SCIENTIFIC REPORTS

OPEN

Comprehensive statistical inference of the clonal structure of cancer from multiple biopsies

Jie Liu [1], John T. Halloran[2], Jeffrey A. Bilmes[2], Riza M. Daza[1], Choli Lee [1], Elisabeth M. Mahen[3,4,5], Donna Prunkard[6], Chaozhong Song[3,4,5], Sibel Blau[3,7], Michael O. Dorschner[3,6], Vijayakrishna K. Gadi[8,9], Jay Shendure[1,10], C. Anthony Blau[3,4,5] & William S. Noble [1,11]

A comprehensive characterization of tumor genetic heterogeneity is critical for understanding how cancers evolve and escape treatment. Although many algorithms have been developed for capturing tumor heterogeneity, they are designed for analyzing either a single type of genomic aberration or individual biopsies. Here we present THEMIS (Tumor Heterogeneity Extensible Modeling via an Integrative System), which allows for the joint analysis of different types of genomic aberrations from multiple biopsies taken from the same patient, using a dynamic graphical model. Simulation experiments demonstrate higher accuracy of THEMIS over its ancestor, TITAN. The heterogeneity analysis results from THEMIS are validated with single cell DNA sequencing from a clinical tumor biopsy. When THEMIS is used to analyze tumor heterogeneity among multiple biopsies from the same patient, it helps to reveal the mutation accumulation history, track cancer progression, and identify the mutations related to treatment resistance. We implement our model via an extensible modeling platform, which makes our approach open, reproducible, and easy for others to extend.

Cancer is heterogeneous in the sense that the cancer cells in a tumor are not genetically identical, but form distinct clones, defined as subpopulations of cancer cells that host the same genomic aberrations. In aggressive and metastatic cancers, these genomic aberrations quickly evolve, resulting in extreme spatial and temporal heterogeneity[1,2]. Therefore, multiple biopsies over different locations and at different time points need to be collected and sequenced in order to capture the complexity of tumor genomic landscapes and provide insight into how tumors evolve and escape treatment[3,4]. Accordingly, computational tools are needed to accurately characterize the clonal structure of cancer and reveal how that structure evolves over time.

In recent years, a large number of computational tools and statistical models have been developed to analyze tumor heterogeneity from DNA sequencing data (Table 1). However, most of these tools only model one type of genomic aberration, such as single-nucleotide variants (SNVs), copy number alterations (CNAs), or structural variants. Restricting the analysis to a single type of mutation not only reduces statistical power to accurately detect the clonal structure within the tumor, but also prevents us from understanding interactions among different types of mutations. Furthermore, many SNV-based methods assume that no copy number changes have occurred, which is extremely improbable. Therefore, their estimation of the prevalence of a given clone can be inaccurate, and the corresponding heterogeneity results may be misleading. Existing methods that capture SNVs and CNAs in the same model (i.e., phyloWGS[5], SPRUCE[6] and Canopy[7]) require running a CNA-calling algorithm before heterogeneity analysis, but accurate CNA characterization also depends on heterogeneity analysis.

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Department of Electrical Engineering, University of Washington, Seattle, WA, USA. [3]Center for Cancer Innovation, University of Washington, Seattle, WA, USA. [4]Institute for Stem Cell and Regenerative Medicine, University of Washington, Seattle, WA, USA. [5]Department of Medicine/Hematology, University of Washington, Seattle, WA, USA. [6]Department of Pathology, University of Washington, Seattle, WA, USA. [7]Northwest Medical Specialties, Puyallup and Tacoma, WA, USA. [8]Department of Medicine/Oncology, University of Washington, Seattle, WA, USA. [9]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, 98109, WA, USA. [10]Howard Hughes Medical Institute, Seattle, WA, USA. [11]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. Correspondence and requests for materials should be addressed to C.A.B. (email: tblau@uw.edu) or W.S.N. (email: william-noble@uw.edu)

1

Software	Year	SNV	CNA	SV	Multiple	Model/Algorithm
OncoSNP[17]	2010		✓			Mixture model, EM, Bayesian methods
TuMult[18]	2010		✓		✓	Breakpoint distance
GRAFT[19]	2012		✓			Partial maximum likelihood
ABSOLUTE[20]	2012	✓	✓			Maximum likelihood
TrAp[21]	2013	✓				Exhaustive search under constraints
THetA[22,23]	2013		✓			Maximum likelihood
CancerTiming[24]	2013		✓			Maximum likelihood
OncoSNP-seq[25]	2013		✓			Mixture model, EM, Bayesian methods
PyClone[26]	2014	✓			✓	Dirichlet Process, beta-binomial/MCMC
SciClone[27]	2014	✓			✓	Beta mixture model/variational Bayes
Clomial[28]	2014	✓			✓	Binomial mixture model, EM
CloneHD[29]	2014	✓	✓		✓	HMM, EM, variational Bayes
MEDICC[30]	2014		✓		✓	Finite state transducer, minimum-event distance
TITAN[9]	2014		✓			Two-chain factorial HMM/EM
SubcloneSeeker[31]	2014	✓	✓	✓		Clustering, enumeration and co-localization prediction
BTP[32]	2014	✓				Binary tree partition
BreakDown[33]	2014			✓		Maximum likelihood
PhyloSub[34]	2014	✓			✓	Tree-structured stick-breaking process prior, MCMC
BayClone[35]	2015	✓			✓	Categorical Indian Buffet Process
PhyloWGS[5]	2015	✓	✓		✓	Tree-structured stick-breaking process prior, MCMC
CITUP[36]	2015	✓			✓	Quadratic integer programming
LICHeE[37]	2015	✓			✓	Clustering and evolutionary constraint network
AncesTree[38]	2015	✓			✓	Integer linear programming
SPRUCE[6]	2016	✓	✓		✓	Combinatorial enumeration
Canopy[7]	2016	✓	✓		✓	MCMC
THEMIS (our work)	2017	✓	✓		✓	Dynamic graphical models

Table 1. Software tools for characterizing within-patient and intra-tumor heterogeneity and their features, including whether they capture SNVs, CNAs and structural variants (SVs), whether they support multiple biopsy analysis, and their key models and algorithms.

Most existing tools are designed to analyze a single tumor biopsy and are not suitable for jointly analyzing multiple biopsies. As DNA sequencing becomes more affordable, we can more easily collect multiple biopsies from a single patient during treatment. If we only perform heterogeneity analysis on the individual biopsies, then we are unable to detect clones that are shared across different biopsies from the same patient, and we fail to address important questions about how the tumor cells evolve, metastasize and escape treatment.

Finally, although most models are free and publicly available, it is difficult to extend them by adding new assumptions and new types of biological data. Even under the best of circumstances, significant effort is required for users to fully understand the source code. In many situations, data structures and computational algorithms prohibit other investigators from modifying the model to accommodate their special needs.

To address these challenges, we propose THEMIS (Tumor Heterogeneity Extensible Modeling via an Integrative System), which allows us to jointly characterize different types of genomic aberrations from multiple biopsies using a dynamic graphical model. We implement our model via an extensible modeling platform, the Graphical Models Toolkit (GMTK)[8], which makes our approach open, reproducible and easy for others to extend. To extend the model, users only have to modify the model specification files; GMTK then automatically handles the required computation. Simulation experiments demonstrate that THEMIS significantly increases the accuracy of recovering tumor subclones and their genotypes, compared with its ancestor, TITAN[9]. Single cell DNA sequencing confirms that individual nuclei can be segregated into one of the two tumor subclones identified by THEMIS. We applied THEMIS to three tumor biopsies from one cancer patient, thereby revealing the mutation accumulation history of the patient, tracking cancer progression, and identifying mutations related to developing resistance following various treatments.

