Wang et al. Genome Biology (2021) 22:70
https://doi.org/10.1186/s13059-021-02291-5

Genome Biology

RESEARCH Open Access

MEDALT: single-cell copy number lineage tracing enabling gene discovery

Fang Wang[1,2†], Qihan Wang[1,3†], Vakul Mohanty[1], Shaoheng Liang[1], Jinzhuang Dou[1], Jincheng Han[4], Darlan Conterno Minussi[5], Ruli Gao[6], Li Ding[7], Nicholas Navin[5] and Ken Chen[1*]

* Correspondence: kchen3@mdanderson.org
†Fang Wang and Qihan Wang contributed equally to this work.
[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Houston, TX, USA
Full list of author information is available at the end of the article

Abstract

We present a Minimal Event Distance Aneuploidy Lineage Tree (MEDALT) algorithm that infers the evolution history of a cell population based on single-cell copy number (SCCN) profiles, and a statistical routine named lineage speciation analysis (LSA), whichty facilitates discovery of fitness-associated alterations and genes from SCCN lineage trees. MEDALT appears more accurate than phylogenetics approaches in reconstructing copy number lineage. From data from 20 triple-negative breast cancer patients, our approaches effectively prioritize genes that are essential for breast cancer cell fitness and predict patient survival, including those implicating convergent evolution.

The source code of our study is available at https://github.com/KChen-lab/MEDALT.

Keywords: Single-cell, scDNA-seq, scRNA-seq, Copy number alteration, Tumor evolution, Lineage tracing, Driver discovery

Background

Aneuploidy, the phenomenon that genomes acquire or lose chromosomal fragments, has been causally implicated in a wide variety of human diseases such as neuropsychiatric disorders and cancer [1–3]. Genetic and phenotypic plasticity resulting from aneuploidy evolution causes treatment resistances and disease recurrences [4–6], which fundamentally challenges current medicine. Recent studies have shown that not only disease tissues, but also pathologically normal tissues may contain a high degree of somatic mosaicisms (e.g., peripheral blood [7] and esophagus [8]). Therefore, defining which copy number alterations (CNAs) cause pathogenesis and which are part of normal variations becomes increasingly important in genome medicine, especially for cancer [9, 10].

Various efforts have been made to obtain comprehensive knowledge of CNAs responsible for cancer diagnostics, prognostics, and targeted therapeutics. Systematic CNA analysis in over 10,000 primary tumor samples in the cancer genome atlas (TCGA) and 2500 samples in the International Cancer Genome Consortium (ICGC) revealed distinct CNA landscapes in different cancer types [11–13]. Comparison of CNAs among autologous tumors obtained at different stage from different histology revealed that CNAs are critical

for tumor evolution across time and space. However, studies based on bulk tissue samples cannot fully depict the history of tumor evolution, which occurs in single-cell resolution [14], and thus have limited power to discover the associated genetic drivers.

Recent advances in single-cell DNA sequencing (e.g., tagmentation-based approach [15] and single-cell CNV solution by the 10x Genomics) have enabled large-scale acquisition of single-cell copy number (SCCN) profiles in tens of thousands of cells at around 100-kb resolution (~ 0.1X sequencing coverage per cell) [16–19]. Other platforms such as single-cell RNA-sequencing [20, 21] and single-cell ATAC-sequencing [22] have also been utilized for SCCN profiling. A set of bioinformatic tools have been developed to call SCCN profiles, taking into consideration various confounding factors [23–25].

These SCCN profiles not only present a rich pool of genetic perturbations that are invisible at tissue level, but also potentiate reconstruction of cellular lineage, based on which the impact of an allele on cellular fitness can be measured. Thus, statistical approaches that integrate cellular lineage tracing with population genetic analysis [26] can enable discovery of novel disease genes and mechanisms of disease progression.

So far, studies performing retrospective lineage tracing from single-cell data have largely been utilizing phylogenetics approaches designed to model species evolution, which is quite different from cellular evolution in terms of duration, scale, genetics, and dynamics [27, 28]. Many existing phylogenetics approaches assume that genomic sites evolve independently and follow the so-called infinite site assumption (ISA) [29]. But in the context of aneuploidy, a genome site can often be altered repeatedly by different CNAs, due partly to constraints on genome and chromatin structures, properties of DNA replication/repairing [30], and functional selection. To apply conventional maximum parsimony approaches on SCCN data, one has to over-segment genomic regions and represent copy numbers as characters in disjoint intervals, which ill-represents the properties of DNAs and distorts evolution propensity across copy number states. Other conventional methods using Euclidean, Hamming, or correlational distances also ill-represent the segmental, non-linear nature of CNA evolution [31], leading to inaccurate inference of tree topology and branch lengths.

A few new phylogenetics approaches have been developed to tackle these limitations by introducing a new distance metric called Minimal Event Distance (MED), which postulates the minimal number and the series of single-copy gains or losses that are required to evolve one genome to another. Particularly, the MEDICC [32] algorithm infers a copy number phylogenetic tree from the allelic copy number profiles of a set of samples. However, the problem is NP-hard [33]. Even the simplified solutions could be applied to only tens of genomes and are not scalable to current single-cell datasets consisting of thousands of cells. Zeira et al. [34] proposed a linear-time solution to the problem based on an integer linear programming (ILP) formulation, but no tool was released.

Having the new distance and efficient tree inference algorithms was a good step forward, but it remains unclear how to identify functional variants, given a cell phylogenetic tree. Intuitively, functional variants affecting cellular fitness should lead to altered variant allele frequencies in the descendant populations, as implicated by previous multiregional tumor phylogenetics studies [10, 35]. However, mathematical procedures [28, 31] have not been developed to quantify the impact of a genomic alteration over a phylogenetic

tree, taking into account sparsity in cell population sampling, multiplicity in subset partitioning, and propensity of the alteration at a particular genomic location, etc. [36]

Results

Overview of the methods

To address these challenges, we propose a new computational framework that performs lineage tracing from SCCN data and detects significant focal (gene resolution) and broad (chromosomal-arm resolution) CNAs associated with lineage expansion (Fig. 1).

