Supplementary Materials for Chowdhury et al., Inferring Models of Multiscale Copy Number Evolution for Single-Tumor Phylogenies

S1 Supplementary Methods

In this section, we provide proofs of correctness for the algorithms presented in the main paper as well as methodological details related to supplementary results omitted from the main paper. We also include a section describing Methods used in simulation experiments whose results appear in Supplementary Results, rather than in the main manuscript.

Our main theoretical result is a method for inferring minimum distances between two states within a copy number phylogeny when duplication/loss of single genes (SD), duplication/loss of all genes on a common chromosome (CD), and duplication of all genes in the full genome (GD) are possible and each event type is associated with a weight parameter. We first establish some mathematical results used by our algorithm for accurate distance computation. This algorithm then becomes a subroutine in a heuristic Steiner tree algorithm for inferring copy number phylogenies in the presence of weighted SD, CD, and GD events. Finally, we develop an iterative algorithm to infer the rate of different event types from the observed data using the weighted Steiner tree inference algorithm as a subroutine.

In what follows, we will consider sequences of SD, SD+CD, SD+GD and SD+CD+GD events, which we call *paths*. A *boundary-insensitive* path is one in which the copy numbers of intermediate configurations can take on any integer values and for which zero copy number is not treated specially. Zero is special, however, because once the copy number of a gene is reduced to zero, it is generally assumed that the gene cannot be gained back to get to copy number one. We also define *boundary-sensitive* paths for which intermediate copy-numbers must lie between positive bounds, denoted by LB and UB. When we present pseudocode for computing paths, we will discuss how zero copy number is handled.

We introduce some notation required for specifying and proving the theoretical results:

1. The observed data consists of copy-number counts of probes g_i for $i = 1, \ldots, d$. Typically, each probe allows one to count the copy number of a particular gene, so we refer to the g_i as *genes*.

2. We define a *configuration* $C(g_1, g_2, \ldots, g_d)$ to be a vector of length d of integers representing the copy numbers of each gene. A configuration is the state that might be observed for a single cell. When the collection $g_1, \ldots, g_d$ is clear from context, we will just write C as a shorthand.

3. $w_{g_i}^{\{g,l\}}, w_{c_i}^{\{g,l\}}, w_d$: Cost/weight of gain ($w_{g_i}^g$) or loss ($w_{g_i}^l$) associated with individual gene g_i, individual chromosome c_i ($w_{c_i}^{\{g,l\}}$) or cost/weight of whole genome duplication event (w_d). The weight for a particular event is derived from the probability p of observing that event by the rule $w = -\log p$.

4. We write $L_1(C^i, C^j)$ to denote the length of the shortest-length boundary-insensitive SD path between configurations C^i and C^j. As we discuss below, the length is precisely the rectilinear, or L_1, distance between the configurations, justifying the notation.

5. The weight of the minimum-cost SD path between C^i and C^j is denoted $R^w(C^i, C^j)$.
6. We let $D_w^{s,ch}(C^i, C^j)$ denote the weight of the minimum-cost boundary-insensitive SD+CD path between C^i and C^j.

A *feasible* configuration is one in which all counts are between LB and UB. A feasible path consists entirely of feasible configurations. An *infeasible* path has at least one infeasible configuration. Every infeasible path is boundary-insensitive, but a boundary-insensitive path may be either feasible or infeasible.

SD and SD+CD events have the desirable property that the order of SD or SD+CD events can be rearranged arbitrarily (Chowdhury *et al.*, 2014); such a property does not hold for paths with GD events. In our previous work, we established the following two lemmas for the unweighted SD and CD cases (Chowdhury *et al.*, 2014):

Lemma S1. *A shortest-length boundary-insensitive sequence of CD and SD events cannot have both a gain of chromosome c_i and a loss of the same chromosome c_i.*

Lemma S2. *A shortest-length boundary-insensitive sequence of events cannot have both a gain of gene g_i and a loss of the same gene g_i.*

We now prove that the natural generalizations of the two lemmas hold in the weighted case too:

Lemma S3. *A minimum-weight boundary-insensitive sequence of CD and SD events cannot have both a gain of chromosome c_i and a loss of the same chromosome c_i.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain and a loss of the same chromosome. Then removing one gain and one loss produces a new sequence that weighs $(w^g_{c_i} + w^l_{c_i})$ less and has the same final state. □

Lemma S4. *A minimum-weight boundary-insensitive sequence of SD and CD events cannot have both a gain of gene g_i and a loss of the same gene g_i.*

Proof. By contradiction. Suppose S is a sequence of events that has both a gain of g_i and a loss of g_i. Then removing one gain and one loss produces a new sequence that weighs $(w^g_{g_i} + w^l_{g_i})$ less and has the same final state. □

From these lemmas, it follows that the length of the shortest-length SD path between configurations C^i and C^j is precisely the rectilinear distance, justifying our use of the notation $L_1(C^i, C^j)$. In contrast, $R^w(C^i, C^j)$ is not a distance because it is not symmetric, but it can be expressed as a sum

$$\begin{aligned} R^w(C^i, C^j) = & \sum_{C^i(g_k) < C^j(g_k)} (C^j(g_k) - C^i(g_k)) w^g_{g_k} \\ & + \sum_{C^i(g_k) > C^j(g_k)} (C^i(g_k) - C^j(g_k)) w^l_{g_k} \end{aligned}$$

For other types of path, the length of the shortest-length paths and the cost of the minimum-weight path are not so easily expressed. Developing algorithms to compute weights of minimum-weight paths is the topic of the rest of this section.

Progression model considering SD and CD events

One of our main theoretical contributions consists of novel theory for inference of minimum-weight paths of single-gene (SD) and single chromosome (CD) events from a starting configuration $C^s(g_1, g_2, \ldots, g_d)$ to a terminal configuration $C^t(g_1, g_2, \ldots, g_d)$. Our model assumes that on division of a tumor cell, the configuration can change either by gain or loss of one copy of a single gene (SD event) or by gain or loss of one copy of each gene on a single chromosome (CD event). For example, a configuration of four genes $(2, 2, 2, 2)$ with the first two genes on the same chromosome might evolve to $(3, 2, 2, 2)$ by a single SD event or to $(3, 3, 2, 2)$ by a single CD event.

First, we establish the correctness and optimality of Algorithm 5, which computes an SD+CD path starting with zigzag paths and finishing with SD steps. Second, we establish the correctness and optimality of Algorithm 6 that determines how many zigzag paths, if any, should be used at the beginning of an optimal path transform $C^s(g_i, g_{i+1}, \ldots, g_j)$ into $C^t(g_i, g_{i+1}, \ldots, g_j)$. All steps after the initial zigzag paths are SD steps.

We focus on a single chromosome because, as explained below, the problem of finding the minimum-weight SD+CD path can be solved one chromosome at a time. In the case in which there is only data for single gene probe on a chromosome, one cannot distinguish CD events from SD events. We treat such cases mathematically by setting the weight of the corresponding CD events to infinity. In practice, one may simply calculate an SD path for those genes.