Results

The Model. From bulk next generation sequencing data, we define two primary observations at each genomic position: the allelic ratio, defined as the proportion of the reads containing a specified allele among all reads aligned to the site, and the log ratio between tumor read depth and normal read depth (Fig. 1a). From these
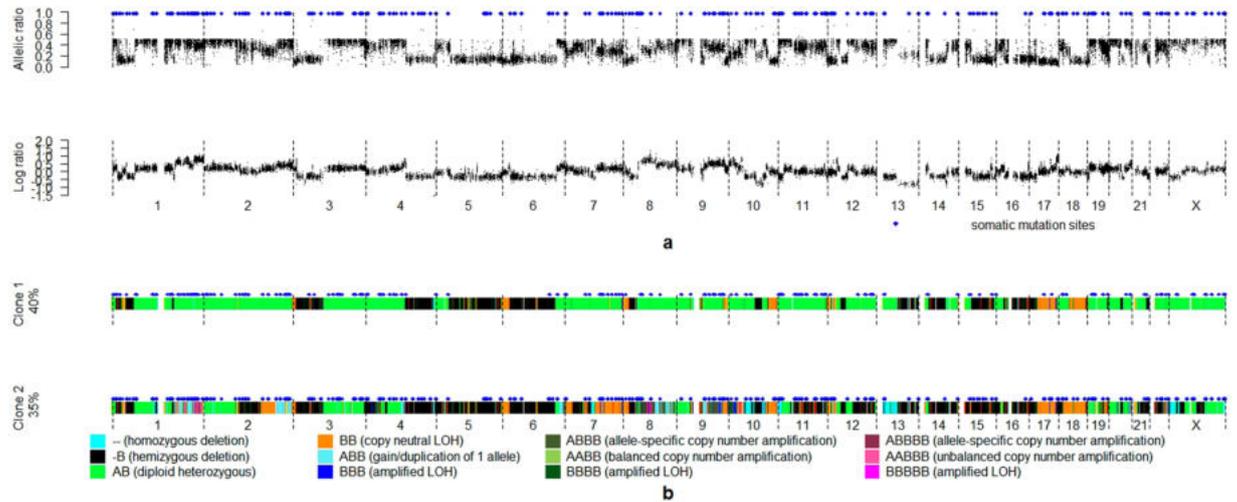
Figure 1. Example THEMIS input observations and the corresponding inferred outputs. (**a**) Inputs to THEMIS, including allelic ratio, log ratio and genomic position information. Somatic mutation sites are indicated by blue diamonds. (**b**) Outputs of THEMIS show that there are two tumor clones in the tumor biopsy, one parent tumor clone with 40% cell prevalence and one child tumor clone with 35% cell prevalence. The CNAs of the two tumor clones are visualized with color bars across the genome, and the genomic positions hosting an SNV are indicated by blue diamonds.

inputs, we aim to infer the number of distinct clones, the full genotype of each clone, and the prevalence of each clone within each biopsy (Fig. 1b). To carry out this inference, TITAN[9] uses a dynamic graphical model, in which each frame represents one genomic position, and the allelic ratio and tumor/normal log ratio are observed at each frame. The backbone of the TITAN model consists of two hidden Markov chains, one representing the genotype of the CNA event at the current position, and the other representing the clone in which the CNA event occurs. Our model, THEMIS, is similar to TITAN in the sense that both models are dynamic graphical models with each frame representing a single genomic position, with CNA events captured by hidden Markov chains. However, THEMIS extends TITAN by (1) jointly accounting for SNVs and CNAs, (2) jointly analyzing multiple biopsies, (3) estimating transition probabilities between hidden states of the model from observed data rather than fixing them at specific values, and (4) using an open and extensible modeling language (GMTK[8]). More details about the THEMIS model, including the modeling choices and assumptions, model's structure, variables and parameters, are provided in Methods section.

Simulation results. We first used simulated data to compare the performance of THEMIS and TITAN[9]. As a starting point for the simulation, we used the genomic positions measured in three tumor biopsies from three patients with triple negative breast cancer (Supplementary Table 1). For each set of genomic positions, we also specified three different sets of tumor subclone compositions. More details about the simulation experiments are provided in Supplementary Note 1. We evaluated (1) the percentage of sites at which the hidden genotype was incorrectly inferred, (2) the percentage of sites at which the clonal/subclonal status was incorrectly inferred, and (3) the percentage of sites at which either the genotype or clonal/subclonal status were incorrectly inferred. THEMIS outperformed TITAN in recovering the clonal/subclonal status and genotypes of the genomic positions in all experiments (two-sided paired t-test, $p = 0.00137$ for genotype recovery, $p = 1.152 \times 10^{-7}$) for clonal/subclonal status recovery, and $p = 0.00198$ for both genotype and clonal/subclonal status recovery), and reduced the recovery error by 13.3% on average (Supplementary Table 2). Not surprisingly, both THEMIS and TITAN performed better when the prevalence of the somatic events was higher.

Validation via single-cell DNA sequencing. The two tumor clones in Fig. 1b were identified from the bulk DNA sequencing data from an involved axillary node in a patient with metastatic triple negative breast cancer. We single-cell sequenced a second sample taken from the same axillary node at the same time, aiming to validate the subclones previously identified from bulk DNA sequencing. From a total number of 143,108 nuclei, 96 nuclei were fluorescence activated cell sorted (FACS) using gating to select for tumor cell nuclei, placed on a 96-well plate, and whole genome amplified (GenomiPhi). Indexed Nextera libraries were sequenced on the NextSeq using PE 150 bp mid-output flow cell. Six cells were removed due to extremely low numbers of reads after adapter removal (Supplementary Fig. 1). The sequencing data from another eleven nuclei was of low quality, as evidenced by a much larger fraction of short reads (Supplementary Fig. 2) and were excluded from our analysis. The remaining 79 cells were used for validation.