The SCCN profiles are represented as an integer-valued matrix using previously pub-

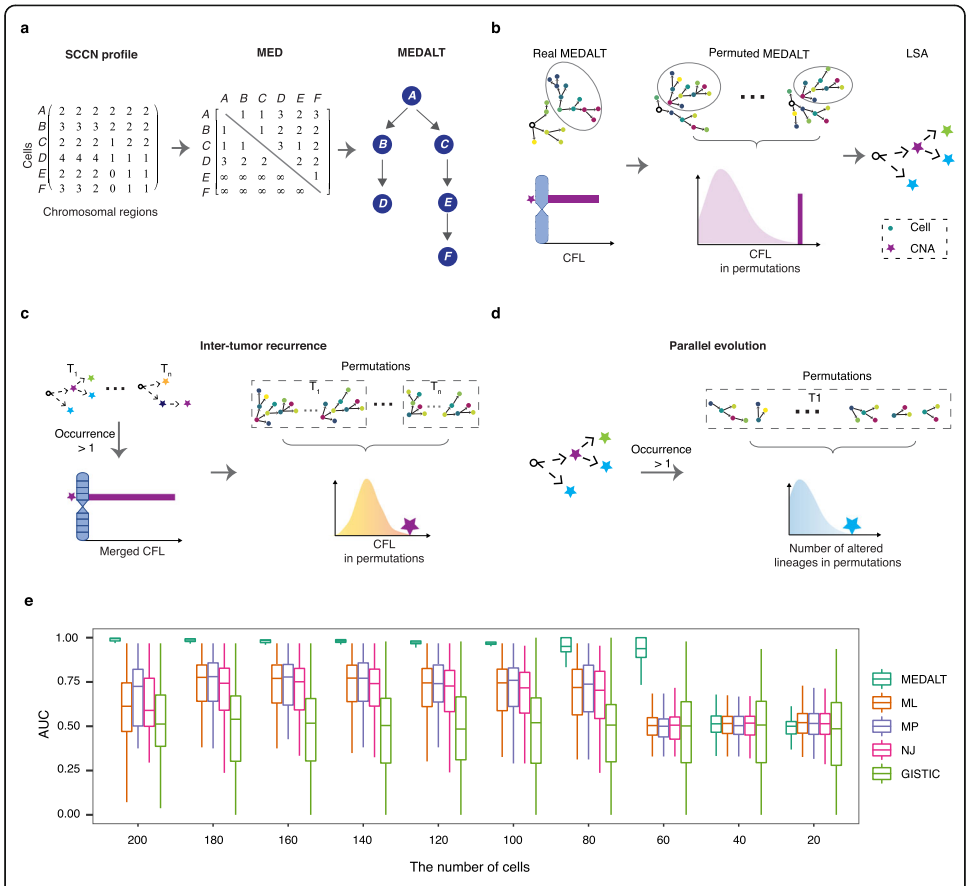

Fig. 1 Algorithm flowchart and evaluation. **a** Algorithm for constructing a MEDALT. **b** Identification of non-random fitness-associated CNAs in an individual sample. **c** Identifying non-random fitness-associated CNAs in a cohort of samples. **d** Identifying parallel/convergent evolution CNAs in an individual sample. **e** AUC of FAA identification based on 100 synthetic datasets without noise. MP, maximal parsimony tree. NJ, neighbor-joining tree. ML, maximum likelihood tree

lished approaches [16, 18], in which each row represents a cell and each column a chromosomal region. We then deduce the minimal number and the series of single-copy gains or losses (i.e., minimal event distance) that are required to evolve the genome of one cell to the next (Additional file 1: Fig. S1a) using an efficient greedy algorithm which is similar and has the same asymptotic bound as Zeira et al. [34] (see the "Methods" section and Additional file 1: Table S1).

We then infer a directed minimal spanning tree, named Minimal Event Distance Aneuploidy Lineage Tree (MEDALT, Fig. 1a), using an adapted version of the Edmond's algorithm that scales polynomially with respect to the number of cells (see the "Methods" section and Additional file 1: Table S2). In a MEDALT, each node represents a cell, each edge represents a kinship between two cells, arrows point towards younger cells, and the root represents a normal diploid cell.

MEDALT allows a genomic region to be repetitively altered by multiple single-copy gains or losses. It provides a parsimonious interpretation, the minimal number of single-copy gains or losses that may have led to the evolution of the entire cell population.

An important constraint is that chromosomal fragments cannot be recovered if completely lost. To reflect that property, the MEDs originating from cells containing homozygous copy number loss are set to infinity.

Since MEDALT describes copy number evolution by segments instead of sites, we expect that it will enable more accurate cellular lineage tracing than do conventional phylogenetics methods (Additional file 1: Fig. S1b; see the "Methods" section).

We further establish a statistical routine, named Lineage Speciation Analysis (LSA), to prioritize CNAs and genes that are non-randomly associated with lineage expansion and thereby have potential functional impact.

To perform LSA, we first iteratively partition cells into lineages (subsets) based on the topology of the lineage tree. For each CNA region in each candidate lineage, we calculate a cumulative fold level (CFL) as the summation of the copy number levels in constituent cells (Fig. 1b and Additional file 1: Fig. S1c). We then assess the statistical significance of the observed CFL with respect to a background distribution established from random lineages of similar sizes (the same or the closest size) obtained from a permutation process (see the "Methods" section; Fig. 1b and Additional file 1: Fig. S1d). The permutation process randomly assigns SCCN profiles by chromosomes into different cells 1000 times and reconstruct a lineage tree from each permutated dataset using the same lineage tracing algorithm. For each lineage from the real data, at least 1000 lineages of similar size (the same or the next closest size) from the permutated trees are selected, since multiple lineages of similar size may exist in each permutated tree. It is important to account for background variations induced by factors unrelated to cellular fitness such as high CNA prevalence at fragile sites or repeats that are non-functional, as shown in previous studies [36, 37] and to account for bias of lineage tracing algorithms. We used three additional statistical approaches as controls, which estimate background distributions without reconstructing trees from permutated SCCN data (see the "Methods" section). LSA clearly outperformed other approaches for identifying CNAs that are non-randomly associated with lineage expansion (Additional file 1: Fig. S1e). The efficiency of the MEDALT algorithm, which is linear with respect to the number of cells and genome size (Additional file 1: Fig. S2), makes it possible to perform a large number of permutations in order to obtain a reasonably accurate background distribution. The statistically significant CNAs and genes so identified may not be causal themselves, but are associated with (e.g., co-occur) with causal fitness-impacting alterations. Thus, LSA distills the massive genome-wide SCCN data into a compact molecular blueprint, consisting of CNAs/genes occurring non-randomly at important moment during the course of the evolution with significant impact on the fitness of the descendant cells.

LSA can also be applied at cohort level to analyze single-cell data obtained from multiple patient samples. In that setting, the method creates meta-lineages combining cells from different patients and prioritizes events non-randomly occurring across background lineages established over the entire cohort (Fig. 1c, Additional file 1: Fig. S1f and see the "Methods" section). Genes that are altered nonrandomly in multiple patients will likely have higher scores than those altered in a single patient.

Additionally, LSA can be applied to prioritize CNAs associated with parallel/convergent evolution [38] (abbr. PLSA) by estimating the chance of a CNA occurring nonrandomly in two or more parallel lineages, as a consequence of positive selection (Fig. 1d, Additional file 1: Fig. S1g and see the "Methods" section). This opens a new way for gene discovery that was substantially underpowered in bulk sample studies.

