolution over the course of a disease in order to optimize treatment protocols. Parallels of tumor evolution have been drawn to that of human evolution and thus similar tools and algorithms are being applied and adjusted to analyze cancer [12]. Phylogenetic trees are being constructed to capture the change occurring during the disease while subclonal structure is identified and analyzed for clinically relevant changes. Several algorithms have been proposed to capture tumor evolution using single cells [13, 14] however these tools do not account for tumor heterogeneity and thus do not pick up on all clones present in a given tumor. Most algorithms consider bulk tumor samples which has the advantage of integrating genetic information from many tumor cells but are challenged by the need to deconvolve the mixture of clones present in any given biopsy [15–21]. Studies have also shown that determining which tree amongst the multitude that are possible is a non-trivial problem [22]. Several of these methods have been adapted for multi-site sample integration but are not specific for longitudinal data [16, 18, 19, 23]. More recently there have been several tools developed that do integrate longitudinal (multi-time) sampling [24, 25]. Although these models are more accurate, they still are limited by their inability to deal with samples with large mutational burdens and are not designed for multi-site samples.

Concerti overview An informative analysis for SARS-CoV-2 would require a method to be able to perform fine-grain evaluations with the ability to differentiate between viral sequences. In addition, the method would need to be able to analyze longitudinal data to capture when co-occurrence transpires. In cancer, sequential liquid biopsies over the course of disease are becoming more common given the ease of collection, lower cost, and greater ability to describe the complete disease profile vs. a subset of mutations present in distinct lesions. Therefore, it is imperative to establish tools that can manage/deconvolve mixed clonal samples, integrate multi-site and longitudinal sampling, and analyze large numbers of mutations with the same level of accuracy as low burden samples. We introduce Concerti[1], a tool for inferring disease evolution phylogenies, at genomic scales, from multiple sites and multiple longitudinal DNA sequencing samples. One of the unique features of Concerti is that

[1] A preliminary version of the algorithm handling only tumor phylogeny was presented at RECOMB-CCB 2020.

Table 1 Exemplar phylogenetic methods

Method	Input data	Data mode	CNV	Multi-time data	Multi-site data	Multi-patient data	Genomic scale	Time-scaled tree
SCITE [13]	single cell	single value	✗	✗	✗	✗	N/A	✗
Pyclone [20]	bulk data	single value	✗	✗	✓	✗	✗	✗
CITUP [18]	bulk data	single value	✗	✗	✓	✗	✗	✗
Calder [25]	bulk data	single value	✗	✓	✗	✗	✗	✓
VERSO [4]	bulk data	single value	✗	✗	✗	✓	✓	✗
BEAST [26]	bulk data	single value	✗	✗	✗	✓	✓	✗
ClonalTREE [27]	bulk data	single value	✗	✓	✓	✗	✓	✗
PhylogicNDT [24]	bulk data	dist.	✓	✓	✓	✗	✓	✗
Concerti	bulk data	dist.	✓	✓	✓	✓	✓	✓

it generates *time-scaled trees*, i.e., trees aligned to actual time scales that capture not only the birth and death of clones, but also acquisition of alterations within the same clone. Concerti uses almost exclusively discrete optimization methods and has the flexibility to provide multiple possible solutions suggested by the patient data. To help with the interpretation of the results, the solutions are ordered by decreasing likelihood. Due to the absence of benchmark data, it is hard to perform a precise comparison of the different tools reported in literature. We provide in Table 1 a succinct summary of the capabilities of eight classes of exemplar tools and highlight the elements of uniqueness in each approach.

We demonstrate the accuracy of Concerti by reproducing and expanding on known results from literature. In particular, we confirmed the results reported in [4] for the viral evolution model and expanded on them by discovering new homoplasies. While for the tumor evolution model using whole-exome sequencing data from patients [7, 28], Concerti constructs phylogenetic trees that accurately describe a tumor's evolution while simultaneously highlighting new post-treatment subclones that likely confer resistance and may serve as new potential drug targets.

Method

Viral Evolution Model. For computational purposes, we assume that all the virions of the same lineage have the same set of alterations (with respect to a reference). Since there is evidence of intra-patient variations with a wide

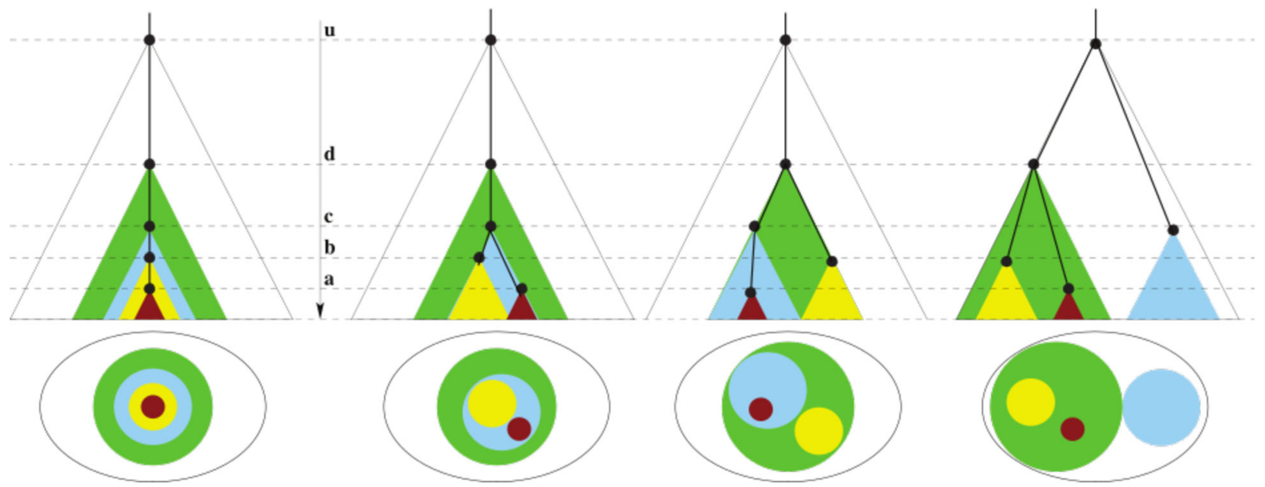

Fig. 1 Let U (white), D (green), C (cyan), B (yellow), A (brown) be pseudoclones with prevalence values: 1.0 ≥ $u > d > c > b > a > 0$ respectively. The top row shows 4 possible evolution trees where the time axis is the *molecular clock*. The bottom row shows the single time-point "fishplot" as appropriately stacked disks. Notice that the leftmost phylogeny suggests that there exist some cells/virion with both A and B alterations while all the other three suggest that there exists no such cell/virion