We use a Bayesian classifier (Supplementary Note 2) to assign each of the 79 cells into one of the three clones identified from the bulk sequencing data by THEMIS—one normal clone, one parent tumor clone and one child tumor clone. As input to the classifier, we use sequencing coverage on three types of genomic regions which are derived from the inferred genome-wide genotype for the two tumor clones (Fig. 1b), namely, clonal 2-copy (i.e. normal) regions, clonal 1-copy (i.e. loss of heterozygosity [LOH]) regions and subclonal 1-copy regions (Fig. 2b
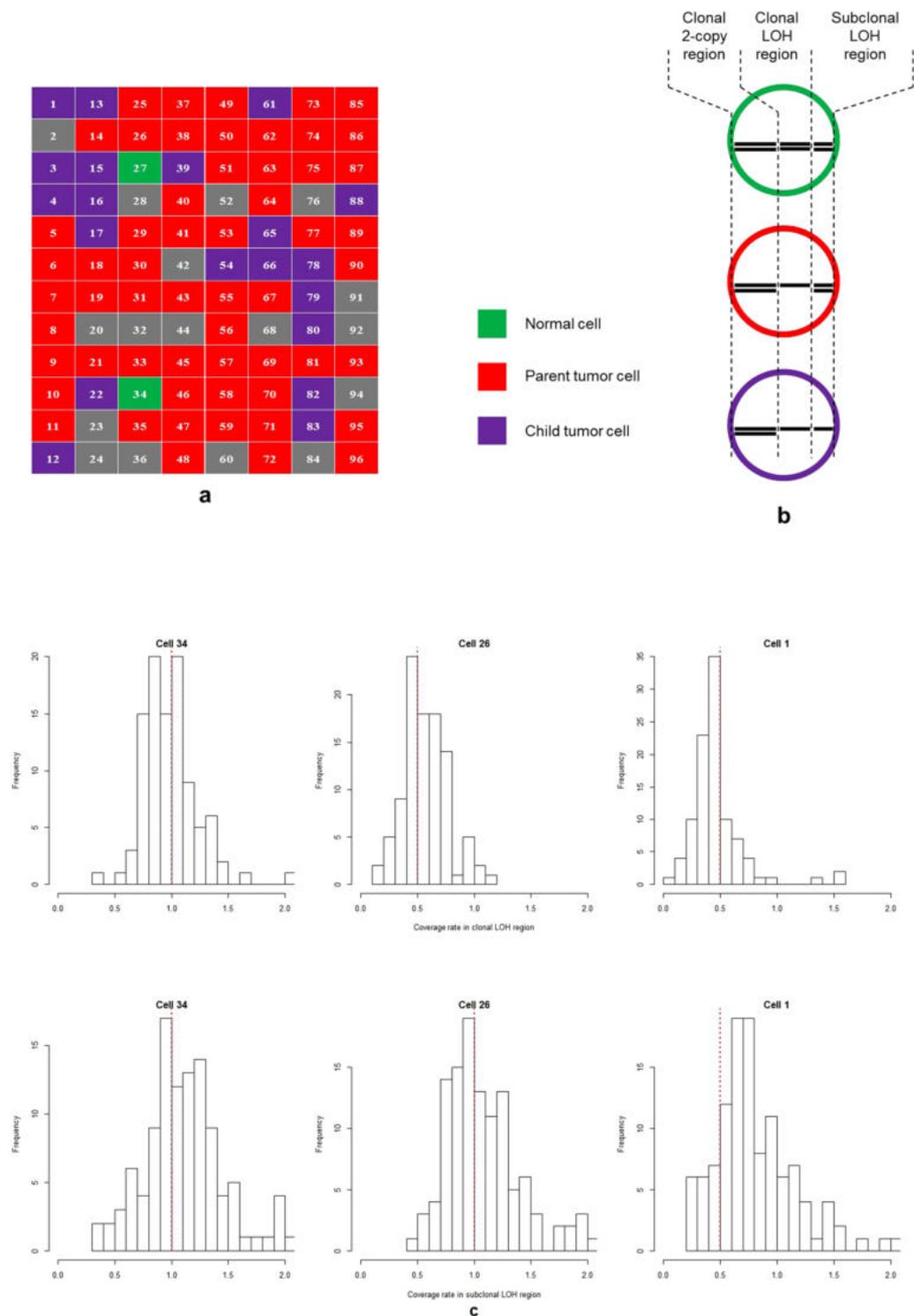
Figure 2. Results of single cell validation experiments. (**a**) Predicted cell category from the Bayesian classifier: normal cell (green), parent tumor cell (red), child tumor cells (purple) and unknown due to low sequencing quality (grey). (**b**) Three different types of regions that we use to distinguish the three types of cells, namely clonal 2-copy regions, clonal LOH regions and subclonal LOH regions. (**c**) Histograms of the relative coverage rate in clonal LOH segments and subclonal LOH segments demonstrate cell 34 is normal, cell 26 is from the parent tumor clone and cell 1 is from the child tumor clone. Red dotted lines in the histograms indicate the expected coverage rates in the cells.

and Supplementary Table 3). Clonal 2-copy regions provide the baseline measurement of sequencing coverage in normal regions. Clonal LOH regions distinguish tumor from normal. Subclonal LOH regions distinguish the child tumor clone from its parent. The Bayesian classifier identifies 2 normal cells (indicating that FACS gating
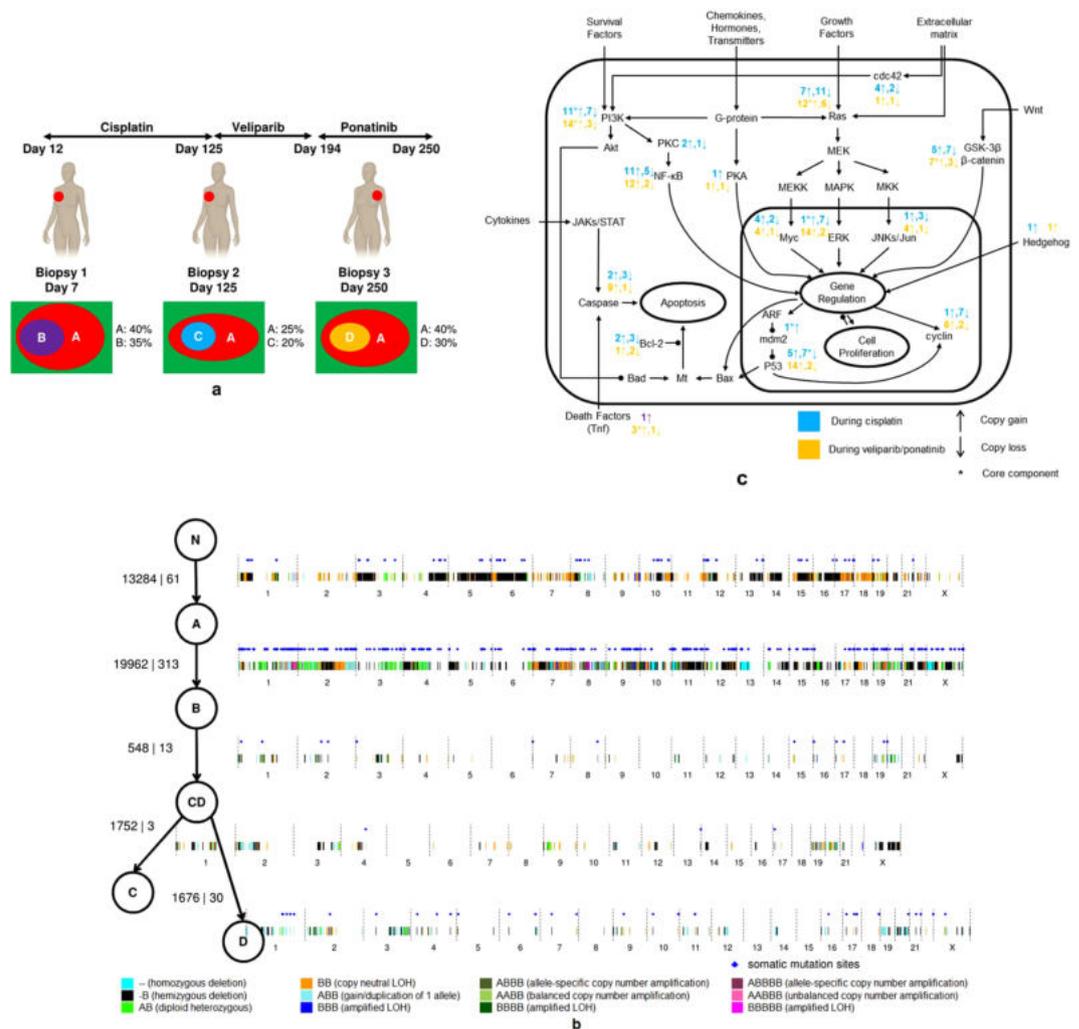
Figure 3. Experiments and results from joint analysis of three biopsies from the same patient. (**a**) The collection of the three biopsies from the same patient during three stages of treatment and the inferred tumor clones in the biopsies. (**b**) The recovered phylogenetic tree and the mutations accumulated on each edge. The mutations are shown on a genome-wide plot, and the two numbers on each edge are the number of germline heterozygous sites affected by these CNAs and the number of SNVs. (**c**) Signal pathways (a simplified version of signal transduction pathways from Wikipedia), and the number of genes with copy number changes (copy gain and copy loss) in different stages of cancer progression. *Denotes at least one of the mutated genes is a core component of the signaling pathway.