In silico evaluation

To evaluate our approaches, we simulated copy number evolution in single cells using a Markov process parameterized by cell fitness parameters (Additional file 1: Fig. S3a and b; see the "Methods" section) [39]. Spiked in randomly were fitness-associated alterations (FAAs), which indicate fitness change in a cell triggering subsequent lineage expansion. Synthetic SCCN profiles were created mimicking various CNA mechanisms such as genome doubling, breakage-fusion-bridge (BFB), tandem duplication, terminal deletion, and unbalanced translocation [30]. We created 100 simulated datasets, each containing around 200 cells. Besides obtaining MEDALTs, we also obtained phylogenetic trees using conventional maximum likelihood (ML), maximum parsimony (MP), and neighbor joining (NJ) approaches (see the "Methods" section). In addition, we ran GISTIC [37] (see the "Methods" section), a method developed to prioritize CNAs in tissue samples by treating the cells as unrelated samples.

We then performed FAA detection in each dataset by performing LSA on individual trees inferred by various methods. We compared the detection performances using area under receiver operating characteristic curves (AUC; see the "Methods" section). Overall, the MEDALT approach achieved substantially better detection performance than the other methods (Fig. 1e). The benefits appeared robust over a range of cell numbers, when we repeated the benchmarking on subsets of the cells via random downsampling, until the number of cells dropped below 60. It appeared that at least 30% of the cells were required to recapitulate the major population structure in this simulation irrespective of the algorithms (Fig. 1e).

We further dissected the contribution of each of the 3 steps in our approach, i.e., MED, MEDALT, and LSA, to the final performance of FAA detection. Compared to MED, the MEDALT and LSA steps had more contribution to the final performance (Additional file 1: Fig. S3c). Therefore, although MED can be affected by noise in the SCCN data, the net effects appeared limited (Additional file 1: Fig. S3d).

Detecting fitness-associated CNAs in disease cohorts

We applied our methods on the single-cell DNA-sequencing data acquired from 20 triple-negative breast cancer patients (TNBCs, Additional file 1: Table S3) [16, 18]. SCCN profiles were downloaded from the original paper which were generated using a variable binning method followed by circular binary segmentation (CBS) [40]

(Additional file 1: Fig. S4; see the "Methods" section). We obtained both MEDALTs and phylogenetic trees for each sample and ran LSA to identify non-random alterations at both sample and cohort levels.

We then compared the accuracy of the trees in inferring cellular timing using data from 4 patients with longitudinal pre-, mid-, and post-treatment (neoadjuvant chemotherapy) samples. We found that MEDALTs ordered cells much more consistently with their biopsy timing than did the phylogenetic trees (Additional file 1: Fig. S5), with pre-treatment cells appearing near the root and post-treatment cells near the leaves.

Consistent with previously studies [16, 18], most of the TNBC samples appeared to have developed through branched evolution via multiple parallel lineages. Interestingly, the MEDALTs indicated that these parallel lineages may have distinct mutation rates (Fig. 2a and b, Additional file 1: Fig. S6), which may be attributable to variable degree of DNA damage repair (DDR) loss (Fig. 2b; see the "Methods" section) [41]. Indeed, when we performed gene set enrichment analysis on genes identified by LSA, we found that the lineages of higher CNA rates have more DDR genes affected by the CNAs than the lineages of lower CNA rates (Additional file 1: Fig. S7).

We identified fitness-associated CNAs at chromosomal and gene resolution using cohort-level LSA (p value < 0.001; see the "Methods" section). For benchmarking, we also performed the same LSA on the MP trees. We also ran GISTIC [37] on the pseudo-bulk copy number profiles generated by averaging the SCCN profiles across the cells in each sample (see the "Methods" section).

Overall, the MEDALT plus LSA approach identified 30 broad CNAs, 80% of which have been functionally associated with breast cancer development and treatment outcome in the literature (Additional file 1: Table S4). The accuracy was at least 13% higher than the results derived by the other methods (Fig. 2c; see the "Methods" section). We independently performed the LSA at gene resolution, focusing on 448 genes from 11 oncogenic pathways including Notch, PI3K, Hippo, RTK/RAS, MYC, cell cycle, p53, Nrf2, Wnt, TGFB, and DDR defined in TCGA Pan-can atlas research [41, 42]. Our approach identified 197 genes, including 109 amplified and 88 deleted genes (Additional file 2). In contrast, the MP plus LSA approach identified 130 genes, 82 of which were amplified and 48 deleted. GISTIC identified 60 genes, 33 of which were amplified and 27 deleted.

By examining the CRISPR knockout screen data in 29 breast cancer cell lines in the DepMap database [43], we found that the 109 amplified genes identified by the MEDA LT plus LSA approach had significantly lower gene knockoff effect scores than those of the 82 amplified genes detected based on the MP trees (one-side Wilcoxon rank-sum test, $p = 2.75 \times 10^{-9}$) and of the 33 genes detected by GISTIC (one-side Wilcoxon rank-sum test, $p = 6.65 \times 10^{-17}$) (Fig. 2d). The scores were also significantly lower than those of oncogenes (one-side Wilcoxon rank-sum test, $p = 1.12 \times 10^{-15}$) and tumor suppressors (one-side Wilcoxon rank-sum test, $p = 2.81 \times 10^{-16}$) reported in the oncoKB [44] and intOGen [45] databases, which are not specific to TNBC, and sets of randomly selected genes of identical size (one-side Wilcoxon rank-sum test, $p = 8.97 \times 10^{-21}$). Not significant were the scoring differences among the sets of deleted genes, due likely to challenges in calling deletions from noisy low-coverage data and in quantifying deleterious effects in lineages of limited cell numbers.