Thus far, we have discussed boundary-insensitive paths, which might contain problematic intermediate cases with zero copy-number or absurd cases with negative copy number. However, if one finds an optimal boundary-insensitive path, one may construct an optimal boundary-sensitive path of the same weight.

Theorem S5. *If there is a minimal-weight boundary-insensitive sequence of SD and CD events between two feasible configurations C^s and C^t, where the smallest feasible copy number is at least one, then there is a boundary-sensitive sequence of SD and CD events with the same weight.*

Proof. Assume without loss of generality that all genes are on the same chromosome. If the boundary-insensitive path contains no CD event then it is an SD path and all gene counts change monotonically, so the theorem holds.

We will suppose that all the CD events are chromosome losses; the argument for CD gains is symmetric. If placing a chromosome loss next on the path would result in an infeasible configuration, then the copy number of some genes must be at LB. If these genes did not have any SD gains in the path, then the ultimate configuration would be infeasible, which by assumption is not. Therefore, we place a single SD gain for each of the genes with copy number at the LB next on the path in arbitrary order. Adding these SD gains cannot produce an intermediate configuration that is infeasible because the genes are at copy number LB, and these gains will just increase the count for each gene by one.

Next, we place the chromosome loss on the path. The resulting configuration is feasible. We repeat the process until there are no more chromosome losses. By construction, once the final chromosome loss has been placed, we attain a feasible intermediate configuration via a feasible path, and all further events are SD gains. Thus, thereafter, we only have monotonic gene gain/losses between two feasible configurations. Thus, the entire path is feasible. □

In the Results section of the main paper, we analyze data with minimum copy number (LB) zero and maximum copy number (UB) nine, but in this subsection we assume positive copy number. Zero copy number is handled as a special case in a later subsection.

Now, we prove the following results, which exclude the possibility of gains and losses of the same chromosome or the same gene on the boundary-sensitive paths.

Corollary S6. *In a feasible optimal path, where the smallest feasible copy number is at least one, there cannot be both a gain and loss of the same chromosome or a gain and loss of the same gene.*

Proof. The proof is by contradiction. Take an optimal feasible path, which is by definition comprised of a boundary-sensitive sequence of operations. It is also a valid boundary-insensitive path. If it contains paired gain/losses, we can rearrange and cancel to produce a lower weight path. By the previous Theorem, there is also a boundary-sensitive path with this lesser weight. Hence, the original boundary-sensitive path could not have been optimal. □

Algorithm for computing the SD+CD distance

The algorithm for computing the SD+CD distance is centered around the concept of a *zigzag subpath* from C^s to C^t, which is so named because its construction focuses on alternations between consecutive gain and loss events. For any *zigzag* path, we first choose a predominant *sense* or *direction* as either *gain* or *loss*, where the sense of the zigzag path is the sense of the CD events in that path. When the *zigzag* sense is loss, then we determine the set of genes that are on the affected chromosome and for which the copy number at C^s is less than or equal to the copy number at C^t. For each such gene, we insert an SD gain in the path in arbitrary order. Then we insert a single CD loss. We define the path symmetrically when the sense of the zigzag path is gain. We use the shorthand "zigzag gain" (respectively, "zigzag loss") to refer to a zigzag path of sense gain (loss).

Thus, a zigzag path is a series of zero or more SD changes followed by a single CD change with opposite sense (gain/loss). The sense of a zigzag step is the sense of the final CD event.

Lemma S7. *There is a CD step on an optimal SD+CD path from C^s to C^t if and only if there is a (possibly different) optimal path from C^s on the way to C^t that starts with a zigzag subpath of the same sense, affecting the same chromosome.*

Proof. We consider a CD loss and argue symmetrically for a CD gain. If there is a CD loss on the optimal path, then by Corollary S6, the only CD events on the optimal path are CD losses. We choose one such CD loss and let that choice determine the affected chromosome.

If the copy number of a gene at C^s is less than or equal to the copy number of that gene at C^t, then the path must contain at least one SD gain for that gene. Thus, one may rearrange the optimal path to create an equal-weight path that starts with a zigzag loss.

The converse is true because a zigzag subpath contains a CD step, by definition. □

Lemma S8. *Consider a specific zigzag path. Let ℓ be a vector such that $\ell_k = 1$, if the copy number of gene g_k changes after the zigzag path and zero otherwise. Let m be another vector for which $m_k = 1$ if gene g_k is on chromosome affected by the CD step, but the copy number of gene g_k does not change after the entire zigzag path. Then,*

$$\ell_k + m_k = \begin{cases} 1 & \textit{if gene } g_k \textit{ is on the chromosome lost or gained} \\ 0 & \textit{otherwise.} \end{cases}$$

Moreover, after a zigzag loss, the copy number of every gene is the same or lower, and after a zigzag gain, the copy number of every gene is the same or higher.

Proof. This Lemma is a consequence of the definition of zigzag paths and their senses. If a gene is on the chromosome affected by a the CD step, but is not matched by a corresponding SD step of the opposite sense, the copy number of the gene must change. The sense of any change in gene copy number is the same as the sense of the CD step. □

Lemma S9. *If C^{int} is an intermediate configuration created by taking a zigzag step of either sense from C^s on the way to C^t,*

$$L_1(C^s, C^t) = L_1(C^{int}, C^t) + \sum_k \ell_k,$$

where ℓ is defined as in Lemma S8.

Proof. We consider only zigzag losses; the argument for zigzag gains is symmetric. By definition of a zigzag loss, exactly those genes with copy number greater at C^s than C^t have a copy number change after the zigzag loss, and the copy number of each of these genes decreases by one. But those genes are precisely the genes with indices k such that $\ell_k = 1$. □

Lemma S10. *If C^{int} is an intermediate configuration reached by taking a zigzag path from C^s on the way to C^t, then*

$$R^w(C^s, C^t) = R^w(C^{int}, C^t) + a^T\ell,$$

where ℓ is defined as in Lemma S8, and a is a vector for which a_k represents the cost of an SD step for gene g_k of the same sense as the zigzag path.

Proof. By definition of a zigzag loss, exactly those genes with copy number greater at C^s than C^t have a copy number change after the zigzag loss, and the copy number of each of these genes decreases by one. By Lemma S9, those genes are precisely the genes for which $\ell_k = 1$. For a zigzag loss, then a_k represents the weight of the loss of gene g_k. Thus, the weight of an optimal SD subpath from C^{int} to C^t differs from the weight of an optimal SD subpath from C^s to C^t by exactly $a^T\ell$.

We can argue symmetrically for zigzag gains. □

Next, we develop the rule indicating when a CD and a zigzag step is possible.

Theorem S11. *When there is no SD+CD path between C^s and C^t of strictly lower weight than an optimal SD path between C^s and C^t, a CD step from C^s will result in an intermediate configuration C^{int} for which*

$$R^w(C^{int}, C^t) + w_c \geq R^w(C^s, C^t),$$

where w_c is the cost of a CD step with the same sense as the zigzag path.