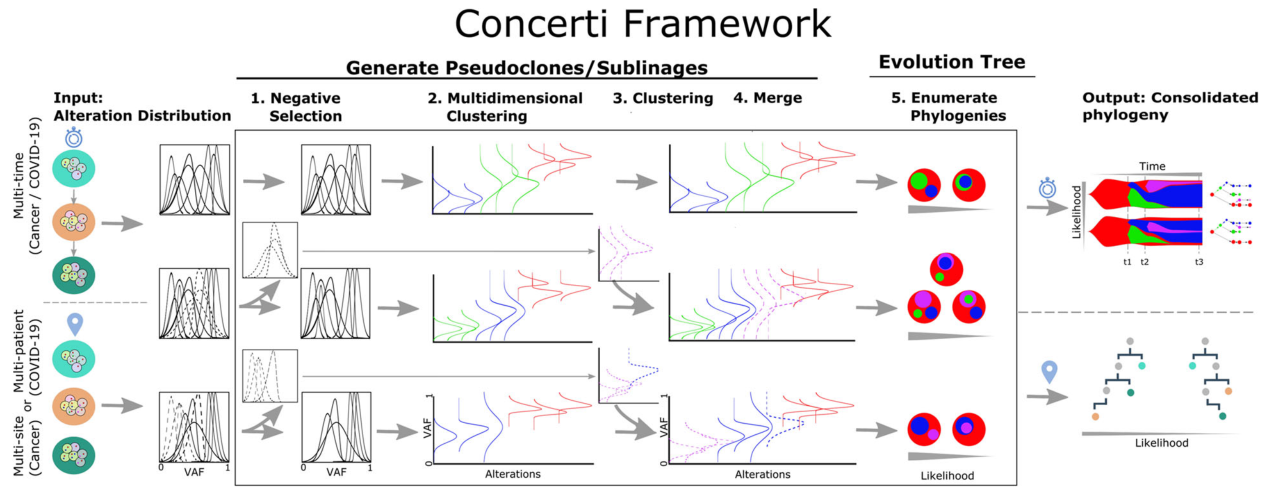
Concerti Framework
Generate Pseudoclones/Sublinages
Evolution Tree
Input: Alteration Distribution
1. Negative Selection
2. Multidimensional Clustering
3. Clustering
4. Merge
5. Enumerate Phylogenies
Output: Consolidated phylogeny
Multi-time (Cancer / COVID-19)
Multi-patient (COVID-19) or Multi-site (Cancer)
VAF
0
1
Alterations
Likelihood
Time
t1
t2
t3

Fig. 2 Schematic of the Concerti Framework. Given a set of multi-patient (COVID-19) or multi-site, multi-time (cancer) genomic samples, the algorithm analyzes the underlying alteration frequency distribution as input and performs a (1) negative selection to filter *appearing* alterations. A (2) multidimensional clustering is done to identify pseudoclones/lineages that will then be enriched by a (3) single sample clustering that (4) merges alterations that were initially negatively selected. (5) All potential phylogenies are generated and assessed for compatibility according to Definition 1 . Finally the set of consolidated phylogenetic structures over time or site are output with likelihood scores

range of allele frequencies [4, 6], we postulate that there is heteroplasmy due to possibly multiple sublineages evolving in this micro-environment. Since the coronavirus is a non-segmented positive RNA virus, we further postulate that it is very unlikely that any recombination occurs during the virus's life cycle: attachment and entry, replicase protein expression, replication and transcription, assembly and release [29].

Tumor Evolution Model. We assume the following model: tumors arise from an altered cell, accumulating additional alterations over time. These changes give rise to populations of cells termed in literature as *clones*. For computational purposes, we assume that all the cells in a clone have the same set of alterations. Furthermore, these clones may alter further over time. Thus multiple clones co-exist in a tumor and some may have an evolutionary advantage over the others within the tumor environment, allowing for growth or shrinkage of a clone over time.

Terminology. The absence of recombination and the accumulation of variants over time are the two salient factors that facilitate a common methodology for inferring evolution in both models. Furthermore for the tumor evolution model, the inferencing may be based on single or multiple DNA sequencing samples: the latter can be *multi-time*, i.e., at multiple timepoints, or can be *multi-site*, i.e., from different lesions possibly collected at the same time. We use the term *data point* for multi-patient (COVID-19) and multi-time, multi-site (cancer). The term *alteration* is applicable to any genetic event including, but not limited to, mutation, single nucleotide variant, copy number variant, etc. In this manuscript CCF (Cancer Cell Fraction) denotes the fraction of cancer cells bearing an alteration in a cancer sample [30]. For the purposes of our algorithm, CCF and VAF (Variant Allele Frequency) are indistinguishable and the precise method of determining alteration frequencies is outside the scope of this paper. For clarity of exposition we use VAF to represent VAF or CCF and SNV to represent all alterations.

Furthermore, in the context of cancer, it is important to note the distinction between *clones* and *pseudoclones*. Clone is a biological entity described as a population of indistinguishable cells. For our purposes, the nuclear DNA is identical for the population and thus a clone can be defined by a set of SNV's. A pseudoclone on the other hand is a subset of these SNVs. In practice they are a maximal collection of SNVs with identical (or similar) VAF values [19, 20, 25]; this value is termed *prevalence* in this paper. This collection of SNVs is meaningful under the assumption that identical VAF values implies these SNVs co-occur in a cell. Note that the converse is not necessarily true, i.e., multiple SNVs within a cell may have varied VAF values. Thus pseudoclone is an algorithmic artefact while a clone is simply the union of some finite pseudoclones. For example in Fig. 1, the distinct colors denote the different pseudoclones, but the biological clone is the union of all the subclones in the path to the root of the evolutionary tree. Hence the (biological) brown clone in the leftmost tree is actually the union of the SNVs that define the brown pseudoclone, the yellow pseudoclone, the cyan pseudoclone and the green pseudoclone. Hence