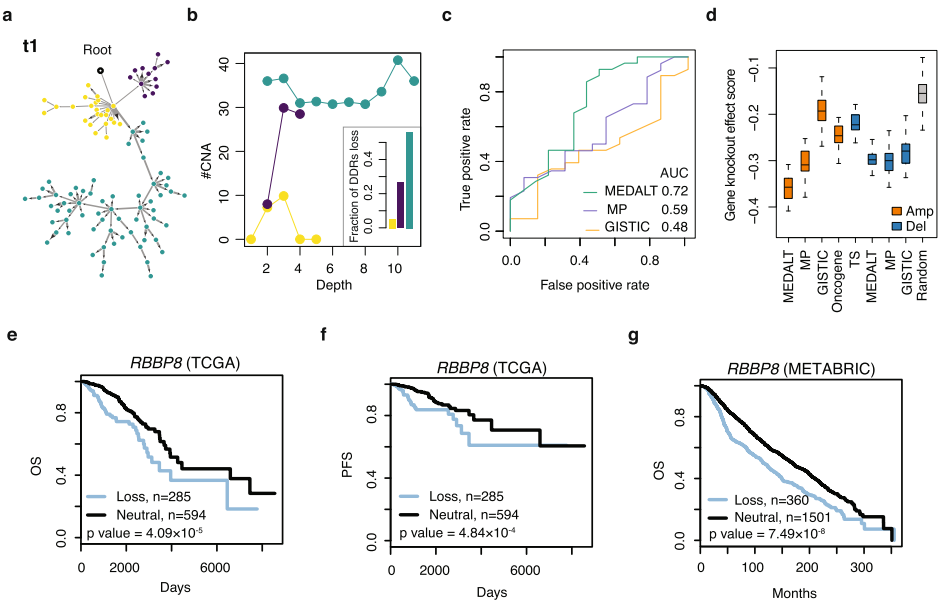

Fig. 2 Application on scDNA-seq data from TNBCs. **a** The MEDALT Inferred from patient t1. The widths of the edges are drawn proportional to the MEDs. Colors (yellow, blue and green) highlight the 3 main branches. **b** The relationship between CNA numbers (*Y*-axis) and the depth on the tree (or distance to the root, *X*-axis). The barplot shows the fraction of DDR genes among the genes with copy number losses in the 3 lineages. **c** ROC curve for identifying functionally important, broad CNAs in the literature (Table S4). **d** The gene knockout effect scores of the gene sets (cohort LSA test *p* value < 0.001) identified based on the MEDALTs, MP trees, and GISTIC in 29 breast cancer cell lines. Included as controls are 100 sets of 197 randomly selected genes, 234 oncogenes, and 269 tumor suppressors (TS) in oncoKB and intOGen. The overall survival (OS) (**e**) and the progression-free survival (PFS) (**f**) of *RBBP8* loss in TCGA breast cancer patients. **g** OS of *RBBP8* loss in the METABRIC

Among the 197 genes MEDALT nominated, some are not reported in the oncoKB [44], COSMIC [46], and intOGen [45] databases (Additional file 3) but supported by functional genomics data in large-scale cancer patient studies (Additional file 1: Fig. S8a). For example, loss of *RBBP8* indicated worse prognosis among the breast cancer patients in TCGA and those in the METABRIC [47] (Fig. 2e to g). *RBBP8* is a potentially interesting target as it interacts with *BRCA1* and modulates its function in transcriptional regulation, DNA repair, and/or cell cycle checkpoint control [48]. In addition, loss of *PPP4R1* indicated worse prognosis in TCGA and the METABRIC as well (Additional file 1: Fig. S8b to d).

In addition, we identified 107 genes that were likely positively selected (PLSA *p* value < 0.001, Additional file 4) by convergent evolution in 7 of the 20 patients (Fig. 3a), by performing PLSA on the MEDALTs derived from individual patients. Among these, 65 genes were amplified. By repeating the same PLSA on the MP trees, we identified 355 genes, 252 of which were amplified. The set of 65 genes identified from the MEDALTs had significantly lower gene knockout effect scores (thus more essential) than those of the set of 252 genes identified from the MP trees (one-side Wilcoxon rank-sum test, p value = 4.07×10^{-9}), of known oncogenes (one-side Wilcoxon rank-sum test, $p = 2.81 \times 10^{-16}$) and sets of randomly selected genes (one-side Wilcoxon rank-sum test, $p = 9.01 \times 10^{-21}$), based on the CRISPR screens of the 29 breast cancer cell lines in the DepMap [43] (Fig. 3b). No significant scoring differences were found between the deleted genes identified from the MEDALTs and those identified from the MP trees,

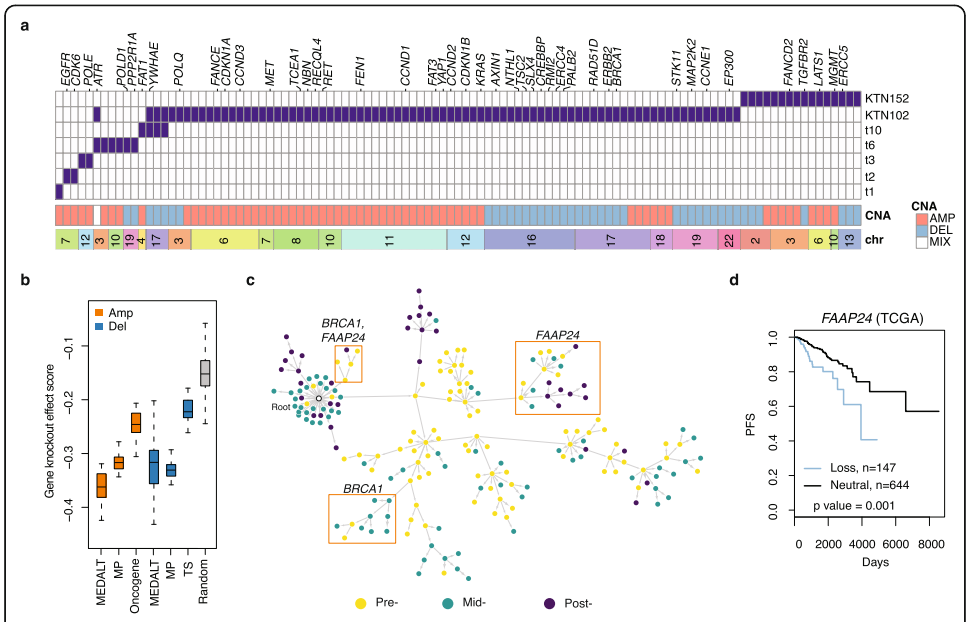

Fig. 3 Convergent evolution in TNBCs. **a** Genes associated with convergent evolution in the 20 TNBC patients. Labeled at the top are known cancer genes. AMP, copy number amplification; DEL, copy number deletions; MIX, genes with amplifications and deletions. **b** The gene knockoff effect scores of gene set identified from the MEDALTs and the MP trees by PLSA (p value < 0.001). Included as controls are 100 sets of 107 randomly selected genes, 234 oncogenes and 269 tumor suppressors (TS) in oncoKB and intOGen. **c** The MEDALT of patient KTN102. Orange boxes highlight lineages under potential convergent evolution. **d** Progression-free survival of breast cancer patients with and without *FAAP24* loss in TCGA

although both sets appeared more essential than the sets of known tumor suppressors and randomly selected genes.

Among the 107 genes identified by PLSA, 42% were known cancer genes, a fraction higher than what we obtained from the cohort-level single-lineage LSA (38%, Additional file 1: Fig. S8e). Loss of *FAAP24* appeared in two distinct lineages in patient KTN102 and was associated with worse progression-free survival (PFS) in TCGA breast cancer data (Fig. 3c and d). Loss of *BRCA1* was also found in two parallel lineages, which were depleted of cells from the post-treatment sample (Fig. 3c). That observation may be explained by the fact that BRCAness tumors often respond to neoadjuvant chemotherapy [49, 50].

Applications on single-cell RNA sequencing data

Our approaches are likely beneficial to characterizing SCCN data derived from single-cell RNA sequencing (scRNA-seq) experiments. To examine that possibility, we collected data obtained from paired primary and metastasis (or relapse) samples of a variety of cancer patients, including 6 head and neck squamous cell carcinoma (HNSCC) [20], 8 multiple myeloma (MM), 2 oral squamous cell carcinomas (OSCC) [51], and 4 ovarian cancer patients (OV) [52] (Additional file 1: Table S5).