Proof. The weight of a CD event is precisely w_c. One may then take an SD path from C^{int} to C^t for total path weight $R^w(C^{int}, C^t) + w_c$. If this is strictly less than the weight of an SD path from C^s to C^t, then this constitutes an SD+CD path that has lower weight than the optimal SD path. □

A similar result holds for zigzag paths. By Lemma S7, if there is a CD step on the *optimal* path, the path may be rearranged so that there is a zigzag path.

Lemma S12. *When there is no SD+CD path between C^s and C^t of strictly lower weight than an optimal SD path between C^s and C^t, a zigzag step of the same sense from C^s will result in an intermediate configuration C^{int} for which*

$$R^w(C^{int}, C^t) + w_c + b^T m \geq R^w(C^s, C^t), \tag{1}$$

where b is a vector such that b_k is the cost of an SD step affecting gene g_k in the sense opposite to the sense of the zigzag path. For instance, if the path is a zigzag loss, b_k is the cost of an SD gain of gene g_k.

Proof. The sum of the weights of the SD steps and the CD step that make up the zigzag step is precisely $w_c + b^T m$. One may then take an SD path from C^{int} to C^t for total path weight $R^w(C^{int}, C^t) + w_c + b^T m$. If this is strictly less than the weight of an SD path from C^s to C^t, then this constitutes an SD+CD path that has lower weight than the optimal SD path. □

Theorem S13. *When there is an SD+CD path between C^s and C^t that has lower weight than an optimal SD path, then for any CD step on the SD+CD path, taking a zigzag step from C^s on the way to C^t of the same sense and affecting the same chromosome results in an intermediate configuration C^{int} for which*

$$R^w(C^{int}, C^t) + w_c + b^T m < R^w(C^s, C^t), \tag{2}$$

where m is defined as in Lemma S8 and b is defined as in Lemma S12.

Proof. We proceed by induction on $L_1(C^s, C^t)$. If $L_1(C^s, C^t) = 0$, then C^s and C^t are the same configuration and $R^w(C^s, C^t) = 0$. Any nonempty SD+CD path has nonnegative weight, and so cannot have lower weight than the optimal SD path, and the claim holds.

The induction hypothesis is that the claim holds whenever $L_1(C^s, C^t) < n$. Suppose therefore that $L_1(C^s, C^t) = n$. Let C^{int} be an intermediate point such that $L_1(C^{int}, C^t) < n$, so that the induction hypothesis applies to paths between C^{int} and C^t.

To simplify the exposition, assume without loss of generality that all genes are on the same chromosome. Furthermore, let us consider the case in which an SD+CD path between

C^s and C^t containing a CD loss has lower weight than an optimal SD path; the argument for paths that contain a CD gain is symmetric.

Let C^{int} be an intermediate configuration generated by a zigzag loss from C^s on the way to C^t. Since we assume inequality (2), taking the zigzag path must result in a change in the copy number of at least one gene. Thus, by Lemma S9, it must be that $L_1(C^{int}, C^t) < n$. Thus, one may apply the induction hypothesis on paths from C^{int} to C^t.

Suppose there is no SD+CD path from C^{int} to C^t that has lower weight than the optimal SD path. In such a case, the SD path is itself optimal as an SD+CD path from C^{int} to C^t. Thus, any path constrained to start with a zigzag path from C^s to C^{int}, and continuing on to C^t, has weight at least $R^w(C^{int}, C^t) + w_c + b^T m$. But by the assumptions of the theorem, there is an SD+CD path containing a CD loss and having weight less than the optimal SD path, and by Lemma S7 such a path may be constrained to start with a zigzag loss. Thus, inequality (2) holds.

Suppose, therefore, that there is an SD+CD path from C^{int} to C^t that has weight less than that of the optimal SD path and that contains a CD loss. Then by Lemma S10,

$$R^w(C^s, C^t) = R^w(C^{int}, C^t) + a^T \ell. \tag{3}$$

Let C^u be another intermediate configuration, generated by taking another zigzag path from C^{int} of the same sense. It holds that

$$R^w(C^{int}, C^t) = R^w(C^u, C^t) + a^T \widehat{\ell}, \tag{4}$$

where $\widehat{\ell}$ is a vector for which $\ell_k = 1$ for each gene whose copy number decreases between C^t and C^u. By the induction hypothesis

$$R^w(C^u, C^t) + w_c + b^T \widehat{m} < R^w(C^{int}, C^t), \tag{5}$$

where $\widehat{m}$ is a vector representing the SD gains in the zigzag step.

It must be that

$$a^T \widehat{\ell} \leq a^T \ell \text{ and } b^T m \leq b^T \widehat{m}, \tag{6}$$

because the nonzero entries of ℓ denote the set of genes with copy number strictly greater at C^s than C^t, and the nonzero entries of $\widehat{\ell}$ denote the set of genes with copy number strictly greater at C^{int} than C^t. But copy numbers can only decrease after a zigzag loss. Furthermore, by Lemma S8, $\ell_k + m_k = \widehat{\ell}_k + \widehat{m}_k = 1$, for all genes g_k on the relevant chromosome.

Combining (3)–(6), we find

$$\begin{aligned}
R^w(C^{int}, C^t) + w_c + b^T m &= R^w(C^u, C^t) + w_c + b^T m + a^T \widehat{\ell} && \text{by (4)}\\
&\leq R^w(C^u, C^t) + w_c + b^T \widehat{m} + a^T \ell && \text{by (6)}\\
&< R^w(C^{int}, C^t) + a^T \ell && \text{by (5)}\\
&= R^w(C^s, C^t), && \text{by (3)}
\end{aligned}$$

completing the induction step and the proof. □

Now, we develop the rule to identify which direction of zigzag step, if any, needs to be applied on the optimal boundary-sensitive path to convert C^s to C^t. We use the following shorthand notation for four possible partial paths:

1. ZZL: the zigzag loss path.
2. ZZG: the zigzag gain path.
3. SDL: SD losses of those genes that have a lower copy number in C^t than in C^s.
4. SDG: SD gains of those genes that have a higher copy number in C^t than in C^s.

Lemma S14. *Taking either ZZL or SDL leads to the same intermediate state. Taking either ZZG or SDG leads to the same intermediate state.*

Proof. The definition of ZZL is that it has one CD loss followed by SD gains of the genes for which the copy number at C^{int} is less than or equal C^t. Those genes have compensating losses and gains. In ZZL, the remaining genes, which have a higher copy number in C^s than in C^t have SD losses. That is the definition of SDL. The proof for ZZG and SDG is symmetric. □

The weight of ZZL is the sum of the cost of a CD loss and the sum of costs of SD gains of genes that have lower or equal copy number in C^s than in C^t. The weight of the SDL subpath is the sum of costs of single SD loss events for those genes which have lower copy number in C^t than in C^s. The weights of the ZZG and SDG partial paths are defined analogously. In the following theorem, we propose two alternative tests to identify the sense of the zigzag partial path to use, if any, as the value of the input parameter σ in Algorithm 6.