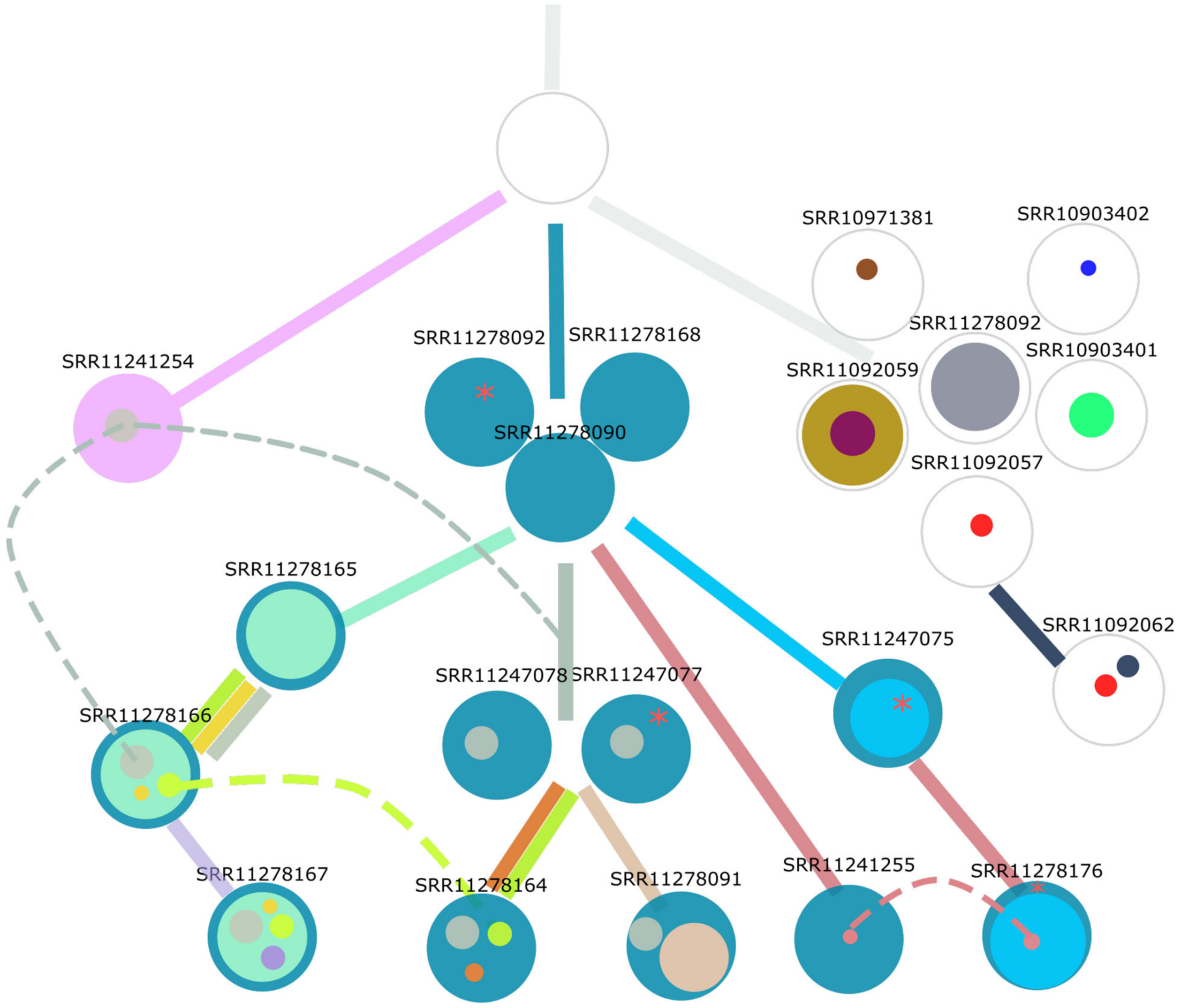

Fig. 3 The 21 COVID-19 patients are shown at different internal and leaf nodes in the phylogeny as stacked disks of different colors. Each color corresponds to a distinct sublineage identified by a set of alterations and the size is roughly proportional to its observed prevalence value. Where possible, the edges of the phylogeny are colored by the emerging sublineage(s). When a node has multiple individuals, it indicates that there is not enough evidence to delineate the distinctions in the phylogeny. The three homoplasies (parallel mutations) are shown by dashed transversal lines. While in two (raspberry, green colors) the alteration event occurred at least twice, in the third (gray color) the alteration occurred at least three times. Furthermore, if the date of collection of a sample at a child node precedes the date at a parent node, it is within a window of a week

for mathematical preciseness, we use the term pseudoclones in the Method sections, yet to avoid clutter, we use the term clone in place of pseudoclone in the Results and discussion section.

With a slight abuse of terminologies, a pseudoclone and clone corresponds to a sublineage and lineage, respectively, in the context of virions. To avoid clutter, we use the terms sublineage and lineage interchangeably.

Method assumptions

We make the following assumptions.

Assumption 1 ***[Infinite Sites Model]*** *A majority of the alterations satisfy the following:*

1. *irreversible, i.e., once the alteration occurs the reverse of turning it back to its original state does not occur (no back mutation).*
2. *unique, i.e., the same alteration does not occur elsewhere in the tumor (no parallel mutation).*

The topology of evolution is a tree. Although Occam's Razor Principle suggests the perfect phylogeny assumptions used most commonly in literature [31], it is important to note that some exceptions to this property of alterations may occur in practice, especially when modeling biology. One such example is the presence of parallel mutations known as homoplasy. In order to capture this natural phenomenon in our trees, we handle this violation of perfect phylogeny as an exception in our algorithm.

Assumption 2 ***[Alteration distribution]*** *Most of the alterations follow i.i.d. (uniform) distribution.*

Again, for algorithmic purposes, it is reasonable to assume that tumor clones would follow the same principles as the individual alteration. But a clone, unlike an alteration, may die, i.e., may be selected against and overrun by other clones. So a clone may change in composition over time, i.e. more alterations can be added to the clone (but, not removed due to Assumption 1). Various selection pressures are in effect on the different clones, whose effect is manifested in the size of the clones: the clone can either grow or shrink in size reflected as an increasing or decreasing VAF value respectively. Thus the following:

Assumption 3 ***[Tumor Clone dynamics]*** *Over time, a clone may*

1. *change in composition / size (additional alterations but not lose alterations)*
2. *change in prevalence values (increase or decrease)*
3. *die or a new clone may be born.*

Method overview

Input The input may come as one of two forms for both SNVs and CNVs. Single value VAF and CCF are taken as a matrix of data points (multi-patient, multi-time, or multi-site) by SNV . CCF distributions are received as a dataframe with data point, SNV, and CCF distribution discretized into x bins ($x = 100$), which relates to the confidence associated to the CCF by the originating algorithm. All values are continuous [0,1]..

See Fig. 2 for an overview of Concerti. Based on our assumptions, the method has two major phases.

1. **Phase I**. We first identify the pseudoclones across all the data-points. However, the pseudoclones are not identical due to the clonal dynamics (Assumption 3).
   - **a.** Due to Assumption 3(3), we gate the alterations that are present in all the samples. This results in some alterations being filtered out and refer to this step as Negative Selection in the outline. We cluster these filtered alterations separately for each sample.
   - **b.** The pseudoclones are preserved in the samples, albeit with some dynamics (Assumption 3(2)). We carry out a multi-dimensional clustering, across all samples, based on the values or distributions of the alterations.
   - **c.** We appropriately merge the clusters of the above two steps to obtain the pseudoclones based on the similarity between the pseudoclone prevalence of the existing, multi-dimensional clusters from *step b* and with the clusters from *step a* in the appropriate samples.
2. **Phase II**. We first deduce the phylogenies of each sample separately and then we relate them with each other.
   - **a.** The sizes of the pseudoclones in each data point admits possibly multiple phylogenies. We enumerate the admissible phylogenies associating a probability with each based on Assumption 2.
   - **b.** Next we consolidate the trees from the multiple data points. This captures the topology as well as the clonal dynamics.

   Concerti offers two types of visualizations: one that captures the change-in-composition dynamics of the clones (as a tree) and the other that captures the change-in-size and birth/death of clones (fishplots [32]). The multiple possible solutions suggested by the data is output in decreasing order of probabilities to ease interpretation.

Exception Handling. Real data is sometimes notoriously perplexing, either due to errors in sequencing or CCF distribution estimation or simply the infraction of some of the assumptions enumerated in the last section. In practice, for intractable cases we handle the exceptions by relaxing the minimum of assumptions by consulting with the domain experts.