worked), 57 parent tumor cells and 20 child tumor cells (Fig. 2a and Supplementary Table 4). The observed ratio of parent to child tumor cells (2.85) does not agree with the ratio inferred by THEMIS (1.14), which is likely attributable to the very small number of examined events and that the single cell analysis was performed in a separate sample taken at the same time. However, the histograms of the normalized coverage rate in the 99 clonal LOH segments and 119 subclonal LOH segments in the three types of cells validate that the three categories of nuclei agree with our model (Fig. 2c). For example, cell 34 has no LOH events in the clonal LOH region nor in the subclonal LOH region, and is therefore a normal cell. Cell 26 displays LOH events in the clonal LOH region, but not in the subclonal LOH region, and is therefore identified as belonging to the parent tumor clone. Cell 1 has LOH events both in clonal and subclonal LOH regions, and is therefore identified as belonging to the child tumor clone. The histograms of the coverage rates from the aggregated 57 parent tumor cells show similar patterns to that of cell 26, whereas the histograms from the aggregated 20 child tumor cells show similar patterns as cell 1 (Supplementary Fig. 3). Therefore, our single cell experiment successfully validates the subclones identified by THEMIS.

Joint analysis over multiple biopsies from the ITOMIC study. The Intensive Trial of OMics In Cancer (ITOMIC-001) enrolls patients with metastatic triple negative breast cancer in whom biopsies of multiple metastatic sites are performed repeatedly over time[10]. For each patient, multiple biopsies are evaluated using next generation sequencing. We performed joint heterogeneity analysis on three biopsies from Patient 1 (Fig. 3a). Patient 1 was originally diagnosed with triple-negative breast cancer in February 2011, and enrolled in ITOMIC-001 in

October 2013. The patient then received three different treatments, including cisplatin (between study days 12 and 125), an investigational PARP inhibitor veliparib (between study days 126 and 194), and the kinase inhibitor ponatinib (from study day 195 until the time of her death on study day 250). The first biopsy B1 was sampled from an involved right axillary lymph node, collected on study Day 7 (before cisplatin). The second biopsy B2 was sampled from the same right axillary lymph node on study Day 125 (after cisplatin). The third sample B3 was from a left peribronchial lymph node, collected at autopsy following the patients death on study day 250 (post ponatinib).

We followed a five-step procedure to recover the phylogenetic tree from the three tumor biopsies (see Methods for details), which involves enumerating all possible candidate phylogenies from individual biopsy analysis and then selecting the best phylogeny by likelihood (Supplementary Note 3). The recovered phylogenetic tree (Fig. 3b) reveals the relationships among the recovered clones and the mutations accumulated at each stage of cancer progression. Parent clone A is shared by all three biopsies, and child clones (B, C, and D) inherit the mutations present in parent clone A and acquire new mutations of their own. Although clones B, C and D occur in separate biopsies, clones C and D appear to be descendants of clone B, meaning that the new mutations detected in later biopsies occurred within the child tumor clones. In addition to clones A, B, C and D, the phylogenetic tree includes an internal node for an inferred intermediate clone (CD) hosting the mutations shared between clone C and clone D, and corresponding to a splitting point between clones C and D in cancer progression. On the phylogenetic tree, we also label the new CNAs and SNVs on the edges. We visualize these mutations on a genome-wide plot, and provide the number of germline heterozygous sites affected by these CNAs and the number of the SNVs (Fig. 3b). The mutations on the edges of the phylogenetic tree reveal the mutation accumulation history of this patient and can help in tracking mutations related to treatment resistance. By the time the patient joined the study, there had been two major phases of mutation accumulation, one corresponding to the mutations accumulated in clone A and the other corresponding to mutations accumulated in clone B. Comparing the two phases, more SNVs emerged in the second phase. After joining the study and receiving further treatments, additional mutations emerged. The mutations on the edges between clone B and clone C emerged during treatment with cisplatin, whereas those on the edge between clone CD and clone D emerged during treatment with the PARP inhibitor veliparib (without response) followed by treatment with the kinase inhibitor ponatinib (which did yield a partial response). Because veliparib failed to affect tumor growth, we attribute the changes associated with the CD → D transition to ponatinib, which was given based on the presence of two activating mutations affecting FGFR2 (S252W;Y375C) (manuscript submitted).

A number of intriguing patterns were revealed when we looked at the mutations associated with different phases of treatment. First, many of the genes in CNA regions are known to be related to cancer. On the CNAs associated with the three treatments, including 147 CNAs during the treatment with cisplatin and 98 CNAs during the treatment with veliparib and ponatinib, we identify 848 genes and 519 genes, respectively. We queried these genes on the NCBI gene website (www.ncbi.nlm.nih.gov/gene) and retrieved 186 genes and 175 genes, respectively, that are known to be related to cancer in the literature. The retrieved genes are related to many important cancer signaling pathways (Fig. 3c). On some of the cancer signaling pathways (i.e., MEK/MAPK/Erk, PI3K/Akt/mTor, NF-κB, and p53), core genes (i.e., MAP2K7, PIK3C2A, PIK3CD, PIK3CB, TNFSF9, TNFSF14, NRAS, GSK3B, Notch3, TP53TG5 and SNAI1) experienced copy-number changes during different phases of treatment. In addition, the genes mutated during different phases of treatment show patterns that are potentially illustrative of different therapeutic responses. The proportion of genes experiencing copy number gains in later stages (i.e., on the edge CD → D, during the treatments with veliparib and ponatinib) is much higher than that in earlier stage (on the edges B → CD and CD → C, during the treatment with cisplatin), and the proportion difference is more dramatic in important cancer signaling pathways, including MEK/MAPK/Erk, PI3K/Akt/mTor and NF-κB. These genes mutated in the different phases of treatment also showed different functional focuses. According to DAVID[11,12], the top function clusters among the genes mutated in the earlier stage are rho GTPase activation, growth factor, DNA damage and ErbB signaling. The top function clusters from the genes mutated in the later stage are DNA damage, ras signaling, nucleotide-binding, zinc-finger, neurotrophin signaling, and endocytosis. Third, a number of SNVs occurred on or near the genes known to be related to cancer signaling pathways, which allows us to investigate them together with the genes mutated due to copy number changes. During treatment with cisplatin, one SNV occurred near an intron/extron boundary within BIN1, which is known to interact with the myc oncoprotein as a putative tumor suppressor. During the treatments with veliparib and ponatinib, SNVs occurred on or near LATS1 (a core component of Hippo-YAP pathway), MGMT (related to DNA damage), IL17RB (related to NF-κB signaling) and APCDD1L (related to wnt signaling), all of which are known to be related to breast cancer. Although additional experiments are needed for further validation, THEMIS provides a powerful computational tool to generate hypotheses from multiple biopsy DNA sequencing data.