Theorem S15. *At most one of the following tests can be successful:*

1. *If the cost of ZZL is lower that that of SDL, take ZZL by setting $\sigma = -1$.*
2. *If the cost of ZZG is lower that that of SDG, take ZZG by setting $\sigma = 1$.*

Proof. The proof is by contradiction. Sort the costs of the four partial paths in increasing order, breaking ties arbitrarily. If both tests 1 and 2 succeed, then the most costly of the four paths must be SDL or SDG. Without loss of generality, assume the most expensive path is SDG, so that both $cost(ZZG) < cost(SDG)$ and $cost(ZZL) < cost(SDG)$. The path SDG has a proper subset of the single-step events in the path ZZL. Therefore, it is not possible that $cost(ZZL) < cost(SDG)$, a contradiction. □

We use the proof of correctness of Algorithm 6 to derive the main theorem of this subsection, which establishes a method to find a minimum-cost sequence of weighted SD and CD events for transforming C^s to C^t. Again, we can consider each chromosome separately since each CD and GD event affects only one chromosome.

Theorem S16. *Partition the gene list by chromosomes such that each chromosome $c_i \in \{c_1, \ldots, c_q\}$ corresponds to a consecutive subset of genes $g_{i,1}, \ldots, g_{i,d_i}$. Let $C^s(g_1, \ldots, g_d) = (s_1, \ldots, s_d)$ and $C^t(g_1, \ldots, g_d) = (t_1, \ldots, t_d)$. Then we can construct a minimum-cost boundary-sensitive sequence of events transforming C^s to C^t by constructing a minimum-cost boundary-sensitive sequence of events $\mathcal{S}_i$ transforming $(s_1, \ldots, s_{i,1}, \ldots, s_{i,d_i}, \ldots, s_d)$ to $(s_1, \ldots, t_{i,1}, \ldots, t_{i,d_i}, \ldots, s_d)$ for each chromosome c_i and interleaving the $\mathcal{S}_i$ in arbitrary order.*

Proof. The weight function can be decomposed into individual parts for genes belonging to distinct chromosomes as follows:

$$D_w^{s,ch}(C^s, C^t) = \sum_{i=1}^{q} D_w^{s,ch}(C^s(s_{i,1}, \ldots, s_{i,d_i}), C^t(s_{i,1}, \ldots, s_{i,d_i})) \tag{7}$$

Because the weight can be decomposed in this way and each CD or SD event contributes to only a single term of the outer sum, we can minimize the cost for each chromosome independently and combine the events from distinct chromosomes in arbitrary order without changing the value of the objective function. Likewise, since each chromosome affects a disjoint subset of genes, boundary-sensitive sequences for each chromosome will yield a boundary-sensitive sequence across all genes. □

Progression model considering SD, CD and GD events

In this subsection, we provide proof of correctness for Algorithm 4, a method for finding minimum-cost SD+GD paths that is called as a subroutine in the full SD+CD+GD method. The algorithm makes use of an observation that shortest-length SD+GD paths have a well-defined structure, established in Theorem S17.

Theorem S17. *For a fixed number $g \geq 1$ of GD steps, the minimum-weight path between $i \geq 1$ and $j \geq 1$ must end with either a GD event, if j is even, or a GD event followed by a single SD event, if j is odd. The result holds whether or not the SD events must have the same sense, or may be mixed.*

Proof. Consider the final GD event in the minimum-weight path and the SD events that follow it, necessarily of all the same sense (otherwise one would cancel a gain and a loss to get a shorter path). Suppose there are two or more SD events following the GD event. If the SD events after the GD event are gains, or the SD events after the GD events are losses but the start point of the GD event is greater than 1, then one could create a minimum-weight path by inserting an SD event of the same sense before the GD event, using that new point as the start point of a GD event and eliminating two of the SD events that follow it. Because we replaced two SD events with an SD event of the same sense, it is irrelevant whether SD events are restricted to a single sense.

The remaining case is when the GD event is from 1 to 2, but is followed by more than one SD loss. Since counts cannot go below zero, this scenario entails that $j = 0$, contrary to the assumptions of this theorem. Therefore, a minimum-weight path must end with a GD event followed by at most one SD event.

Since the endpoint of a GD event must be an even number, if j is an even number, it would not be possible to arrive at j by following a GD event with a single SD event. Thus, if j is even, it must be the endpoint of the GD event. Similarly, if j is odd, a single SD event must follow the final GD event. □

This theorem provides the basis for Algorithm 4, which constructs a table of shortest paths for all pairs of taxa i and j by enumerating over possible numbers of duplication events. An immediate corollary of the theorem is that for a maximum copy number of nine, which we use in our experiments, it suffices to consider at most four genome duplication events.

Corollary S18. *If $1 \leq j \leq 2^m$, there can be at most m GD steps in a minimum-weight SD+GD path between $i \geq 0$ and j.*

Proof. By induction on j. If $j = 1$ then there are no GD steps in a minimum-weight SD+GD path from i to j and the result holds. Otherwise, if j is even, then a minimum-weight path from i to j must end with a GD event starting at $k = j/2$. But then $k < 2^{m-1}$ and the induction hypothesis may be applied. Similarly, if j is odd, then the path must end with a GD step to $j - 1$ or $j + 1$, followed by an SD step. In either case, the GD step must start from $k \leq (j + 1)/2$. Thus $k < 2^{m-1}$, and again the induction hypothesis applies. □

Because we restrict copy numbers to be at most nine, four genome duplication events is sufficient. For a fixed $k > m$, there is still a minimum length SD+GD path with k genome duplications, but such a path must have a cycle where a copy number of 1 is duplicated by a GD event to become 2 which is followed by an SD loss to return to one. Though for a singe gene probe, a path containing a cycle would not be optimal, when a shortest-length SD+GD path is computed for several genes, some of the genes may exhibit such a cycle.

For $k = 0, \ldots, m$, Algorithm 4 creates a table of duplication points for the shortest-length SD+GD path between copy numbers of an individual probe for given a fixed number of genome duplication events. For our code $m = 4$ because $UB = 9$.

Simulation Methods

We performed simulation experiments to test how accurately our algorithms can infer the tumor progression models for data with known ground truth parameters and phylogenies. We initially generated ground truth trees for four different combinations of SD, CD and GD rates. For each combination of parameters, we generated 100 trees. Each node in the tree represented a configuration of six probes, two of which were considered to be located on the same chromosome. Each tree was generated from a diploid root node by a branching process. The number of children belonging to each node was generated by using a geometric distribution with mean 0.6. Each child was generated from the parent by executing an SD, CD or GD event, according to a particular probability distribution. The process was terminated when all leaf nodes were assigned 0 children. Afterwards, FISH data were generated by uniformly sampling 300 cells from the nodes in the tree. The simulated data corresponds to counts of probes for each sampled cell in the tree. We refer to the six genes in the simulation as *G1, G2, G3, G4, G5* and *G6.* The genes *G5* and *G6* are on the same chromosome and the other four genes are on separate chromosomes. We then used our algorithms to infer the tumor phylogenetic trees and collected the inferred parameter values for each event type.