Phase I: generate pseudoclones/Sublineages

We define the distance between a pair of CCF distribution g_1 and g_2 as

$$\mathrm{dis}(g_1, g_2) = \int_{-\infty}^{+\infty} |g_1(r) - g_2(r)| dr.$$

However, for algorithmic efficiency, we use the similarity measure defined as

$$\mathrm{sim}(g_1, g_2) = 1 - \frac{1}{2}\mathrm{dis}(g_1, g_2).$$

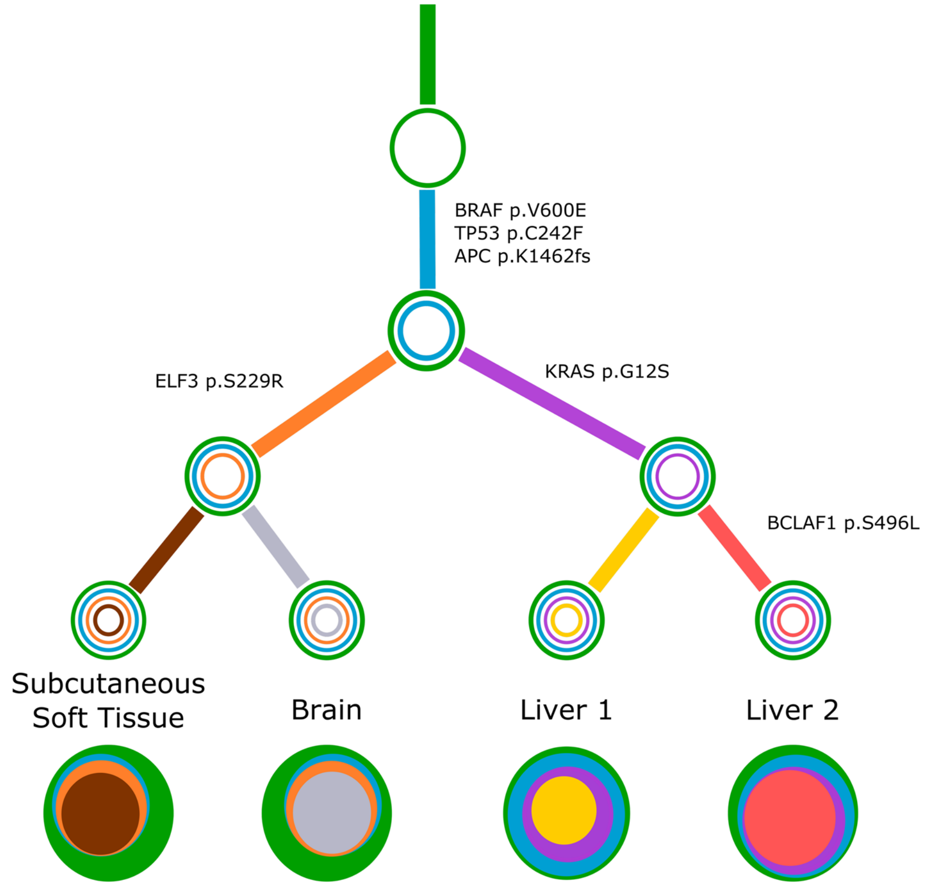

Fig. 4 Concerti fishplot and tumor evolution tree **T** for patient CLL1 multi-time data. The fishplot width corresponds to approximate tumor size using ALC (absolute lymphocyte count) values. Clones are colored and sized proportionally to their prevalence. The corresponding tumor tree is aligned by timepoint and highlights to the birth of brown clone, which occurred prior to the annotation of clinical relapse. Node sizes correspond to prevalence. The edges of the **T** are labeled by the known cancer genes and the colors denote the distinct pseudoclones estimated by Concerti. Red asterisks indicates acquisition of new alterations to clone

Thus identical distributions have a similarity of 1 while distinct distributions have similarity value zero. In practice, since the probability density function is specified as a discrete set of pairs of values and its probability, we compute the similarity as follows:

$$\sum_{r=l}^{u} \min(g_1(r), g_2(r)), \quad (1)$$

where $[0 <= l, u <= 1.0]$ is the maximal interval where both g_1 and g_2 have non-zero values.

We first perform a negative selection where alterations not present in all samples are removed. Let S_a be the set of alterations present in all samples and S_b the removed alterations. We carry out a hierarchical clustering [33] of the alteration set S_a using the similarity function (1). We cluster the group of alterations in $S_b$that are present in the same set of samples separately and then merge with the multi-dimensional clustering of S_a to produce the pseudoclones. The *prevalence* of a pseudoclone is approximated by the mean of the mean value of each constituent alteration (CCF) distribution. We merge the clusters from S_a and S_b to obtain the final set of pseudoclones if the similarity between the clusters is less than *thd* (e.g. in this manuscript we used $thd = 0.1$) for all respective samples.

Phase II: generate evolution tree(s) of pseudoclones/Sublineages

Enumerate Admissible Trees. Assume without loss of generality (WLOG) u=1.0. Prevalence values $0 \leq a, b \leq 1$ can be viewed as some sub-intervals (*sticks*) of $[0, 1]$ of lengths a and b respectively. Then in a cell-population realization either sticks A (with prevalence value a) and B (with prevalence value b) are nested or disjoint but may not straddle. To remove possible computational artefacts, pseudoclones with prevalence $v < 0.05$ and less than 3 SNVs for all samples are discarded. Finally, given a pseudoclone A with a prevalence value a, then for each x pseudoclone nested directly in A with prevalence v_x, the sum of their prevalence is $\sum_x v_x \leq a$. Figure 1 shows a simple example. With bulk-sequencing (i.e. no viral isolates or single-cell), the multiple possible scenarios cannot really be teased apart. But using Assumption 2, the probability of each admissible tree can be estimated.

We present a recursive algorithm, called Stick-Stack (Algorithm 1), that enumerates all possible ways of stacking the sticks (or sub-intervals) corresponding to pseudoclones. The algorithm computes in steps 4-14 the probabilities of the trees using Assumption 2. In steps 15-19 all possible trees are generated as follows. Assuming we have already seen j pseudoclones and need to assign $j+1$, we iterate all possible j pseudoclones and determine whether the v_{j+1} can be added as child. If yes we adjust the prevalence of v_i and recursively call the algorithm to analyze the next pseudoclone. In the output of Stick-Stack (A, B) denotes nesting of B in A. Stick-Stack call is initiated as $(1, tr = \emptyset, pr = 1.0, v_1 > ... > v_k > 0)$ with the n prevalence values of the k pseudoclones and the output is a set of trees where each tree tr is a collection of (parent,child) pairs, with probability pr. It is easy to verify from the algorithm description below that the probabilities of all the admissible trees sum up to 1.

Mutual Comparison. While a single data point may suggest the relative relationship between pseudoclones, clonal dynamics can only be captured from multiple data points, be it multi-time or multi-site.

Observation 1 ***[Clone Dynamics]*** *Based on Assumption 3 the clusters of filtered alterations of Step 1 of Phase I provide the clone dynamics.*

a. *If such a cluster merges with the multidimensional cluster, then this indicates a change in composition of the pseudoclone and provides labels for the edges of the evolution phylogeny.*
b. *If a new cluster is generated (i.e., it does not merge with clusters from the multi-dimensional clustering) then this indicates the birth of a new pseudoclone.*

A clone acquiring new alterations (case a. above) is shown as asterisk in the COVID phylogeny in Fig. 3 or tumor phylogeny in Figs. 4 and 5 . The birth and death of clones (case b. above) are also illustrated in the latter two figures.