We obtained SCCN profiles from the scRNA-seq data using the *inferCNV* program [53], which derives CNAs by exploring expression intensity of genes across position of tumor genome in comparison to a set of normal cells. We calculated average copy number levels in non-overlapping genomic 30-gene windows to infer MEDALT (see

the "Methods" section). We then obtained a MEDALT for each patient, including cells in both the primary and the metastasis samples. For comparison, we also performed analysis for each patient using Monocle v3.0 [54], which was designed to reconstruct the transcriptomic (and phenotypic) trajectory of a developing cell population. Since the cells in the primary samples were most likely born before the cells in the metastasis (or relapse) samples, they should be arranged closer to the root of the lineage trees. Indeed, in the MEDALTs, the cells from the primary samples were placed significantly (one-side Wilcoxon rank-sum test, $p = 0.0098$) closer to the root than the cells from the metastasis (or relapse) samples (Fig. 4a). In contrast, the pseudotime estimated by Monocle did not significantly (one-side Wilcoxon rank-sum test, $p = 0.51$) delineate the two types of cells (Fig. 4b). Meanwhile, cells in the MEDALT lineages had more homogenous SCCN profiles than those in the Monocle clusters (Fig. 4c and Additional file 1: Fig. S9; see the "Methods" section). The result from this experiment indicated that our approaches are potentially more accurate in characterizing genome evolution from cancer scRNA profiles than approaches that are designed for transcriptomic trajectory reconstruction. This may not be entirely surprising as DNA copy number data have demonstrated useful for cancer cell chronology inference [12] while RNA data are known subject to complex transcriptional regulation.

We performed cohort-level LSA for gene set estimated from inferCNV on the MEDALTs and identified 75 fitness-associated genes (Additional file 5, p value < 0.001), which included 45 amplified and 30 deleted genes from the 20 patients. In contrast, Monocle identified 3412 differentially expressed genes between the cell clusters.

We found that the amplified genes identified by our approach are significantly more essential than those identified by Monocle (one-side Wilcoxon rank-sum test, $p = 2.35 \times 10^{-186}$; Fig. 4d), based on the CRISPR screens of 524 cancer cell lines in the DepMap [43].

Discussions

Advances in single-cell technologies present new challenges and opportunities for making biological discovery. Single-cell studies often involve large numbers of cells, which

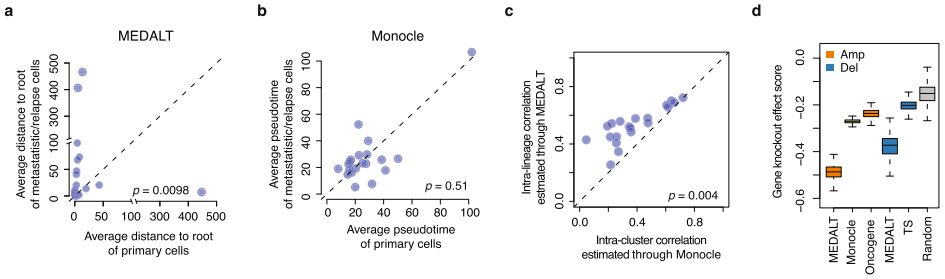

Fig. 4 Application on scRNA-seq data from cancer patients. **a** Average distance to root of the cells in the primary samples (*X*-axis) and those in the metastasis/relapse samples (*Y*-axis) estimated from the MEDALTs. **b** Average pseudotime of the cells in the primary samples (*X*-axis) and those in the paired metastasis/relapse samples (*Y*-axis) estimated by Monocle. **c** Pearson's correlation coefficients between the SCCN profiles of the cells in the same lineages dissected from the MEDALTs (*Y*-axis) versus those in the same cell states clustered by Monocle (*X*-axis). Each dot in **a**, **b**, and **c** represents a cancer patient. All the *p* values were estimated by one-side Wilcoxon rank-sum test. **d** Average DepMap gene-knockoff effect scores of the gene set identified by MEDALT and those by Monocle from the 20 patients. Also included as controls are 100 sets of 75 random genes, 234 oncogenes, and 269 tumor suppressors (TS) in oncoKB and intOGen

are powerful at characterizing cellular heterogeneity, but small numbers of biological samples, which are underpowered for discovering common disease genes. It has been shown by recent genome-wide association analysis that it is possible to enable new discovery by performing association analysis at cell-type resolutions [55]. For cancer and genetic diseases driven by somatic mutations, being able to obtain genetic footprint at various time and conditions can enable discovery of genes responsible for disease progression and resistance to therapy.

However, it remains unclear what analytical strategies should be deployed to achieve the benefits. Even more challenging it gets when CNAs are being considered, as CNAs affect large regions of the genome and are difficult to trace using phylogenetics methods.

In our study, we demonstrated that it is possible to achieve the benefit by reconstructing copy number evolution history as a lineage tree, i.e., MEDALT, and performing permutation-based statistical analysis, i.e., LSA, to identify fitness-associated CNAs and genes.

We have learned several important lessons in our study.

First, it is important to perform accurate lineage tracing. Although the single-copy gain and loss model that we implemented in deriving MEDALTs is limited in complexity, it already performed substantially better than conventional phylogenetics algorithms such as MP that assumes infinite sites and NJ that employs naïve distance metrics, as shown in our simulation and in real data analysis. It is conceivable that further development of methodology that incorporates more complex genome evolution mechanisms such as chromothripsis [56] can lead to better results.

An important goal was to represent convergent evolution that is likely prevalent in the lens of CNAs [10, 57]. Conventional phylogenetics algorithms strictly prohibit the expression of convergent evolution by disallowing an alteration to occur multiple times in a course of evolution [28]. Several new algorithms relaxed such limitation but were designed for analyzing point mutation data [58]. As shown in our analysis of the TNBC patients, genes identified based on convergent evolution analysis (i.e., PLSA) had an even higher fraction of known cancer genes than those identified based on cohort-level single-lineage LSA. Our result suggests that examining convergent evolution is likely a key component towards fully unleashing the power of single-cell studies.

Unlike canonical phylogenetic trees, MEDALTs are minimal spanning trees that do not contain unobserved internal ancestral nodes. Representing evolution using minimal spanning trees instead of phylogenetics trees was our deliberate choice, as it allowed us to develop polynomial-runtime solutions that are scalable to real datasets containing thousands of cells. It also allowed us to conveniently implement biologically meaningful MED and enforce directionality constraints. Phylogenetics algorithms are likely effective when the numbers of cells are small and that the alterations are simple to trace. None of these conditions apply to available SCCN datasets that have CNAs evolving nonlinearly in hundreds of cells. Moreover, we have shown in our simulation that for the purpose of detecting fitness-association alterations, our method outperformed phylogenetics approaches in a wide range of sample sizes.