We initially used the following four set of parameter values, intended to stress ability of the code to discriminate between different rates of gain and loss of single genes, different rates between distinct genes, and different relative rates of gene, chromosome, or whole-genome gain or loss:

- Combination 1: The SD gain and loss probabilities of each of the genes are 0.075, gain and loss probabilities of the chromosome are 0.035 and the probability of whole genome duplication is 0.03.

- Combination 2: For each gene, the SD gain and loss probabilities are 0.1 and 0.05 respectively. Gain and loss probabilities of the chromosome are 0.035 and the probability of whole genome duplication is 0.03.

- Combination 3: The SD gain and loss probability values for *G1, G2, G3, G4, G5* and *G6* are (0.20,0.03),(0.13,0.02), (0.1,0.05), (0.1,0.05), (0.07,0.02) and (0.08,0.02) respectively. Gain and loss probabilities of the chromosome are 0.05 and 0.02 respectively. The probability of whole genome duplication is 0.06.

- Combination 4: The SD gain and loss probability values for *G1, G2, G3, G4, G5* and *G6* are (0.03,0.15), (0.10,0.04), (0.05,0.11), (0.05,0.11), (0.10,0.03) and (0.06,0.03) respectively. Gain and loss probabilities of the chromosome are 0.07 and 0.02 respectively. The probability of whole genome duplication is 0.05.

Next, we performed simulation experiments to assess the effect of sample size on the ability of the algorithm to infer the rate parameters accurately. For this experiment, we used the same rate parameters as in combination 4, but instead of sampling 300 cells from each phylogeny, we sampled 100 and 200 cells separately and then used FISHtrees to infer the tumor phylogenetic model.

We next performed simulation experiments to observe the effect of having two, rather than one, chromosome that had multiple gene probes. We simulated 100 trees from a fifth scenario with eight probes *G1-8*. Genes *G5, G6* and *G7, G8* were considered to be located on the first (refer to here as *CHR1*) and second (*CHR2*) chromosome respectively. Genes *G1-4* were modeled as residing on chromosome distinct from each other and distinct from chromosomes *CHR1* and *CHR2*. For generating the trees, we used SD gain and loss probabilities of (0.15,0.03), (0.07,0.03), (0.05,0.02), (0.05,0.03), (0.07,0.03), (0.07,0.04), (0.03,0.05) and (0.06,0.02). Gain and loss probabilities for *CHR1* and *CHR2* are (0.06,0.03) and (0.03,0.05) respectively. The probability of whole genome duplication is 0.03. We refer to this combination of parameter values as Combination 5.

We also quantified how accurately our algorithm can reconstruct the evolutionary trees by comparing the inferred trees with the simulated trees. We first pruned the real trees to remove any subtrees for which no cell was sampled. This step was performed to avoid penalizing for the impossible problem of inferring subtrees unsupported by any data. We then computed the set of nontrivial bipartitions for the nodes in each tree by removing each of the internal edges of the tree one at a time. We then built a complete bipartite graph using the sets of bipartitions in the real and inferred trees as nodes of that graph and computing the edge weight as the cardinality of the set of nodes shared between the corresponding bipartitions. We computed maximum matching of edges between simulated and inferred trees and used the following formula for computing the reconstruction error E of the inferred trees (Chowdhury *et al.*, 2014):

$$E = (1 - \frac{W}{|T|(|P_r|+|P_i|-W)}) \times 100$$

Here W denotes weight of the maximum matching, $|T|$ is the set of taxa common between the real and inferred trees, and $|P_r|$ and $|P_i|$ denote the number of nontrivial bipartitions in the real and inferred trees respectively. Traditional distance measures, such as Robinson-Foulds (RF), cannot be used in this case, since RF is not defined for trees with different node sets. As in our prior work, we used a matching-based tree distance similar to that of Lin et al. (Lin *et al.*, 2012) but modified to deal with the complication that observed nodes in our data often include internal nodes of the tree, unlike in conventional species-tree phylogenetics.

We compared the performance of our algorithm with two standard phylogeny building algorithms, Neighbor Joining (NJ) and Maximum Parsimony (MP), that are scalable to similar numbers of taxa and have been previously used in inferring single-cell tumor phylogenies. We used implementations of both algorithms in MEGA version 6 (Tamura *et al.*, 2013). We tested the accuracy of NJ and MP on inference of 50 real trees simulated using the parameter combination 3. We used Euclidean distance between the taxa for computing the pairwise distance matrix in NJ, and for MP, we treated each copy number as arbitrary characters. We used the weighted matching based method to compute the mean percentage reconstruction error E between the real and inferred trees. We compared the performance of NJ and MP with FISHtrees on the same set of trees.

S2 Supplementary Results

In this section, we provide some additional results that are not essential to the major results or conclusions of the paper but provide additional validation or support for the methods introduced. We describe and analyze two additional real-life datasets that each has the characteristic that there is at least one pair of gene probes that share a chromosome. Therefore, these datasets allow us to test the utility of modeling chromosome gain/loss (CD) events. In each case, the pair of co-located genes are one oncogene and one tumor suppressor, so our weighted models need to balance between CD events in which both genes have the copy number changing in the same direction and SD events, which would usually be gains for the oncogenes and usually be losses for the tumor suppressors. From a clinical point of view, each of the four datasets (two in the main document, two in this Supplementary Materials document) was derived from a study with a distinct study design. Yet, for all four data sets, we can derive meaningful qualitative inferences, suitable to the particular study design, from our tumor progression models.

Breast Cancer (BC) Dataset:

We analyzed a breast cancer dataset (BC, (Heselmeyer-Haddad *et al.*, 2012)) that consists of paired DCIS and IDC samples collected from 13 breast cancer patients. Each sample was probed on eight genes, out of which *COX2, MYC, HER2, CCND1* and *ZNF217* are oncogenes and *DBC2, CDH1* and *P53* are tumor suppressors. Out of the eight genes, *DBC2* and *MYC* reside on chromosome 8 and *HER2* and *P53* reside on chromosome 17. Other genes reside on distinct chromosomes.