The mutual comparisons reveal whether the different samples are related or independent. When related, it is possible to reconstruct *consolidated* tree(s) that capture the evolution across the multiple data points.

```
Input: $v_0 > v_1 > \ldots > v_j, tr, pr, v_{j+1} > \ldots > v_c$
Result: Compute all possible configurations for a
        given sample
1  if $v_{j+1}$ == 0 then
2      return tr, $pr$;
3  else
4      $tot = 0$ ;
5      for $i = 0 \ldots j$ do
6          if $v_i \geq v_{j+1}$ then
7              $tot = tot + v_i$ ;
8          end
9      end
10     for $i = 0 \ldots j$ do
11         if $v_i \geq v_{j+1}$ then
12             $pr_i = v_i / tot$ ;
13         end
14     end
15     for $i = 0 \ldots j$ do
16         if $v_i \geq v_{j+1}$ then
17             $v_i = v_i - v_{j+1}$;
18             Stick-Stack($v_0 > v_1 > \ldots > v_{j+1}$,
                tr+$(v_i, v_{j+1}), pr * pr_i, v_{j+2} > \ldots > v_c$);
19         end
20     end
21 end
```

Algorithm 1: Stick-Stack: Given the prevalences, the algorithm enumerates all possible admissible trees, each with an estimated probability of occurrence.

Stick-Stack algorithm produces each tree as a set of two-tuples corresponding to each edge as (parent,child), where the parent and child are both pseudoclones. Formally, if there exist $k > 0$ parent-child pairs as $(C_0, C_1), (C_1, C_2), ..., (C_{k-1}, C_k)$, then C_0 *precedes* C_k or $C_0 \prec C_k$. Let (-, C_i) denote that C_i has no parent.

Definition 1 ***[Incompatible]*** *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two trees with three sets of (possibly empty) pseudoclones: A_i that occur in both $\mathcal{T}_1$ and $\mathcal{T}_2$; D_i that occur in $\mathcal{T}_1$, but not in $\mathcal{T}_2$, and, B_i that occur in $\mathcal{T}_2$ but not in $\mathcal{T}_1$.*
$\mathcal{T}_1$ and $\mathcal{T}_2$ are incompatible if at least one of the following conditions does not hold:

1. *WLOG if $A_1 \prec A_2$ in $\mathcal{T}_1$ then $A_1 \prec A_2$ in $\mathcal{T}_2$.*
2. *WLOG each D_i is of the type (-, D_i) and if D_i has a child then it occurs as (D_i, D_j) in $\mathcal{T}_1$.*
3. *WLOG each B_i is of the type (-, B_i) and if B_i has a child then it occurs as (B_i, B_j) in $\mathcal{T}_2$.*

Once all possible configurations of the pseudoclones for any given time point are generated independently, the next step is to extract the possible compatible trees between all time points.

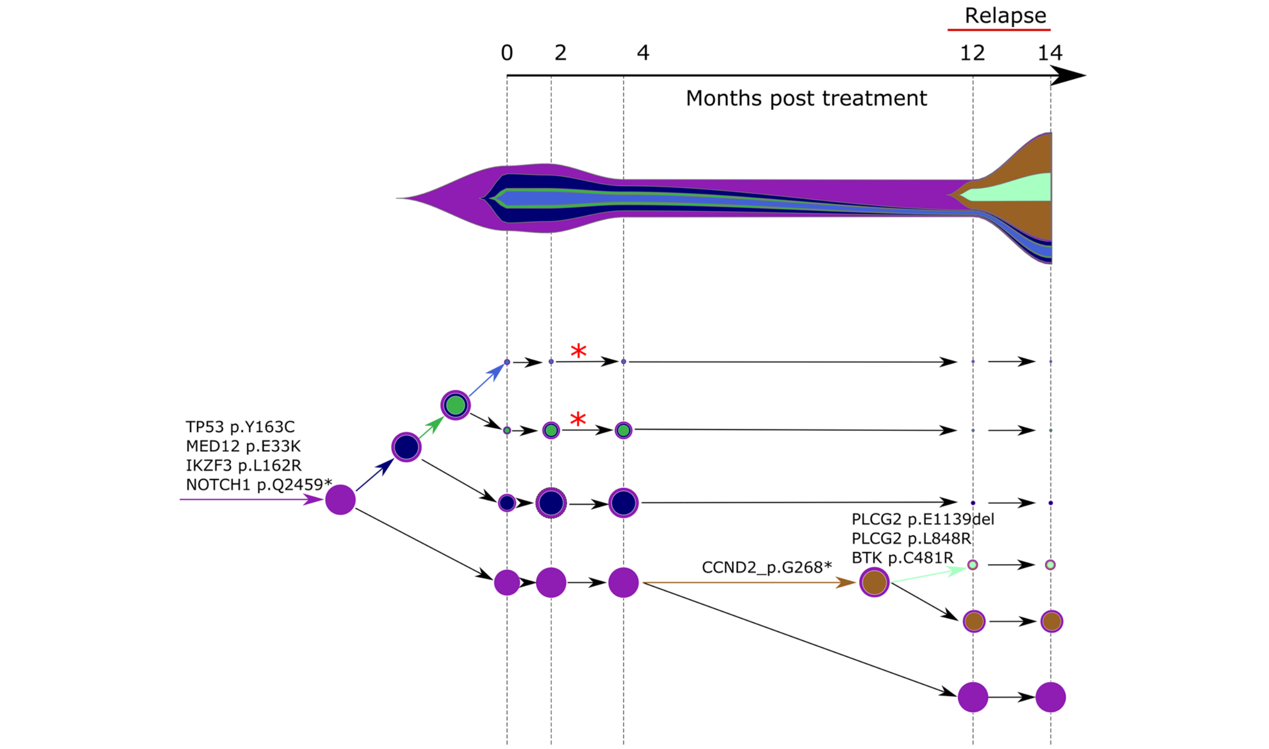

Fig. 5 Concerti fishplot and tumor evolution tree **T** for patient CLL2 multi-time data. The fishplot width corresponds to approximate tumor size using ALC values. Clones are colored and sized proportionally to their prevalence. The corresponding tumor tree is aligned by timepoint and highlights to the birth of brown clone, which occurred prior to the annotation of clinical relapse. Node sizes correspond to prevalence. The edges of the **T** are labeled by the known cancer genes and the colors denote the distinct pseudoclones estimated by Concerti. Red asterisk indicates acquisition of new alterations to clone

Definition 2 ***[Consolidated phylogeny]*** *Let $E(\mathcal{T})$ be the set of the two tuples (parent,child) of $\mathcal{T}$. If $\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_K$ are mutually compatible, then* **T***, the consolidated phylogeny of the K trees, is defined by the following set*

$$E(\mathbf{T}) = \cup_{k=1}^{K} E(\mathcal{T}_k).$$

Notice that by conditions 2 and 3 of the incompatible definition above, **T** does not have any nodes with multiple parents and **T** is a tree.

Let $\mathcal{T}_{i,j}$ be the jth compatible tree at data point i with probability $p_{i,j}$ as estimated by Stick-Stack. Then the relative probability of a compatible evolution tree $\mathbf{T}^k = (\mathcal{T}_{1,j_1}, \mathcal{T}_{2,j_1}, ..., \mathcal{T}_{n,j_1})$ over the n datapoints is given by

$$\mathbb{P}(\mathbf{T}^k) = \frac{p_{1,j_1}}{\sum_k p_{1,k}} \times \frac{p_{2,j_1}}{\sum_k p_{2,k}} \times \ldots \times \frac{p_{n,j_1}}{\sum_k p_{n,k}}. \qquad (2)$$

Note that $\sum_k \mathbb{P}(\mathbf{T}^k) = 1$ where k is over all possible compatible configurations. Thus the probability of a $\mathbf{T}^k$ may be underestimated. However, it preserves the ordering of the possible multiple solutions which is used here.