Discussion

THEMIS offers a powerful and extensible modeling framework to jointly capture different types of genomic aberrations in the analysis of multiple biopsies. The integration of CNAs and SNVs in the heterogeneity analysis increases the accuracy of clonal inference relative to previous methods that consider only single types of mutations. For example, if we observe an allelic ratio 0.3 at one genomic position, then the cell prevalence of the SNV should be 60% if the corresponding genotype is AB, but the prevalence should be 85.7% if the genotype is AAB. In such cases, methods that fail to jointly consider copy number information and SNVs can be misled. In addition, the integration of multiple types of mutations allows us to understand cancer comprehensively, and to address important questions such as how the different types of mutations cooperate with each other and what roles they play at different stages of cancer progression.

The joint analysis over multiple biopsies from the same patient provides a complete picture of mutation progression in the patient, which may shed light on how tumor cells escape treatment and metastasize. The ability to

analyze multiple biopsies jointly will be increasingly important as DNA sequencing costs continue to decrease. The current turnaround time of analyzing the three biopsies with our model, including both data preprocessing and model running, is just a few hours.

Because THEMIS is built using a general purpose graphical models toolkit, the approach is easy to extend to alternative model architectures. For example, during review of this manuscript, one reviewer suggested that the THEMIS model likely over-segments the genome. We verified this effect empirically and then demonstrated the flexibility of the THEMIS framework by modifying the model to incorporate a user-specified constraint on the number of segments (Supplementary Note 4). In addition, GMTK provides flexible calculation in both estimation and inference, including both exact and approximate inference algorithms. Based on the available computing resources, the user can easily trade memory with running time. Using this modeling and algorithmic flexibility, we plan in future work to extend THEMIS to account for more complex types of mutations, such as chromo-thripsis and chromoplexy. We also plan to incorporate a more principled phylogeny reconstruction method into THEMIS. Ultimately, THEMIS will provide a testbed for model development by us and others interested in modeling the full complexity of tumor evolution.

Data availability.   The bulk DNA sequencing data and the single cell DNA sequencing data used in our analysis can be downloaded from Sequence Read Archive with accession SRP102304.

Software availability.   THEMIS is available at https://github.com/jieliu6/THEMIS.

Methods

Data preprocessing.   We assume that next generation sequencing data was mapped to the reference genome, and the mapped BAM files are ready for analysis. Pre-processing of the data consists of three steps (Supplementary Fig. 4). First, we identify the genomic sites that will be included in the model. Our model captures both CNA events and SNV events; therefore, two types of genomic sites are included. For CNAs, we consider *germline heterozygous sites* since we can monitor not only absolute copy number changes (via tumor-normal read depth difference), but also what happens to the two individual copies (via allelic imbalance). For SNVs, we consider the *somatic mutation sites* which host an SNV event in any of the tumor biopsies. From the germline (normal) BAM file, we use Samtools to identify germline heterozygous sites. From the tumor and normal BAM files, we use MuTect[13] to identify somatic mutation sites. Second, we filter out unreliable sites and reads using MuTect[13]. Third, we adjust for GC content and mappability. Short reads from next generation sequencers are not uniformly distributed across the genome—more reads are expected to be obtained from regions with higher GC content and mappability. The bias cannot fully be adjusted by normalizing with another next generation sequencing library (e.g. from a normal biopsy) from the same patient[14]. We therefore use HMMcopy[15] to adjust GC content and mappability in the read counts.

Modeling choices in THEMIS.   Unlike previous methods such as phyloWGS[5], SPRUCE[6] and Canopy[7], which capture CNA or SNV events as the entities in the model, our model THEMIS and its predecessor TITAN directly model individual genomic positions as the entities in the model and therefore have the ability to perform CNA calling during tumor heterogeneity analysis. Both THEMIS and TITAN are dynamic graphical models with each frame representing a single genomic position, with CNA events captured by hidden Markov chains. Therefore, THEMIS inherits five key assumptions from TITAN:

1. Two primary observed variables—allelic imbalance and the tumor-normal read depth ratio—reflect the underlying somatic genotype of the tumor at germline heterozygous sites.
2. CNA events span multiple contiguous germline heterozygous sites.
3. The observed NGS data comes from heterogeneous cellular populations, including normal cells and tumor subpopulations.
4. Two mutation events are observed at the same cellular prevalence if and only if the two events come from the same subpopulation.
5. Only one CNA event can arise in only one tumor subpopulation at each genomic position.

Note that Assumption 4, although used by many tumor heterogeneity models, can be invalid if two different tumor subclones in a tumor have the same cellular prevalence. The purpose of introducing Assumption 5 is to make the heterogeneity model simple and identifiable; however, this assumption does prevent us from modeling more complicated situations in which multiple CNAs arise in the same genomic region.

We usually have around 30–50 thousand germline heterozygous sites and several hundred somatic mutation sites in whole-exome sequencing data from a single biopsy. With reasonable sequencing depth (greater than ~100 reads per position, on average) the underlying genotypes (i.e. the type of the CNA event) estimated from the contiguous germline heterozygous sites can be inferred accurately. Integrating the somatic mutation sites and germline heterozygous sites using two factorial Markov chains allows us to model sites that harbor both a CNA event and a somatic mutation. In the situation when the observed variables at one somatic mutation site suggest that the genotype or the subclone assignment at that site disagrees with the neighboring germline heterozygous sites, THEMIS can still infer the correct hidden genotype and subclone assignment based on the observed variables at the somatic mutation site. Furthermore, because there will typically be many contiguous germline heterozygous sites before and after this somatic mutation site, the disagreement will not be propagated to nearby germline heterozygous sites.
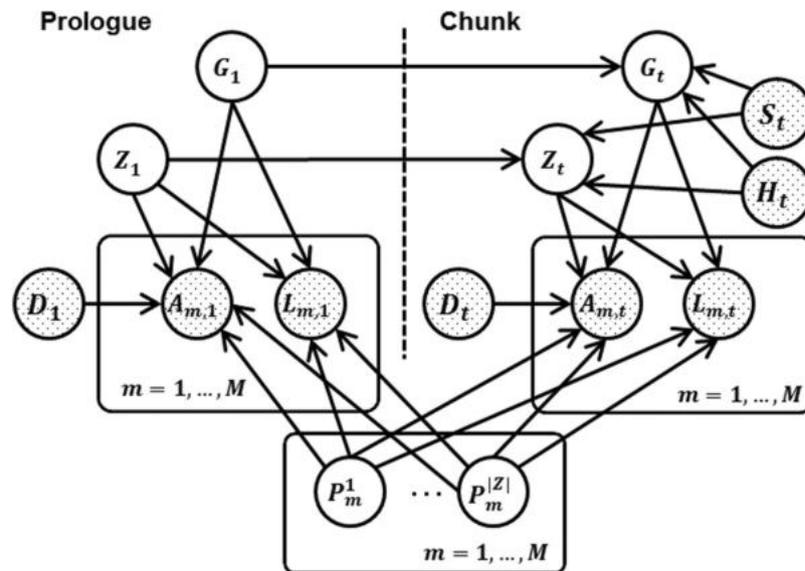
Figure 4. The THEMIS model.

We adopted these particular modeling choices and assumptions based on the sequencing quality and depth in our data. However, we encourage users to adjust these modeling choices and assumptions as appropriate for their own data. The extensible modeling platform employed by THEMIS should make it easy to implement variants of the model proposed here.

Structure of the THEMIS model. The THEMIS dynamic graphical model can be represented using a standard "plate" representation. In Fig. 4, the "Prologue" represents the start of the model, and the "Chunk" represents all random variables associated with a single genomic position. In practice, the chunk is copied multiple times so that the length of the model matches the length of the observed data (i.e., the total number of genomic positions). In the figure, each vertex represents one random variable at a particular genomic position. If a vertex is shaded, then the corresponding random variable is observed; otherwise, the random variable is hidden. The variables and parameters used in the model are explained in the next two sections and summarized in Supplementary Table 5. We use capital letters to denote random variables and the corresponding lower-case letters to denote the particular values of the random variables. If a lower case letter has a bar on top, then it is observed; otherwise, it is inferred.