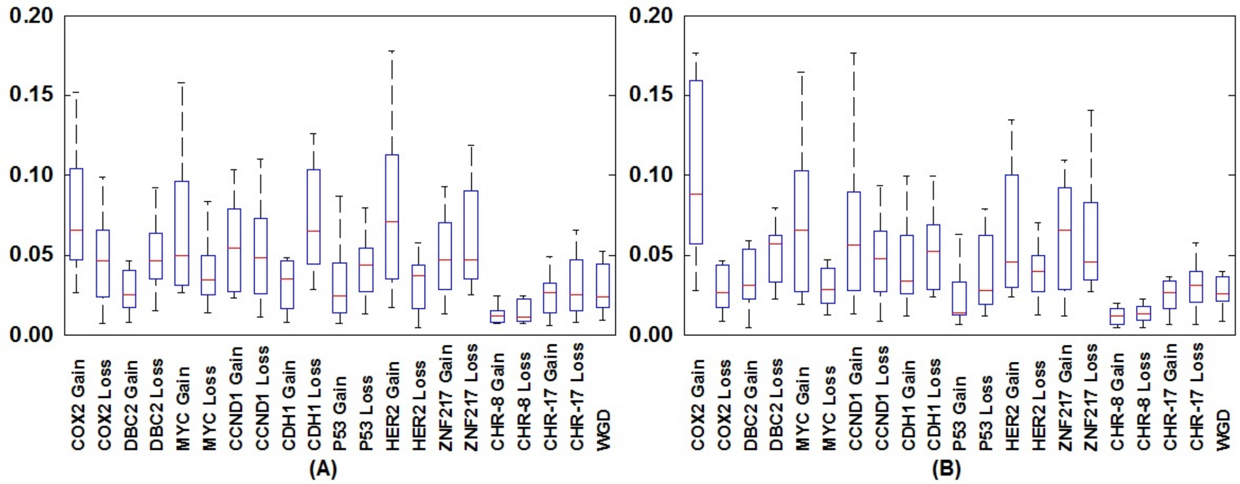

Figure S1: Inferred parameter values across ductal carcinoma in situ (DCIS) (A) and invasive ductal carcinoma (IDC) (B) samples.

We applied our tree building and rate estimation algorithm on the BC dataset. Boxplots for the estimated rates of each event are shown in Figure S1 for DCIS and IDC samples separately. As a first statistic, we ranked the different events across all of the 13 DCIS and 13 IDC samples separately based on their median parameter values. The most frequent events (with median parameter values ≥ 0.05) for both of the DCIS and IDC cases were "Gain of *COX2*", "Gain of *MYC*" and "Gain of *CCND1*". "Loss of *DBC2*" and "Gain of *ZNF217*" appeared as the most frequent events in the IDC samples only.

Similarly to the CC1 dataset, whose analysis is shown in the main document, we next performed an edge count-based statistical test of separation of the DCIS and IDC samples based on the gain/loss values of individual genes. "Gain of *MYC*" (total 5 out of 13 pairs of samples), "Gain of *COX2*"(4) and "Gain of *CCND1*"(2) showed statistically significant separation of DCIS and IDC samples more times than any other event. This is interesting because "Gain of *MYC*" did not appear to have a significant effect in the progression dynamics of DCIS to IDC in our previous analysis using an unweighted single gene duplication model (Chowdhury *et al.*, 2013), but was shown by other methods to have important effects during progression from DCIS to IDC (Heselmeyer-Haddad *et al.*, 2012). Our current analysis, based on the more realistic model of tumor progression developed in the present work, supports the conclusion from (Heselmeyer-Haddad *et al.*, 2012).

Second Cervical Cancer (CC2) Dataset:

We also examined an additional cervical cancer dataset, dubbed CC2. Dataset CC2 consists of following set of cervical samples: (a) one early pre-cancerous lesion (denoted by CIN1), (b) ten late pre-cancerous lesions (denoted by CIN3) and (c) ten cancerous lesions (denoted by CA). Each sample was probed on eight genes: *COX, ING5, FHIT, TERC, TERT, MYC, CHEK1, ZNF217. ING5, FHIT, CHEK1* are tumor suppressors and all others are oncogenes. *FHIT* and *TERC* lie on chromosome 3, while the others all reside on distinct chromosomes.

We applied our algorithm on each of the 21 samples in the CC2 dataset. Boxplots for inferred parameter values across the pre-cancerous and cancerous samples separately are shown

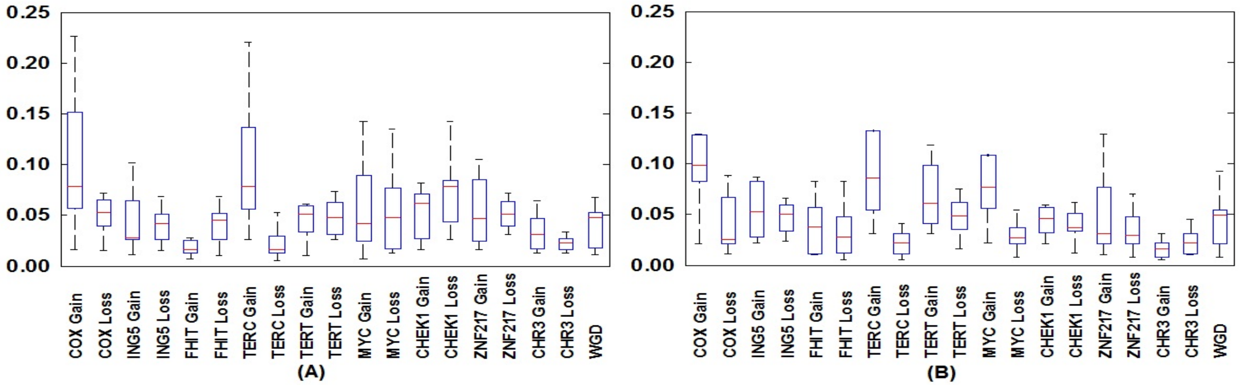

Figure S2: Inferred parameter values across pre-cancerous (A) and cancerous (B) cervical tumor samples.

in Figure S2. Similarly to the BC dataset, we again ranked the events based on the median parameter values across all of the pre-cancerous and cancerous samples separately. The most frequent events (with median parameter values ≥ 0.05) across both of the pre-cancerous and cancerous samples were "Gain of *COX*" and "Gain of *TERC*". "Loss of *CHEK1*" appeared more frequently in the cancerous samples only.

Distribution of cells across primary and metastatic CC1 trees:

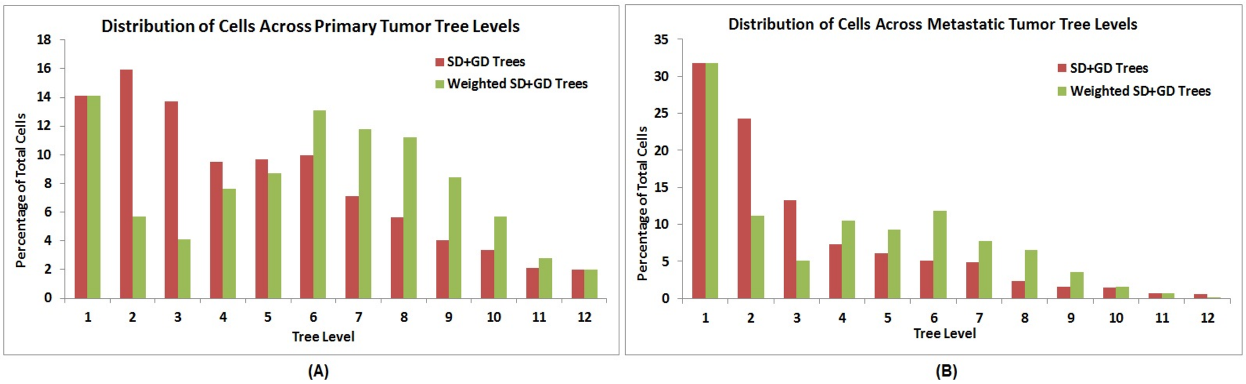

Figure S3: Distribution of cells across different tree levels of primary (A) and metastatic (B) tumor phylogenies.