For a concrete example, consider Fig. 6. The four sites are labeled as subcutaneous soft tissue (subcu), brain, liver1 and liver2. For each site, Concerti produces exactly one tree, with eight pseudoclones across all the four sites: green (G), cyan (C), orange (O), purple (P), ash (A), yellow (Y), red (R), brown (B). Then Stick-Stack produces the following four trees:

$$\begin{aligned}
\mathcal{T}_{\text{subcu}} &= \{(\text{G,C}), (\text{C,O}), (\text{O,B})\},\\
\mathcal{T}_{\text{brain}} &= \{(\text{G,C}), (\text{C,O}), (\text{O,A})\},\\
\mathcal{T}_{\text{liver1}} &= \{(\text{G,C}), (\text{C,P}), (\text{P,Y})\},\\
\mathcal{T}_{\text{liver2}} &= \{(\text{G,C}), (\text{C,P}), (\text{P,R})\}.
\end{aligned}$$

It can be verified that the four trees are mutually compatible. Then the unique consolidated phylogeny is given by

$$\begin{aligned}
\mathbf{T} &= \mathcal{T}_{\text{subcu}} \cup \mathcal{T}_{\text{brain}} \cup \mathcal{T}_{\text{liver1}} \cup \mathcal{T}_{\text{liver2}}\\
&= \{(\text{G,C}), (\text{C,O}), (\text{O,B}),(\text{O,A}),(\text{C,P}), (\text{P,Y}),(\text{P,R})\}
\end{aligned}$$

T has only one connected component suggesting that all the data points are genetically related.

In multi-time data, a consolidated tree **T** is stretched out with the given time points appropriately marking the tree (see Figs. 4 and 5 for example). Additionally, a fishplot is output to visualize the dynamics of the pseudoclones (growth or shrinkage, including birth and death).

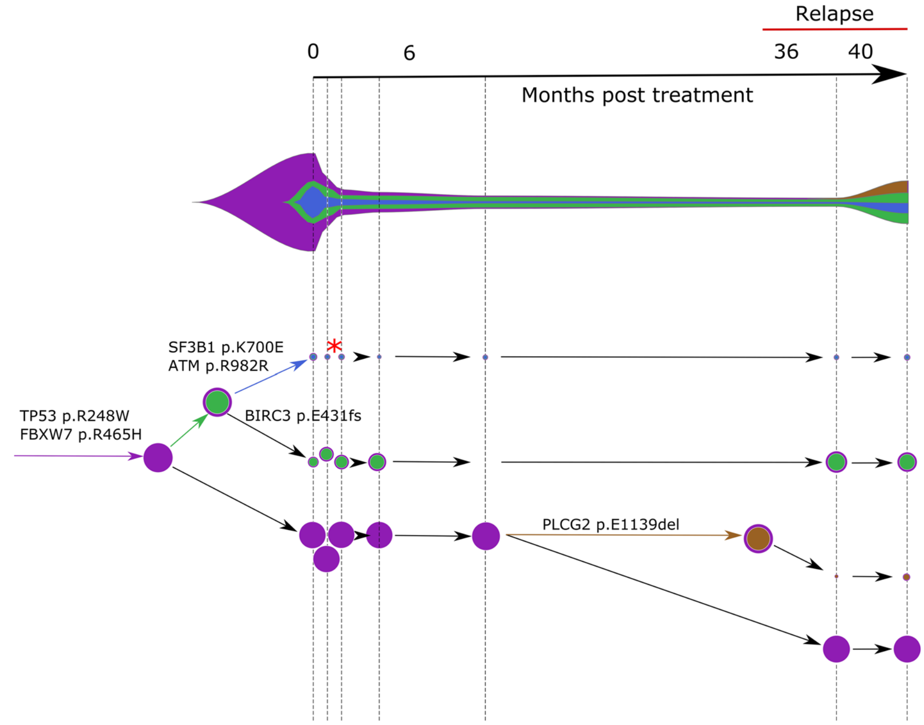

Fig. 6 Concerti tumor evolution tree **T** for patient GI1. Tumor evolution tree **T** for colon cancer patient GI1 multi-site data. The edges of the **T** are labeled by the known cancer genes and the colors denote the distinct pseudoclones estimated by Concerti. Leaf nodes represent each of the distinct lesion sites. The single site trees $\mathcal{T}$ are shown at the bottom as stacked discs and the sizes are proportional to the prevalence values

Results and discussion

We applied Concerti on publicly available COVID-19 sequencing data as well as multi-site and multi-time cancer sequencing data (see Availability if data and materials section).

COVID-19 data

For our study we sought COVID-19 patient samples with access to the raw reads in order to assess the alterations at varying allele frequencies. Using an established reference MN908947.3 obtained from the 'first' patient sequenced in Wuhan, China, we found 41 distinct variants in 21 patient samples (see the Availability if data and materials section for details on the patient samples and the variant calling pipeline). 18 of the patients were also analyzed in [4], albeit the variants therein were derived based on a different reference sequence and protocol. Hence, the set of variants do not exactly match the set we obtain. Our data set has three additional patients from the Wuhan seafood market [34]. We applied Concerti to the data and the resulting phylogeny is shown in Fig. 3. Note that all samples with dark blue sublineage in the figure were collected in the USA, the one with the dark-grey sublineage (SRR11278092) was collected in Nepal, while the remaining were collected in China. The figure shows the other lineages that were identified. The phylogeny is not fully resolved based on the variants of this set of patients; this is shown as clusters of patients in two internal and one leaf node in the tree. The phylogeny also uncovers three parallel mutation events (shown as dashed lines with the corresponding color): 404:A>T (raspberry), 29039:A>T (grey) and 4229:A>C (green). The first two were reported in [4] while the third is discovered in this study.

Cancer data

Using Concerti, we analyzed three patients,two sampled over time (Figs. 4 and 5) one sampled across multiple sites (Fig. 6). We first applied Concerti to longitudinal sequencing data from two relapsed chronic lymphocytic leukemia (CLL) patients. Patient CLL1 had five biopsies taken over the course of treatment with ibrutinib and rituximab and relapsed 12 months after treatment initiation (Fig. 4). Before treatment, the dominant clone contained mutations in several known cancer genes including TP53, MED12, IK2F3, and NOTCH1. Two small clones (red asterisks) continued to evolve as evidenced by the acquisition of additional mutations after two months post-treatment. The fishplot highlights the correspondence between the emergence of a resistant clones and the increase in tumor size. Concerti's time-scaled phylogenetic tree and fishplot captures the birth of this clone, before relapse was clinically documented, that harbored three mutations in genes associated with resistance to ibrutinib, including BTK, PLCG2, and known cancer driver CCND2.