The backbone of the model consists of two Markov chains, i.e., two hidden variables at each site t, corresponding to the unknown genotype of the mutation (G_t) and an indicator (Z_t) of which clone the mutation occurs in. The two Markov chains capture the phenomenon that CNA events span multiple contiguous genomic positions, and the corresponding most probable states that we infer from the observed variables are essentially the output of our THEMIS model. At site t and in biopsy m, a set of observed variables represent the allelic ratio ($A_{m,t}$) and the log ratio between tumor read depth and normal read depth ($L_{m,t}$). Other useful information about the genomic position is also captured in the model as observed variables, including the type of site (D_t), an indicator variable for the first site of a chromosome (S_t), and the distance from its previous site (H_t). For each biopsy m, the model contains an additional set of $|Z|$ hidden variables, $P_m^1, P_m^2, \ldots, P_m^{|Z|}$, denoting the prevalence levels of the clones.

The conditional independence relationships among the variables are encoded by the edges either within a genomic position or between two adjacent genomic positions. At each site, the model specifies the probability of the observed variables given the hidden variables, which captures how different genotypes and the occurrence in different clones, in combination, make the allelic ratio and log ratio different in tumor biopsies. At a germline heterozygous site (where $D_t = 0$), the allelic ratio reflects how the allelic ratio is different from 0.5, which is expected in a normal cell. At a somatic mutation site (where $D_t = 1$), the allelic ratio reflects how the allelic ratio is different from 0, which is expected in a normal cell. Therefore, the parents of the allelic ratio ($A_{m,t}$) include the genotype (G_t), the clone index (Z_t), the prevalence levels ($P_m^1, P_m^2, \ldots, P_m^{|Z|}$) and the type of site ($D_t$). Because the log ratio of the tumor-normal read depth difference ($L_{m,t}$) does not depend on the type of site, its parents include the genotype (G_t), the clone index (Z_t), and the prevalence levels ($P_m^1, P_m^2, \ldots, P_m^{|Z|}$). Between any two adjacent sites, we specify the transition probability between genotypes and the transition probability between clones. The variable H_t is the distance (in base pairs) between site t and its previous site $t-1$. We set transition parameters (for both G_t and Z_t) as functions of $\bar{h}_t$ (the observed value of H_t) to capture the phenomenon that the chance G_t and G_{t-1} agree decreases as $\bar{h}_t$ increases (and similarly for Z_t and Z_{t-1}). Therefore, H_t is a parent of G_t and Z_t at a non-start-of-a-chromosome site t. When the current site is the start of a chromosome ($S_t = 1$), G_t and Z_t do not depend on G_{t-1} and Z_{t-1}, but follow prior distributions (π_G and π_Z). Therefore, S_t is also a parent of G_t and Z_t.

Variables in the THEMIS model. The variables in the THEMIS model are either observed variables or hidden variables. The observed variables are directly obtained from the data, whereas the most probable states of the hidden variables must be inferred, given the observed variables and the trained parameters. Each frame of the THEMIS model contains five observed variables. Two of these are key signals to detect mutations, and they are modeled as Gaussians:

- The allelic ratio $A_{m,t}$ at site t in biopsy m ($\forall m = 1, \ldots, M$) is modeled as a Gaussian variable, i.e.,

$$P(A_{m,t} = a | Z_t = z, G_t = g, P_m^z = p_m^z, D_t = \bar{d}) = \mathcal{N}(a; \mu_{A,m,g,z,p_m^z}, \sigma_{A,m,\bar{d}}^2). \tag{1}$$

The mean parameter μ_{A,m,g,z,p_m^z} associated with this Gaussian is set to

$$\mu_{A,m,g,z,p_m^z} = \frac{n_g^{alt} p_m^z + n_N^{alt}(1 - p_m^z)}{n_g p_m^z + n_N(1 - p_m^z)}, \tag{2}$$

where n_g is the DNA copy number in tumor cells with genotype g, n_N is the DNA copy number in normal cells, n_g^{alt} is the copy number of the alternative allele in tumor cells with genotype g, and n_N^{alt} is the copy number of the alternative allele in normal cells (Supplementary Table 6). The mean parameter μ_{A,m,g,z,p_m^z} is not estimated from data, but determined by the states of the hidden variables G_t, Z_t, and $P_m^{Z_t}$ and the observed $\bar{d}_t$. The variance parameter $\sigma_{A,m,\bar{d}}^2$, however, is estimated from data.

- The log ratio of tumor-normal read depth at site t in biopsy m ($\forall m = 1, \ldots, M$), denoted by $L_{m,t}$, is modeled as a Gaussian variable, i.e.,

$$P(L_{m,t} = l | Z_t = z, G_t = g, P_m^z = p_m^z) = \mathcal{N}(l; \mu_{L,m,g,z,p_m^z}, \sigma_{L,m}^2). \tag{3}$$

The mean parameter μ_{L,m,g,z,p_m^z} is set to

$$\mu_{L,m,g,z,p_m^z} = \log_2 \frac{n_g p_m^z + n_N(1 - p_m^z)}{n_N} + c_m, \tag{4}$$

where n_N is copy number in normal cells (set to be 2 by default), and n_g is the DNA copy number in tumor cells with genotype g. The parameter c_m captures the sequencing depth difference in the tumor biopsy and the normal biopsy and the read number discrepancy due to ploidy change in the tumor biopsy. Therefore, the mean parameter μ_{L,m,g,z,p_m^z} is also not estimated from data, but determined by the states of the hidden variables G_t, Z_t, and $P_m^{Z_t}$. The variance parameter $\sigma_{L,m}^2$ again is estimated from data.

The remaining three observed variables provide information about the current genomic position, and they are discrete:

- The variable H_t is the distance (in base pairs) between site t and site $t - 1$. The effect of H_t is interesting since, while there is an equal number of graphical model frames between any two sites, the actual duration between them still effects the statistics of the underlying Markov chains since both G_t and Z_t directly depend on H_t. THEMIS, therefore, expresses a kind of irregularly spaced dynamic graphical model within the frame work of a regularly spaced dynamic graphical model. Moreover, THEMIS does this more efficiently than an alternative where the number of graphical model frames between sites t and $t - 1$ is proportional to $\bar{h}_t$, an approach that would be significantly more costly computationally. Supplementary Fig. 5 shows a histogram of $\log(\bar{h}_t)$ indicating that there is a diverse set of lengths between successive sites — the diversity suggests that inter-site length can have a significant influence on the Markov chains' transition matrices.
- The site type D_t at site t is a Boolean with $D_t = 0$ denoting a germline heterozygous site and $D_t = 1$ denoting a somatic mutation site which hosts a SNV event.
- The Boolean variable S_t denotes the start of a chromosome. When the current site is the start of a chromosome ($S_t = 1$), G_t and Z_t do not depend on G_{t-1} and Z_{t-1}, but follow uniform prior distributions.

In addition, each frame of the THEMIS model contains two hidden variables.

- The genotype G_t at site t is a discrete variable, which corresponds to all possible genotypes up to a certain number of copies. We consider all possible genotypes up to five copies (Supplementary Table 6).
- The clone index variable Z_t at site t is a discrete variable of $|Z|$ possible values (i.e., $Z_t \in \{1, \ldots, |Z|\}$), where $|Z|$ is pre-specified by the user.