We previously showed that the differential selective pressures working on different stages of cancer is reflected in the distribution of cells located across different levels of the trees built on samples collected from these sites of tumors. We tested how this cell distribution across tree levels is changed in trees built using our new method. In Figure S3, we show the distribution of cells across different phylogenetic tree levels of the primary and metastatic samples separately. Comparison between the weighted and unweighted versions of the SD+GD tree cell distributions reveals different dynamics of tumor progression under these two models. In both primary and metastatic cases, the unweighted event trees have more cells located in the first few levels in comparison to the weighted SD+GD trees, where we observe a bimodal distributions of cells with one mode located at deeper levels in the trees. In the primary case, 72.88% of the total cells are located in the first six levels of the unweighted SD+GD trees, while for the weighted SD+GD trees, this value is 53.29%. Similar bias in cell distribution is observed for metastatic trees, too, where 87.96% and 79.65% of the total cells are located in the first six levels of the unweighted and weighted SD+GD trees, respectively. Our newly proposed model leads to results more consistent with the Nowell model of tumor progression, which predicts that initially tumor cells undergo a brief period of heterogeneity followed by expansion of one or more clones (Nowell, 1976).

Simulation Results

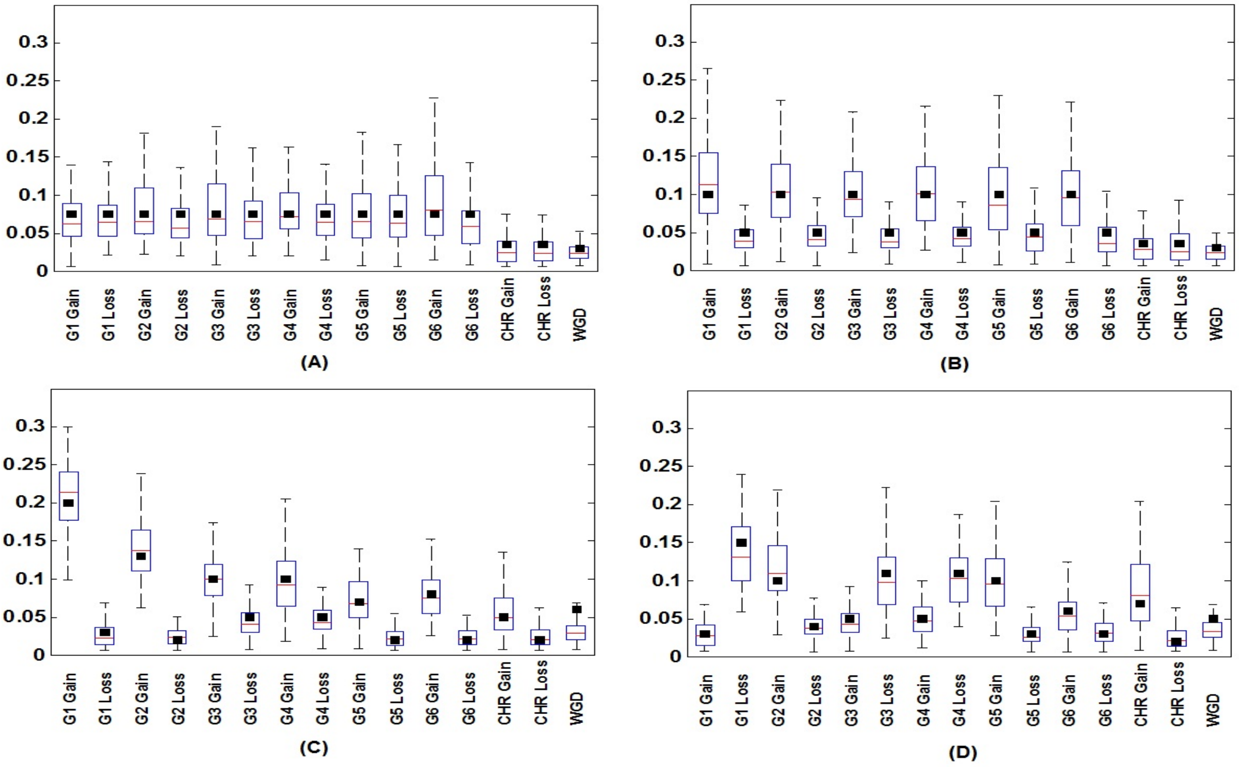

Figure S4: Boxplots showing inferred parameters from the simulated data for parameter value combinations 1 (A), 2 (B), 3 (C) and 4 (D). The true parameter values used in the simulation are shown as black squares in each plot.

We first examine performance on our four initial combinations of parameter values, for which a single pair of genes is found on a common chromosome. Figure S4 shows boxplots of distributions of inferred parameters across 100 simulated trees for each scenario as compared to the true parameter values used in generating the simulated data. From these plots, we can see that FISHtrees parameter inferences are highly responsive to changes in true parameter values. While there can be appreciable variance in estimates, median values closely track real values across data sets in almost every case. In combinations 3 and 4, FISHtrees underestimates the probability of whole genome duplication event by a small margin. This observation is an artifact of our use of an upper bound on the gene probe copy number values (9 by default) in FISHtrees, causing it to replace GD events that result in copy number profiles with higher gene copy number values with SD and CD events.

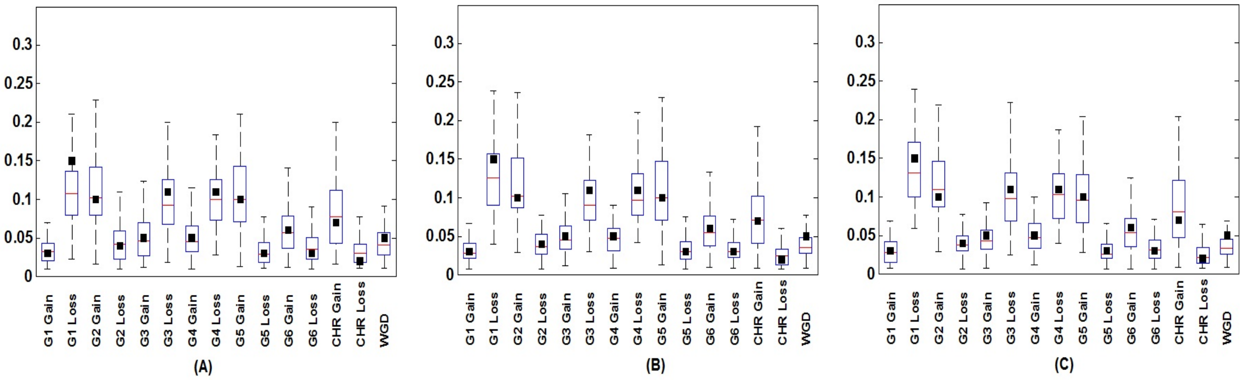

Figure S5: Inferred parameter values from simulated data using parameter set 4 for sample sizes of (A) 100, (B) 200 and (C) 300 cells. The true parameter values used in the simulation are shown as black squares.