A particular challenge in developing and evaluating computational lineage tracing methods is the lack of exact ground truth. Although various experimental technologies

have been developed [59, 60], we are not aware of any that can be applied to trace copy number evolution in patient samples. To circumvent this, we utilized in silico simulation that mimics several prevalent CNA mechanisms to evaluate the accuracies of the reconstructed lineages and fitness-associated alterations. We also utilized longitudinal datasets on which we knew the biological stages of the cells to evaluate the chronological accuracy of the inference results. Although these strategies are unlikely sufficient to validate all the edges and lengths in the trees, they are objective and sufficient to discriminate various approaches.

Second, it is important to control biases in statistical inference. It is challenging to detect fitness-associated genes, as CNAs often affect a large number of genes and that the sample sizes are often small. Passenger CNAs that occur naturally in non-functional regions such as those near fragile sites or repeats could easily cloud the discovery. In addition, lineage tracing algorithms are unlikely to be perfect and could introduce distinct biases. To address these challenges, we employed LSA, which randomly permutes SCCN profiles into different cells to reduce the biases introduced by background genomic variations and technical noises. And we reconstructed trees from permutated datasets to alleviate biases introduced by the lineage tracing algorithms. The evolutionarily meaningful MED metrics and constraints help our analyses to focus on biologically relevant hypotheses, given limited computational resources. These procedures appeared important to achieve the accuracy. Further exploration of different ways to permute the data and to estimate the background distribution will likely lead to better results.

We assessed the functional impact of the identified genes using cell-line CRISPR essentiality screen data. We confirmed that the set of fitness-associated, amplified genes discovered by our methods are significantly more essential than other control gene sets in cancer cell lines. We also nominated novel genes that appear to have prognostic values in TCGA and the METABRIC datasets. These assessment strategies likely have false positives and negatives. Further comprehensive, well-controlled and targeted experiments will likely be required to fully assess the functional impact and clinical values of these genes.

Lastly, it was exciting to observe benefits of our methods on both the scDNA-seq and the scRNA-seq data. Although RNA-derived copy number profiles may not be as accurate as those derived from DNAs, previous studies [61] suggested that they can reasonably distinguish tumor clones. Our study further revealed the value of scRNA-seq data in lineage tracing and supported the notion that genomic profiles, even approximations, are more accurate than transcriptomic profiles in determining biological timing of cells. Our results opened doors towards utilizing scRNA-seq as a platform to understand genetics underlying developmental processes and perform gene discovery.

Conclusions

In this study, we describe two innovative algorithms: MEDALT based on MED tracing tumor lineage evolution and LSA discovering lineage expansion associated genetic drivers. We examined the algorithms using synthetic datasets, longitudinal scDNA-seq data obtained from TNBC patients and scRNA-seq data of HNSCC, MM, OSCC, and OV patients. Compare to conventional algorithms, our approach effectively improves.

Methods

Inferring minimal event distance

We use a modified parsimony scoring method to score the distance between two copy number profiles, which can be considered as non-negative integer arrays. We assume a copy number alteration (CNA) (event) can affect adjacent genomic regions (one single entry or k adjacent entries in array) by increasing or decreasing their values by 1. We define the minimal event distance (MED) between two arrays a and b to be the minimal number of CNAs needed to transition from a to b (Additional file 1: Fig. S1a).

We propose a greedy algorithm (Additional file 1: Table S1) which guarantees to find an optimal solution within a runtime of $O(m)$ (Additional file 6), where m is the size of the array [34]. We add an additional restriction that MED equals to infinity, if the copy number at any site is going from 0 to any other number. In addition, the amplification cannot span across the site with 0 copy number. This is different with Zeira et al., which utilized a zero-skipping rule [34].

Constructing Minimal Event Distance Aneuploidy Lineage Tree (MEDALT)

The optimal aneuploidy lineage tree is a rooted directed minimal spanning tree (RDMST) with the least number of CNAs. We use an implementation of Edmond's algorithm to infer RDMST (Additional file 1: Table S2). Our algorithm runs in $O(VE)$, where V is the node set and E is edge set. That is approximately as $O(n^3)$, where n is the size of the node set.

Lineage speciation analysis

We propose a statistic routine named lineage speciation analysis (LSA), which performs permutation tests on the topology of MEDALT or phylogenetics trees to identify CNAs that are non-randomly associated with cellular lineage expansion in a developmental process. In LSA, we start from the root node and iteratively remove edges to obtain all possible lineages (subsets of cells). For the i-th lineage, we calculate a cumulative fold level (CFL) for the j-th CNA event that sums together the copy number alteration level in constituent cells (Additional file 1: Fig. S1c).

$$CFL_{ij} = \sum_{k=1}^{K} \tag{1}$$

where CN_{ijk} is the copy number level in the k-th cell and K is the size of the lineage.

We treat the amplifications and deletions separately so that a region can be amplified in some samples but deleted in others. This is necessary because some oncogenes and tumor suppressors locate in close proximity and can get binned into the same regions.

We estimate the statistical significance of an observed CFL by comparing its value to a background distribution obtained through permutation (Additional file 1: Fig. S1d). In the default mode, SCCN data are randomly shuffled by chromosomes into different cells. They are not further shuffled by sites within each chromosome, because chromosomal context plays an important role in determining where and how a CNA occur.

In order to obtain an empirical background distribution, we permute SCCN data 1000 times and construct a lineage tree for each permutated SCCN dataset (Additional file 1: Fig. S1d). Similar to the process for the real tree, we dissect each permutated tree

into a collection of lineages. For each lineage from the real tree, we select the lineages in the permutated trees of identical (or very similar) size. If there is no lineage which has the same number of cells in one permutated tree, we will select the lineage with the next closest size. Thus, for each real lineage, at least 1000 lineages from permutated trees are selected. We compute CFLs of each CNA event in these selected lineages using Eq. (1) and construct corresponding background distribution to calculate an empirical p-value (tail probability) of the observed value:

$$p = \frac{\sum_{r=1}^{R} I(S_r \geq S_o) + 1}{R + 1} \tag{2}$$

where R is the number of background lineages from the permutation data, S_r, S_o are respectively the CFLs of the CNA event in the permutation and the real data.

To evaluate the performance of LSA for controlling biases in statistical inference, we estimated the significance using three additional ways:

(1) Rather than reconstructing a tree from each permuted SCCN matrix, estimate CFLs of cells from real lineage using the by-chromosome-permuted SCCN matrix from the real three.
(2) Same as (1) except using the SCCN matrix permuted by chromosomal bins within each cell (similar to GISTIC) instead of by chromosomes across different cells.
(3) One-side Wilcoxon signed-rank test to estimate if the levels of CNA is significantly higher/lower in cells from a lineage than those from other lineages in the same tree.

For (1) and (2), it is similar with LSA that we construct background distribution of CFLs and estimate empirical p-value using Eq. (2).

Cohort-level LSA

In a cohort containing multiple individuals, we can estimate whether a recurrent CNA identified at individual level occurs non-randomly at the cohort (population) level. To do so, we construct meta-lineages by merging lineages dissected from different individuals and calculate a CFL for each meta-lineage through Eq. (1). We then estimate a statistical significance for each observed CFL through Eq. (2), based on a background distribution obtained from corresponding meta-lineages derived from individually permuted trees in the entire cohort (Additional file 1: Fig. S1f).