Finally, the model contains a set of hidden variables that are "tied" across frames. For clone z in biopsy m ($\forall m = 1, \ldots, M$), the prevalence level variable P_m^z is a discrete variable of $|P|$ possible values, where $|P|$ is pre-specified by the user. The default $|P|$ is 20, corresponding to 20 equally spaced prevalence levels between 0 and 1 (i.e. 0.05, 0.10, …, and 1.00). THEMIS does not model clone prevalence as a continuous variable because clone prevalence is a parent of other variables (e.g. allelic ratio), and a hidden continuous variable cannot appear as a parent of other variables in a dynamic graphical model.

Parameters in the THEMIS model. Some parameters in the THEMIS model need to be specified by the user, whereas other parameters are estimated from data. Specifically, the user must specify the following parameters: the number of biopsies used in the analysis (M), the number of subclones ($|Z|$), the maximum copy number in the mutations (c_{max}^T), the number of prevalence levels ($|P|$), and the log-ratio offset in biopsy m due to ploidy and sequencing depth change ($c_m, \forall m = 1, \ldots, M$). In our experiments, we first run THEMIS with an initial estimate of c_m derived by examining the bivariate plot of allelic ratio and log ratio at germline heterozygous sites (Supplementary Fig. 6). Specifically, c_m is estimated by identifying the center of the normal genotype cluster on the log ratio axis. After running THEMIS, we re-estimate c_m as the average log-ratio on the sites whose genotypes are predicted to be "AB" (i.e. no CNA). This new estimate is used in a second run of THEMIS. In practice, the user can also leverage other ploidy estimation tools to get the initial estimate of c_m or run THEMIS multiple times with multiple initial estimates and choose with the one with highest likelihood. Three sets of parameters are estimated from data via the standard expectation-maximization (EM) algorithm for dynamic graphical models:

1. The variance of allelic ratio in biopsy m ($\forall m = 1, \ldots, M$) on germline heterozygous sites, denoted by $\sigma_{A,m,0}^2$, the variance of allelic ratio in biopsy m ($\forall m = 1, \ldots, M$) on somatic mutation sites, denoted by $\sigma_{A,m,1}^2$, and the variance of log ratio in biopsy m ($\forall m = 1, \ldots, M$) on each site, denoted by $\sigma_{L,m}^2$.

2. The transition probability from genotype j at site $t-1$ to genotype i at site t ($i, j \in \{1, \ldots, |G|\}$), denoted by $Q_G(i, j; \overline{h}_t)$ and transition probability from clone j at site $t-1$ to clone i at site t ($i, j \in \{1, \ldots, |Z|\}$), denoted by $Q_Z(i, j; \overline{h}_t)$. We model $Q_G(i, j; \overline{h}_t)$ and $Q_Z(i, j; \overline{h}_t)$ as parametric functions of $\overline{h}_t$, the distance (in base pairs) between site t and site $t-1$. We first define the probability of staying at the same genotype j, denoted by $\rho_G(j; \overline{h}_t)$, as

$$\rho_G(j; \overline{h}_t) = \frac{\mathcal{N}(\log \overline{h}_t; 0, \sigma_{G,j}^2)}{\mathcal{N}(0; 0, \sigma_{G,j}^2)} \frac{|G| - 1}{|G|} + \frac{1}{|G|},$$

(5)

where $\mathcal{N}(x; 0, \sigma_{G,j}^2)$ is the probability density of a Gaussian distribution with mean 0 and variance $\sigma_{G,j}^2$ at the point x. Then

$$Q_G(i, j; \overline{h}_t) = \begin{cases} \rho_G(j; \overline{h}_t), & \text{if } i = j \\ \dfrac{1 - \rho_G(j; \overline{h}_t)}{|G| - 1}, & \text{otherwise.} \end{cases}$$

(6)

Similarly, we define

$$\rho_Z(j; \overline{h}_t) = \frac{\mathcal{N}(\log \overline{h}_t; 0, \sigma_{Z,j}^2)}{\mathcal{N}(0; 0, \sigma_{Z,j}^2)} \frac{|Z| - 1}{|Z|} + \frac{1}{|Z|},$$

(7)

and

$$Q_Z(i, j; \overline{h}_t) = \begin{cases} \rho_Z(j; \overline{h}_t), & \text{if } i = j \\ \dfrac{1 - \rho_Z(j; \overline{h}_t)}{|Z| - 1}, & \text{otherwise}. \end{cases}$$

(8)

Note that we estimate $\sigma_{G,j}^2$ ($\forall j = 1, \ldots, |G|$) and $\sigma_{Z,j}^2$ ($\forall j = 1, \ldots, |Z|$) from data in a maximum likelihood fashion. The way we parameterize the transition probabilities captures the phenomenon that the probability of staying at the same genotype decreases as a monotone function of the distance from the previous site, and the rate it decreases is parametrized by $\sigma_{G,j}^2$. Therefore, we estimate the decreasing speed (i.e. $\sigma_{G,j}^2$) adaptively from the data (unlike user-prespecified transition probabilities in TITAN), and the speed is different for different genotypes since different mutation events may occur with different lengths on the genome (unlike the tied transition probabilities in TITAN).

3. The prior distributions of genotypes, clones and the cell prevalence levels of the clones (denoted by π_G, π_Z and π_P, respectively). These prior distributions are responsible for the frames that correspond to the starts of the chromosomes. These prior distributions are initialized as uniform distributions, and trained from the data along with other parameters.

We use Ω to denote the set of parameters in the joint distribution specified by the model, namely $\Omega = \{\sigma_{A,m,0}^2, \sigma_{A,m,1}^2, \sigma_{L,m}^2, \sigma_{G,j}^2, \sigma_{Z,j}^2, \pi_G, \pi_Z, \pi_P, \forall m = 1, \ldots, M, \forall j = 1, \ldots, |Z|\}$. In the estimation step, we use the EM algorithm to estimate the parameters and find a (local) maximum, denoted by $\hat{\Omega}$, for

$$\underset{\Omega}{\mathrm{argmax}} \Pr(A_{1:M,1:T} = \overline{a}_{1:M,1:T}, L_{1:M,1:T} = \overline{\ell}_{1:M,1:T}, D_{1:T} = \overline{d}_{1:T}, S_{1:T} = \overline{S}_{1:T}, H_{2:T} = \overline{h}_{2:T}; \Omega).$$

(9)

In the inference step, we infer the most probable states of the hidden variables given the estimated parameters $\hat{\Omega}$, i.e.,

$$\operatorname*{argmax}_{g_{1:T}, z_{1:T}, p_{1:T}^z} \Pr(A_{1:M,1:T} = \overline{a}_{1:M,1:T}, L_{1:M,1:T} = \overline{l}_{1:M,1:T}, D_{1:T} = \overline{d}_{1:T}, S_{1:T} = \overline{s}_{1:T},$$

$$H_{2:T} = \overline{h}_{2:T}, G_{1:T} = g_{1:T}, Z_{1:T} = z_{1:T}, P_{1:M,1:T}^z = p_{1:M,1:T}^z; \hat{\Omega}). \tag{10}$$

The inferred most probable sequence of hidden variables, as the output of our algorithm, provide the heterogeneity analysis results, i.e., a number of subclones and their genotypes and cell prevalences (Fig. 1b).