In Figure S5, we show the boxplots for the parameters when the number of cells sampled from the trees are varied. Figure S5(A),(B) and (C) show the parameter values for sample sizes of 100, 200 and 300 cells, respectively. From the plots, we see that our algorithms can infer the parameters accurately even for sample sizes as low as 100 cells, with similar accuracy and variance of estimates across a range of realistic data sizes.

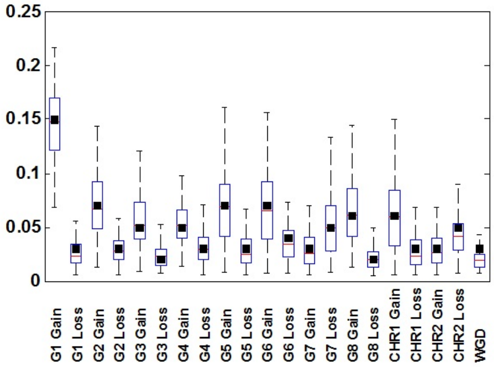

Figure S6: Boxplots showing inferred parameter values for the simulated dataset with two chromosomes each containing two genes. The real parameter values used in the simulation are shown as black squares.

In Figure S6, we show the inferred parameter values when the number of chromosomes with multiple probes is increased from 1 to 2. It can be seen that the algorithm can infer the parameters with high accuracy in this case, too. So, the results show that the algorithm can recapitulate the parameter values with high accuracy across a range of parameter values and sample sizes. The variance across trees is sufficiently high to suggest, however, that making reliably accurate estimates for single trees may require larger numbers of single cells than are yet available in real data.

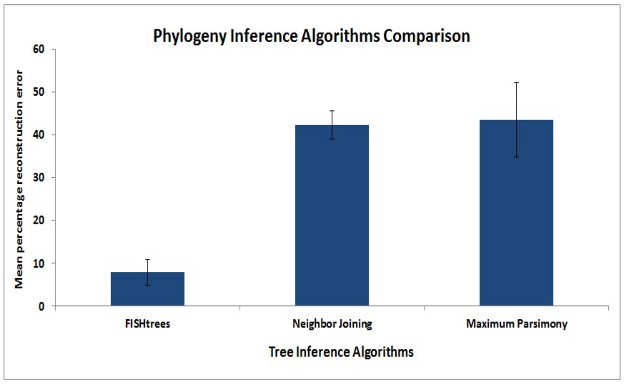

Figure S7: Accuracy of phylogeny reconstruction for our FISHtrees algorithm, neighbor joining (NJ), and maximum parsimony (MP), as assessed by reconstruction error E, for parameter combination 3.

We next examine accuracy of tree reconstruction on these data sets. Figure S7 compares accuracy at the level of phylogenetic bipartitions for our FISHtrees algorithm, NJ, and MP. FISHtrees outperforms NJ and MP by a large margin in inferring the phylogenies. The mean reconstruction error of FISHtrees, NJ and MP are 7.92% (s.d. 2.97%), 42.22% (s.d. 3.28%) and 43.49% (s.d. 8.72%) respectively. These results show that error is substantially reduced in inferring tumor phylogenies with algorithms that consider tumor-like chromosome abnormalities rather than using off-the-shelf traditional phylogeny building algorithms.

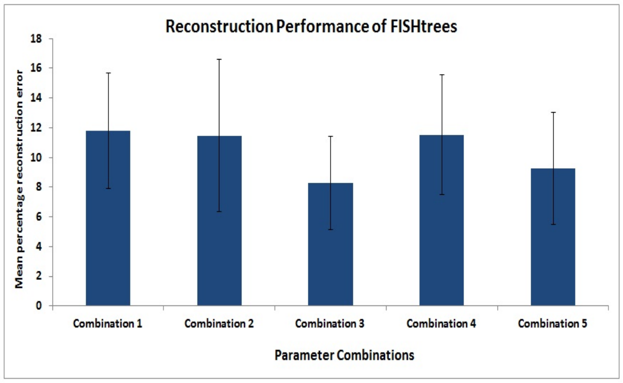

Figure S8: Accuracy of phylogeny reconstruction for our FISHtrees algorithms across the five parameter combinations considered.

In Figure S8, we show the reconstruction accuracy of FISHtrees across the five different combinations of parameters. This plot shows that our algorithm yields comparable accuracy across the ranges of parameter values considered. While there is some variability across the parameter combinations, there are no obvious trends suggesting what features might make a particular phylogeny easier or harder to infer.

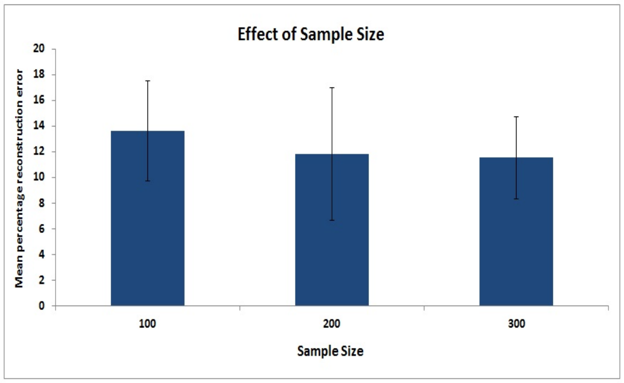

Figure S9: Reconstruction performance of FISHtrees across different sample sizes.

Finally, in Figure S9, we show the performance of FISHtrees in inferring the phylogenies when the number of samples sizes are varied across 100, 200 and 300 cells per tumor. From this plot, we can see that FISHtrees can infer the phylogenetic trees with high accuracy when the number of sampled cells is as low as 100. The accuracy does, however, continue to improve as the number of sampled cells is increased.

We note that run times were modest for the data sets examined in this work. For the 300 cell tests, FISHtrees required an average of 20.45 minutes of wall clock runtime on a standard desktop Linux computer per set of 100 replicates, averaged across the five simulated parameter sets. Thus, it can be seen that FISHtrees can generate the tumor phylogenies across a diverse range of parameter values on realistic sizes of data set in relatively short computational time.

References

Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., and Schwartz, R. (2013). Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*, **29**(13), i189–i198.

Chowdhury, S. A., Shackney, S. E., Heselmeyer-Haddad, K., Ried, T., Schäffer, A. A., and Schwartz, R. (2014). Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comp. Biol.*, **10**(7), e1003740.

Heselmeyer-Haddad, K., Berroa Garcia, L. Y., Bradley, A., Ortiz-Melendez, C., Lee, W., Christensen, R., Prindiville, S. A., Calzone, K. A., Soballe, P. W., Hu, Y., *et al.* (2012). Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity, yet conserved genomic imbalances and gain of *MYC* during progression. *Am. J. Pathol.*, **181**(5), 1807–1822.

Lin, Y., Rajan, V., and Moret, B. M. E. (2012). A metric for phylogenetic trees based on matching. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**(4), 1014–1022.

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, **194**(4260), 23–28.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biol Evol*, **30**(12), 2725–2729.