Identifying parallel evolution event

The lineage speciation analysis (LSA) can be used to identify potential presence of parallel (aka. convergent) evolution (PLSA), i.e., finding CNAs that occur independently in multiple parallel lineages during the evolution of a cell population (Fig. S1g). We can assess the statistical significance of such events using the same permutation framework. Instead of examining each lineage independently, we deploy an algorithm that exhaustively searches for parallel lineages that are formed by disjoint sets of cells with identical CNAs or genes.

We then estimate the probability of observing such multi-lineage CNAs over random chance through permutation (as described above, Additional file 1: Fig. S1g):

$$p = \frac{\sum_{r=1}^{R} I(L_r \geq L_o) + 1}{R + 1} \quad (3)$$

where L_r, L_o are respectively the number of lineages containing the CNA of interest in the real and the permuted trees and $L_o \geq 2$. R is the number of permutations. In this analysis, only CNAs tested positive in the LSA are being further considered for the PLSA.

Simulating single-cell copy number evolution

Simulating cell birth-and-death process

In order to evaluate the accuracy of copy number lineage reconstruction, we implement a Markov process to simulate the cell growth under the influence of CNAs [39, 62]. The simulation process starts from an ancestor cancer cell, which divides and dies at rate b and d, respectively. All the descent cells have the same division and death rates as do their ancestors, unless they are mutated.

The cell growth dynamics follow the following differential equation:

$$\frac{dn(t)}{dt} = b \cdot n(t) \quad (4)$$

where $n(t)$ is the number of cells at time t. We assume that there are one root and 2 children after the first division: $n(0) = 1$, $n(1) = 2$. That leads to $b = 0.69$ as the initial value based on Eq. (4).

The distribution of the time intervals Δt between any two jumps in a Markov process with continuous time is exponentially distributed with the mean $E(\Delta t) = 1/(b + d)$ [63]. Here, we assumed $E(\Delta t) = 1$ and the death rate $d = 1 - b$. When a jump occurs, it results in a birth with a probability $b/(b + d)$ or a death with a probability $d/(b + d)$. This cell birth-and-death process can be depicted as a rooted directed tree in which nodes are cells.

We simulated 100 independent runs, each of which has a population size of 200 cells.

Simulating the occurrence of CNA events

CNAs accumulate among tumor cells at an appreciable rate [64]. The CNAs in a cell at time t_i not only include the alterations it inherits from its parent, but also newly acquired ones from t_{i-1} to t_i (Additional file 1: Fig. S3a). We assume that the CNA rate per site/region varies in several levels $\mu \in \{0.02, 0.05, 0.1, 0.15, 0.2\}$ [32] and determine the number of CNAs (K) accumulating in Δt based on a Poisson distribution (Additional file 1: Fig. S3a):

$$K \sim Poisson(\lambda = \Delta t * \mu * G) \quad (5)$$

where G is the total number of sites/regions in the genome. In our simulation, we set $G = 100$.

Simulating genomic structural rearrangements

We assume that CNAs can be generated by various types of genomic structural rearrangements (GSR), such as terminal deletion (TER), interstitial deletion (DEL), unbalanced translocation (UT), tandem duplication (TD), inverted duplication (ID), and

breakage fusion bridge (BFB) [30]. In addition, different GSRs could occur at differential rate in cancer [65, 66]. Thus, we determine the numbers of various GSRs based on a multinomial distribution [32].

$$X = \{x_1, x_2, ..., x_K\} \sim Multi(p_{TER}, p_{DEL}, p_{UT}, p_{TD}, p_{ID}, p_{BFB}) \quad (6)$$

where we empirically set $p_{TER} = p_{DEL} = 0.1$, $p_{UT} = 0.15$, $p_{TD} = 0.5$, $p_{ID} = 0.05$, $p_{BFB} = 0.1$. We also required that $K = \sum_{k=1}^{K} x_k$ during the period of Δt (Additional file 1: Fig. S3a).

Simulating the location of a CNA

CNAs affect contiguous sites/regions in a chromosome. They often exhibit two modes: (1) focal, affecting a relatively small (<MB) region [67], and (2) broad, encompassing large chromosomal regions (e.g., chromosomal arms) [68]. Broad CNAs often result from chromosomal mis-segregation during mitosis [64], which is a hallmark of cancer. Both focal and broad CNAs are important in oncogenesis. While broad CNAs often manifest through dosage effects [13], focal CNAs often target driver genes directly and result in protein structural changes [69].

We determined the size r of a CNA in X by sampling a zero-truncated Geometric distribution:

$$g(r, p) = p \cdot (1 - p)^{r-1} \quad (7)$$

where r is the number of genomic sites/regions that a CNA occupies and p the probability that a region is affected by the CNA (Additional file 1: Fig. S3a). We set $p = 0.5$ in our simulation.

We encode the simulated CNAs as sequences of non-negative integers in corresponding cells (Additional file 1: Fig. S3b). Our model allows single-copy gains and losses. A copy number gain increases the corresponding values by 1 and a copy number loss decreases the values by 1 (Additional file 1: Fig. S3b).

Simulating fitness-associated alterations

Some CNAs may themselves alter the fitness of a cell, or occur simultaneously with the driver mutations. We call them fitness-associated alterations (FAAs). We simulate the occurrence and the impact of FAAs in the evolution. At each generation, we determine if a FAA would occur through a Bernoulli distribution ($p = 0.5$). If a FAA occurs, we randomly select τ cells to carry the FAA, where τ follows a binomial distribution $B(\zeta, p = \frac{1}{\zeta})$ and ζ is the number of cells in the generation. The selected cells would increase their birth rates by s, which follows a uniform distribution $U(0, 1)$.

In order to estimate the effects of noise, we added noise at different levels in the simulated copy number profile based on a $Poisson(\lambda)$ model, where λ represents the mean number of randomly selected bins with increased or decreased CN values (by 1) in each cell. We set $\lambda = 0, 2, 4, 6, 8, 10$ with 0 being no noise, 10 corresponding to 10% of the genome.

Constructing phylogenetic trees

We construct phylogenetic trees using the R package *phangorn* [70], which implements widely used versions of the maximal parsimony, neighbor-joining and maximum likelihood approaches. To apply the maximal parsimony approach, the SCCN data are re-segmented by the collection of breakpoints detected in each cell, so that each column in the data matrix corresponds to a genomic interval that is uninterrupted by any GSR in any cell. The GSR breakpoints in individual cells are determined by the R package *copynumber* under default parameters. To apply the neighbor-joining approach, Hamming distances are calculated from each pair of the SCCN profiles. To apply the maximal likelihood approach, random trees are chosen as the initial solutions.