Selecting the number of subclones. THEMIS requires the user to specify the number of subclones in the biopsies before running the model. There are three ways of identifying the number of subclones from the data. The first method is a naive visualization method. If the biopsy is well sequenced, then the number of subclones can be directly identified from the bivariate plot (allelic ratio against log ratio at germline heterozygous sites) of the biopsy by observing different prevalence levels of the LOH events. We take one tumor biopsy (biopsy B1 in ITOMIC study) as an example, whose bivariate plot is provided in Supplementary Fig. 6. It can be observed that there are two major LOH prevalence levels in the plot. Therefore, we can assume that there are two tumor subclones in the biopsy.

Another way of choosing the number of subclones is to use the Bayesian information criterion (BIC)[16]. BIC is defined as

$$BIC = -2 \ln L + k \ln(n), \tag{11}$$

where $\ln L$ is the log likelihood of the data, k is the degree of freedom, and n is the number of data points. We choose the subclone number which produces the smallest BIC score. When we run the 2-subclone model and the 3-subclone model on the tumor biopsy B1, BIC scores are $-476,725$ and $-464,710$, respectively. Therefore, we can assume there are two tumor subclones in the biopsy based on the BIC scores.

A third way of choosing the number of subclones is to use cross-validation. Suppose that we use three-fold cross-validation. We randomly partition the chromosomes into three sets. In each training-testing split, we use the data from two sets as the training data and the remaining set as the testing data. With the estimated parameters from training data, we run Viterbi algorithm on the testing data, and choose the subclone number which produces the largest averaged log-likelihood (a.k.a., the Viterbi score in GMTK) on testing data. In both simulated data (in Simulations 1 and 2, we simulated two tumor subclones and three tumor subclones, respectively) and real data (biopsy B1 in the ITOMIC study), we observed that the three methods provide the correct results (Supplementary Table 7).

In practice, one may use any of the three methods or a combination of the three methods to set the number of subclones. Note that although the naive visualization method is straightforward, it may produce inaccurate estimates if the sequencing depth of the biopsy is low or when the prevalence levels of two subclones are close to each other. Cross-validation is more robust compared with BIC, but requires additional computational cost. When BIC and cross-validation are being used, we recommend starting with a small number of subclones (e.g. 2) and increase the number until the evaluation criteria deteriorate. For example, if a 2-subclone model produces a lower BIC (or higher averaged $\ln L$ in cross-validation) than a 3-subclone model, it is not necessary to run the 4-subclone model.

Joint analysis of multiple biopsies from the same patient. The joint analysis of multiple biopsies in THEMIS is done by first enumerating candidate phylogenetic trees, encoding each tree in the conditional probability tables associated with variables $A_{m,t}$ and $L_{m,t}$, and then selecting the tree whose associated model yields the highest likelihood. During the enumeration phase, we make three assumptions. First, we assume that we have the statistical power to discern all the clones from individual biopsies and estimate their prevalences. Second, we assume that we can identify shared clones between biopsies by computing and thresholding similarities between the clones. Third, if the sum of the prevalences (p_a and p_b) of clones a and b is greater than 1.0 in at least one biopsy, and $p_a > p_b$ in all biopsies where clones a and b are present, then we assume that clone a is an ancestor of clone b. The first two assumptions ensure that the ground truth structure is contained in the candidate structures. The third assumption helps us reduce the number of candidate structures. Users are also encouraged to use other information, such as the time and physical locations of the biopsies, to eliminate candidate structures. Joint analysis over multiple biopsies can be carried out in the following five steps.

1. Analyze biopsies separately with THEMIS and identify the genotype and prevalence of each clone within each biopsy.
2. Compute similarities between all pairs of clones from different biopsies and merge similar clones.
3. Identify consistent parent-child relationships based on the individually estimated prevalences using the third assumption above.
4. Enumerate all phylogenies consistent with those relationships and run THEMIS accordingly.
5. Select the phylogeny with maximum likelihood.

References

1. Shah, S. P. *et al.* The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399 (2012).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
3. Gerlinger, M. *et al.* Cancer: Evolution within a lifetime. *Annual Review of Genetics* **48**, 215–236 (2014).

4. Frampton, G. M. *et al*. Development and validation of a clinical cancer genomic profiling test based on massively parallel dna sequencing. *Nature Biotechnology* **31**, 1023–1031 (2013).
5. Deshwar, A. G. *et al*. Phylowgs: Reconstructing subclonal composition and evolution from whole genome sequencing of tumors. *Genome Biology* **16** (2015).
6. El-Kebir, M., Satas, G., Oesper, L. & Raphael, B. J. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems* **3**, 43–53 (2016).
7. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences* **113**, E5528–E5537 (2016).
8. Bilmes, J. & Zweig, G. The Graphical Models Toolkit: An open source software system for speech and time-series processing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (2002).
9. Ha, G. *et al*. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research* **24**, 1881–1893 (2014).
10. Blau, C. A. *et al*. A distributed network for intensive longitudinal monitoring in metastatic triple-negative breast cancer. *Journal of the National Comprehensive Cancer Network* **14**, 8–17 (2016).
11. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols* **4**, 44–57 (2009).
12. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1–13 (2009).
13. Cibulskis, K. *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31**, 213–219 (2013).
14. Benjamini, Y. & Speed, T. P. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Research* **40**, gks001 (2012).
15. Ha, G. *et al*. Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research* **22**, 1995–2007 (2012).
16. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).
17. Yau, C. *et al*. A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biology* **11**, R92 (2010).
18. Letouzé, E., Allory, Y., Bollet, M. A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology* **11**, 1–19 (2010).
19. Greenman, C. D. *et al*. Estimation of rearrangement phylogeny for cancer genomes. *Genome Research* **22**, 346–361 (2012).
20. Carter, S. L. *et al*. Absolute quantification of somatic dna alterations in human cancer. *Nature Biotechnology* **30**, 413–421 (2012).
21. Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research* **41**, e165 (2013).
22. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology* **14**, R80–R80 (2013).
23. Oesper, L., Satas, G. & Raphael, B. J. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics* **30**, 3532–3540 (2014).
24. Purdom, E. *et al*. Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29**, 3113–3120 (2013).
25. Yau, C. OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics* **29**, 2482–2484 (2013).
26. Roth, A. *et al*. Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods* (2014).
27. Miller, C. A. *et al*. Sciclone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology* **10**, e1003665 (2014).
28. Zare, H. *et al*. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Computational Biology* **10**, e1003703 (2014).
29. Fischer, A., Vázquez-Garca, I., Illingworth, C. J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Reports* (2014).
30. Schwarz, R. F. *et al*. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Computational Biology* **10**, e1003535 (2014).
31. Qiao, Y. *et al*. Subcloneseeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biology* **15**, 443 (2014).
32. Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* **30**, i78–i86 (2014).
33. Fan, X., Zhou, W., Chong, Z., Nakhleh, L. & Chen, K. Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinformatics* **15**, 299 (2014).
34. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
35. Sengupta, S. *et al*. Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 20, 467 (World Scientific, 2015).
36. Malikic, S., McPherson, A. W., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349–1356 (2015).
37. Popic, V. *et al*. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology* **16**, 1 (2015).
38. El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **31**, i62–i70 (2015).

Acknowledgements

Author Contributions

J.L., J.A.B. and W.S.N. conceived the method. J.L. and J.T.H. performed the analysis. J.L., W.S.N., and C.A.B. wrote the manuscript. E.M.M., C.S., M.O.D., and C.A.B. contributed experimental results. S.B., V.K.G., and C.A.B. contributed to the clinical trial. R.M.D., C.L., E.M.M., D.P., P.R., and J.S. acquired single cell sequencing data.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16813-4.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.