The second CLL patient Concerti analyzed was similarly treated with and developed resistance to ibrutinib (Fig. 5). For patient CLL2, seven blood biopsies were taken over the treatment course including before-treatment, on-treatment, and at time of relapse. Several truncal mutations in known cancer genes were identified in the pre-treatment samples, including TP53 and FBXW7. After initiation with ibrutinib, a clone with a BIRC3 mutation (green) increased in prevalence. At the time of relapse, Concerti's phylogenetic tree identifies the emergence of a new clone harboring a mutation in PLCG2, a known mechanism of resistance to ibrutinib therapy, and which goes on to grow in prevalence. Clones with ATM and SF3B1 did not have noticeable clonal dynamics during the treatment or relapse intervals suggesting they are not selected for under ibrutinib therapy. The interested reader is referred to Additional file 1 for a comparison between Concerti, CITUP [18], and Calder [25] where we show how Concerti outperforms the other methods for these two patients. In both CLL patients, the birth of these resistant clones in response to treatment was only able to be identified because of Concerti's unique integration of time-scaled trees.

We then applied Concerti to a multi-site case, GI1, a 53 year old male with metastatic colon cancer who was part of a rapid autopsy study where multiple metastatic samples were taken at the time of death. Additional clinical details and the description of the sequencing method can be found in [7]. Samples were taken from different anatomical sites including lesions in the liver, brain, and subcutaneous soft tissue. Time of lesion development was not documented radiologically and thus no longitudinal time-ordering of the samples could be performed. Concerti's generated tumor evolution tree gives a clinically plausible explanation as to the mutational development of this disease and is supported by the original study's PhylogicNDT trees and their clinical findings offering a measure of validation (Fig. 6). The phylogenetic tree characterizes several truncal clones shared across all samples (green and cyan) and then identifies two sibling clones (orange and purple) that are tissue specific. One clone captures both liver samples and contains the KRAS p.G12S allele. The other clone, which contains the ELF3 p.S229R allele, goes on to develop two daughter clones each specific to the brain or subcutaneous soft tissue. Thus, Concerti's integration of multi-site samples enables the phylogenetic tree to capture a tumor's broad spatial heterogeneity and allows for a treatment course to be designed to be locally or broadly targeted.

Conclusion

In this paper we introduce Concerti, an algorithm for inferring evolutionary phylogenies. Concerti's ability to extract and integrate information from multi-point, whether multi-site, multi-time, or combination thereof samples, enables the discovery of clinically plausible phylogenetic trees that capture the heterogeneity known to exist both spatially and temporally. These models can have direct therapeutic implications since they can highlight: "births" of clones that may harbor resistance mechanisms to treatment, "death" of subclones with drug targets, and acquisition on functionally pertinent mutations in clones that may have seemed clinically irrelevant. By considering a phylogenetic analysis that steps back from the original disease context, novel relationships can be discovered before re-contextualization and interpretation in the patient context and highlights a strength of Concerti's applicability across biological contexts. We demonstrate in this paper how Concerti can be applied to any genomic sequencing dataset with varying allele frequencies, whether it be cancer or the new SARS-CoV-2 virus causing the COVID-19 pandemic, and the results can have profound disease-specific clinical implication.

Identifying the presence of multiple viral strain infecting a single host can have significant impact on how we approach treatment, vaccine development, and mitigation strategies. The results for COVID-19 patients demonstrate Concerti's ability to distinguish between viral strains based on difference allele frequencies and discover the presence of new homoplasies. Thus, Concerti's results addresses the overwhelming challenges researches face when developing therapeutics and may help facilitate the key to effective vaccine development. Accurately monitoring tumor evolution over the course of a disease can lead to the identification of new drug targets and therapeutic approaches that can stabilize this complex disease and manage the selective pressures introduced by treatment exposure and tumor-environment changes. These results for patients CLL1, CLL2 and GI1 demonstrate how Concerti's specific integration of multi-point data can facilitate better treatment plans that can both be more locally targeted and optimized for treatment responsivity.

Abbreviations

SNV: Single-nucleotide variant; CLL: Chronic Lymphocytic Leukemia; VAF: Variant Allele Frequency; CCF: Cancer Cell Fraction; ALC: absolute lymphocyte count

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07660-9.

Additional file 1: Experimental comparison with other tumor phylogeny reconstruction methods.

Acknowledgements

We are very grateful to Ignaty Leshchiner, Gad Getz, and Catherine J. Wu for providing CLL patient data.

About this supplement

This article has been published as part of BMC Genomics Volume 22 Supplement 5 2021: Selected articles from the 19th Asia Pacific Bioinformatics Conference (APBC 2021): genomics The full contents of the supplement are available at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-22-supplement-5.

Authors' contributions

FU and LP designed this study. FU implemented Concerti and perform the experiments. FU, CL and KR performed the analysis on the experimental data. All authors have wrote, read and approved the manuscript.

Funding

The author(s) received no specific funding for this work. Publication costs are funded by IBM Research.

Availability of data and materials

All data used in the paper are available with the original publications. In particular, patient CLL1 and CLL2 correspond to B06 and A43 of [28], respectively. While the GI1 colon cancer patient used for the multi-site analysis was previously published as TPS037 in [7]. The COVID-19 patient data come from 5 NCBI BioProjects: PRJNA601736, PRJNA603194, PRJNA610428, PRJNA605983 and PRJNA608651. The reads were trimmed with Trimmomatic (v. 0.39) and then mapped with bwa on MN908947.3 GenBank sequence. This sequence was taken in December from a 57 year old woman, who sold shrimp at the Wuhuan seafood market, appears to be the earliest case with COVID-19. Variant calling was performed generating mpileup files using SAMtools and then running VarScan (min-var-freq parameter set to 0.01). Finally to remove possible sequencing artifacts, we retain SNV that: show a VarScan significance p-value <0.05 (Fisher's Exact Test on the read counts supporting reference and variant alleles) and VAF >10%, resulting in a list of 41 SNVs in 21 patients that are used in the paper. Concerti's binaries are available at https://github.com/ComputationalGenomics/Concerti.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors are listed (together or partially) as co-inventor of 16/022088 patent application currently pending review at the USPTO and US patent 20200075170 related to the Concerti.

Received: 25 April 2021 Accepted: 28 April 2021
Published online: 16 November 2021

References

1. Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, Mentis A. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. bioRxiv. 2020. https://doi.org/10.1101/2020.03.27.009480. https://www.biorxiv.org/content/early/2020/03/28/2020.03.27.009480.full.pdf.
2. Korber B, Fischer W, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Foley B, Giorgi E, Bhattacharya T, Parker M, Partridge D, Evans C, de Silva T, LaBranche C, Montefiori D. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020. https://doi.org/10.1101/2020.04.29.069054. https://www.biorxiv.org/content/early/2020/04/30/2020.04.29.069054.full.pdf.
3. Lu J, Plessis LD, Liu Z, Hill V, Kang M, Lin H, Sun J, Francois S, Kraemer MUG, Faria NR, McCrone JT, Peng J, Xiong Q, Yuan R, Zeng L, Zhou P, Liang C, Yi L, Liu J, Xiao J, Hu J, Liu T, Ma W, Li W, Su J, Zheng H, Peng B, Fang S, Su W, Li K, Sun R, Bai R, Tang X, Liang M, Quick J, Song T, Rambaut A, Loman N, Raghwani J, Pybus O, Ke C. Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. medRxiv. 2020. https://doi.org/10.1101/2020.04.01.20047076. https://www.medrxiv.org/content/early/2020/04/04/2020.04.01.20047076.full.pdf.
4. Ramazzotti D, Angaroni F, Maspero D, Gambacorti-Passerini C, Antoniotti M, Graudenzi A, Piazza R. Characterization of intra-host SARS-CoV-2 variants improves phylogenomic reconstruction and may reveal functionally convergent mutations. bioRxiv. 2020. https://doi.org/10.1101/2020.04.22.044404. https://www.biorxiv.org/content/early/2020/04/26/2020.04.22.044404.full.pdf.
5. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, Zhou Z, Yang J, Zhong J, Yang D, Guo L, Zhang G, Li H, Xu Y, Chen M, Gao Z, Wang J, Ren L, Li M. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. Clin Infect Dis. 2020. https://doi.org/10.1093/cid/ciaa203. https://academic.oup.com/cid/advance-article-pdf/doi/10.1093/cid/ciaa203/33167020/ciaa203.pdf.
6. van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol. 2020104351. https://doi.org/10.1016/j.meegid.2020.104351.
7. Parikh AR, Leshchiner I, Elagina L, Goyal L, Levovitz C, Siravegna G, Livitz D, Rhrissorrakrai K, Martin EE, Van Seventer EE, Hanna M, Slowik K, Utro F, Pinto CJ, Wong A, Danysh BP, de la Cruz FF, Fetter IJ, Nadres B, Shahzade HA, Allen JN, Blaszkowsky LS, Clark JW, Giantonio B, Murphy JE, Nipp RD, Roeland E, Ryan DP, Weekes CD, Kwak EL, Faris JE, Wo JY, Aguet F, Dey-Guha I, Hazar-Rethinam M, Dias-Santagata D, Ting DT, Zhu AX, Hong TS, Golub TR, Iafrate AJ, Adalsteinsson VA, Bardelli A, Parida L, Juric D, Getz G, Corcoran RB. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. Nat Med. 2019;25(9):1415–21. https://doi.org/10.1038/s41591-019-0561-9.
8. Morgillo F, Corte CMD, Fasano M, Ciardiello F. Mechanisms of resistance to EGFR-targeted drugs: lung cancer. ESMO Open. 2016;1:000060.
9. Shaw AT, Friboulet L, Leshchiner I, Gainor JF, Bergqvist S, Brooun A, Burke JB, Deng YL, Liu W, Dardaei L, Frias RL, Schultz KR, Logan J, James LP, Smeal T, Timofeevski S, Katayama R, Iafrate AJ, Le L, McTigue M, Getz G, Johnson TW, Engelman JA. Resensitization to crizotinib by the lorlatinib alk resistance mutation l1198f. N Engl J Med. 2016;15:54–61.
10. Yates LR, Knappskog S, Wedge D, Farmery JHR, Gonzalez S, Martincorena I, Alexandrov LB, Loo PV, Haugland HK, Lilleng PK, Gundem G, Gerstung M, Pappaemmanuil E, Gazinska P, Bhosle SG, Jones D, Raine K, Mudie L, Latimer C, Sawyer E, Desmedt C, Sotiriou C, Stratton MR, Sieuwerts AM, Lynch AG, Martens JW, Richardson AL, Tutt A, Lonning PE, Campbell PJ. Genomic evolution of breast cancer metastasis and relapse. Cancer Cell. 2017;32:169–84.
11. Allred DC, Brown P, D DM. The origins of estrogen receptor alpha-positive and estrogen receptor alpha-negative human breast cancer. Breast Cancer Res. 2004;6:240–5.
12. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;1984:23–8.
13. Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biol. 2016;17:86.
14. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biol. 2016;17:69.
15. Deshwar AG, S V, Yung CK, Jang GH, Morris LSQ. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015;16:35.
16. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015;31:62–70.
17. Hajirasouliha I, Mahmoody A, Raphael BJ. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. Bioinformatics. 2014;30(12):78–86.

18. McPherson SMAW, Donmez N, Cenk S. Clonality inference in multiple tumor samples using phylogeny. Bioinformatics. 2015;31:1349–56.
19. Miller CA, White BS, Dees ND, Griffith M, Welch JS, Griffith OL, et al. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput Biol. 2014;10: 1003665.
20. Ross EM, Markowetz F. PyClone: statistical inference of clonal population structure in cancer. Nat Methods. 2014;11:396–8.
21. Aguse N, Qi Y, El-Kebir M. Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. Bioinformatics. 2019;35(14):408–16. https://doi.org/10.1093/bioinformatics/btz312. https://academic.oup.com/bioinformatics/article-pdf/35/14/i408/28913337/btz312.pdf.
22. Qi Y, Pradhan D, El-Kebir M. Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors. Algorithms Mol Biol. 2019;14(1). https://doi.org/10.1186%2Fs13015-019-0155-6.
23. Jiao W, Vembu S, Deshwar AG, et al. Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics. 2014;15(1):35. https://doi.org/10.1186%2F1471-2105-15-35.
24. Leshchiner I, Livitz D, Gainor JF, Rosebrock D, Spiro O, Martinez A, Mroz E, Lin JJ, Stewart C, Kim J, Elagina L, Bozic I, Mino-Kenudson M, Rooney M, Ou S-HI, Wu CJ, Rocco JW, Engelman JA, Shaw AT, Getz G. Comprehensive analysis of tumour initiation, spatial and temporal progression under multiple lines of treatment. bioRxiv. 508127. https://doi.org/10.1101/508127.
25. Myers MA, Satas G, Raphael BJ. CALDER: Inferring phylogenetic trees from longitudinal tumor samples. Cell Syst. 2019;8(6):514–522.e5. https://doi.org/10.1016%2Fj.cels.2019.05.010.
26. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4(1):. https://doi.org/10.1093/ve/vey016. https://academic.oup.com/ve/article-pdf/4/1/vey016/25028516/vey016.pdf.
27. Ismail WM, Tang H. Clonal reconstruction from time course genomic sequencing data. BMC Genomics. 2019;20(Suppl 12):1002. https://doi.org/10.1186/s12864-019-6328-3.
28. Landau DA, et al. The evolutionary landscape of chronic lymphocytic leukemia treated with ibrutinib targeted therapy. Nat Commun. 2017;8: 2185. https://doi.org/10.1038/s41467-017-02329-y.
29. Fehr AR, Perlman S. In: Maier HJ, Bickerton E, Britton P, editors. Coronaviruses: An Overview of Their Replication and Pathogenesis. New York: Springer; 2015, pp. 1–23.
30. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine DA, Lander ES, Meyerson M, Getz G. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21.
31. Fernàndez-Baca D. The perfect phylogeny problem. In: Cheng XZ, Du DZ, editors. Steiner Trees in Industry. Boston: Springer; 2001. p. 203–34.
32. Miller CA, McMichael J, Dang HX, et al. Visualizing tumor evolution with the fishplot package for r. BMC Genomics. 2016;17(1):. https://doi.org/10.1186%2Fs12864-016-3195-z.
33. Jain AK, Dubes RC. Algorithms for Clustering Data. Englewood Cliffs: Prentice-Hall; 1988.
34. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, G.-F. X, Shi Z-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270–3.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.