Estimating the accuracy of lineage partitioning

The cell birth-and-death process we simulate can be expressed as a rooted directed minimal spanning tree (RDMST). To compare RDMST with phylogenetic trees, we convert RDMSTs into dendrograms, which are fully comparable with the phylogenetic trees in that observed cells are represented as leaves in both types of representations [71]. From each dendrogram or phylogenetics tree, we calculate a metric, termed lineage partitioning accuracy (LPA), which measures how accurately cells are partitioned into lineages (subsets). Given a dendrogram, we performed lineage partitioning as follows:

We iteratively remove each branch in the dendrogram to obtain all the bi-partitions, i.e., the two disjoint subsets resulting from removing a branch. Each subset corresponds to a cellular lineage. All lineages can be described as a binary sequence $l = \{c_1, c_2, \cdots, c_N\}$, $c_i = 1$ if the i-th cell is in lineage l and $c_i = 0$, otherwise.

In the simulation experiments, the lineages partitioned from the simulated cell growth trees are considered as the ground truth. The LPA of a given MEDALT or phylogenetics tree is calculated as the fraction of lineages that exist in the ground truth over the total number of predicted lineages.

Accuracy of FAA detection in simulation

We randomly spike in FAAs in the simulation experiments, which are used as the ground-truth to assess the accuracy of the MEDALT and the phylogenetic trees. For each CNA, we calculate its p value through LSA and identify the minimal p value over all the lineages containing the CNA. We use $-\log(\textit{minimal } p)$ as the prediction score. We then characterize the accuracy of each approach on FAA detection using AUC values, which are calculated by tallying the positive and the negative hits at various prediction score cutoffs from 0 to the maximal values.

Identifying significant CNAs using GISTIC

We apply the GISTIC algorithm on the simulated and the real SCCN datasets to identify significant CNAs [37]. The following steps are taken:

i) Calculate the occurrence frequency (f) and the amplitude (Δ) of each alteration
ii) Define a G-score as a function of f and Δ: $G = f \times \log_2(\Delta + 2)$

iii) Assess the statistical significance of each alteration by comparing the observed *G*-score to a background distribution of *G*-scores obtained from permuted (by regions) copy number profiles

On the simulated datasets, we regard each cell as an individual sample and apply GISTIC at the cell level.

On the TNBC dataset, we average the SCCN profiles across the cells in each patient sample to create a pseudo-bulk copy number profile for each sample. We then run GISTIC on these pseudo-bulk profiles to identify significant CNAs, similarly to how GISTIC is applied in TCGA study.

Integer copy number profiles from single-cell DNA sequencing data

The SCCN profile is an integer-valued matrix. The SCCN profiles from single-cell DNA-sequencing data of triple-negative breast cancer are downloaded from the original paper [16, 18] which estimated using a variable binning method, as detailed in previous studies [18, 72]. Briefly, sequencing reads are counted in 11,927 genomic bins with variable start and stop coordinates, which are optimized to receive even read counts across the bins. The median genomic length spanned by the bins is 220 kbp. Cells with < 50 median reads per bin are excluded. Loess normalization is used to correct for GC bias [40]. Copy number profiles are segmented using circular binary segmentation (CBS) [73] followed by MergeLevels [74] to joint adjacent segments with non-significant differences in segment ratios (parameters alpha = 0.0001 and undo.prune = 0.05). Integer copy numbers are calculated by scaling segment ratios with average DNA ploidy determined by flow sorting indexes and rounding to closest integers [18].

Dissecting MEDALT into disjoint lineages

To characterize CNA rate variation and genetic organization of a cell population, we dissect it into disjoint lineages (cell subsets) based on the corresponding MEDALT. For each internal node v in MEDALT, the subtree rooted at v is denoted as T_v, which consists of all the descendants of v. The number of nodes in T_v is denoted as S_v, the size of the subtree. To ignore small lineages that cannot be confidently characterized, we set a minimal subtree size cutoff s ($s = 5$ in our analysis of the scDNA-seq and the scRNA-seq data) and define an internal node set $IV = \{v | S_v > s, v \in V\}$, where V represents the node set of the MEDALT. We arrange the node sets in IV in an increasing size order:

$$IV = \{v_1, v_2, \ldots, v_k | S_{v_1} < S_{v_2} < \ldots < S_{v_k}\} \tag{8}$$

To obtain disjoint lineages, we remove the internal nodes that lead to redundant lineage assignments. For each $v_i \in IV$, $1 \le i \le k$, its parent node v_j ($j > i$) should exist in IV. If a parent node v_j has more than one child in IV, remove the parent node v_j from IV; otherwise, remove the child node v_i. We iterate through all the nodes in IV until no node can be removed. We then split the MEDALT into subtrees rooting at nodes remaining in IV. All the nodes that are not yet included are assigned into a control lineage.

Estimation of CNA rate and fraction of DDR loss

We estimate CNA rate in a lineage as the average number of CNAs, i.e., average MED, between the cells in the lineage. DNA damage repair (DDR) genes play key roles in maintaining genome stability. In our analysis, we download the list of DDR genes from Knijnenburg et al.'s study [41], based on which we estimate the proportion of DDR genes with copy number loss in each lineage.

Characterizing chromosomal level CNAs identified in TNBCs

We perform cohort-level LSA per genomic bin at a 220-kb resolution. We define the chromosomal arm is significant if more than half of bins in the arm are significantly associated with lineage expansion. Average p value of these significant bins corresponds to the significance level of the chromosomal arm. In order to benchmark the accuracy of chromosomal (arm) level CNA detection in the TNBC data, we search biomedical literature exhaustively and create a list of chromosome-arm-level CNAs that have reported relevance to TNBC biology or clinical utilities (Additional file 1: Table S4). We treat this list as the ground truth.

For each chromosomal level CNA in a lineage tree, we used the $-\log(p)$ estimated via the cohort LSA as its prediction score. We then estimate AUC values, respectively for the MEDALT, the MP, and GISTIC approaches.

Inferring copy number profile from single-cell RNA sequencing data

We use R package *inferCNV* (https://github.com/broadinstitute/infercnv) to identify somatic large scale chromosomal CNAs from single-cell RNA sequencing (scRNA-seq) data [53]. *InferCNV* detects CNAs by exploring expression intensity of genes across positions of tumor genome in comparison to a set of reference "normal" cell. The CNAs at gene level relative to reference cell are estimated under default parameters of inferCNV. According to inferred gene-level relative copy number profile, we calculate average relative CNA values in non-overlapping genomic bins, each consisting of 30 genes. Within each bin for each cell, we calculate an integer copy number by multiplying the relative CNA value by 2 (diploid) and then rounding the results off to closest integers.

Estimating genetic homogeneity

We compute a metric, called genetic homogeneity level (GHL) to compare the accuracies of MEDALTs with those of Monocle trajectories in tracing genetic evolution from scRNA-seq data. For each cell lineage (subset) partitioned from a MEDALT (see the "Dissecting MEDALT into disjoint lineages" subsection of the "Methods" section), we calculate pair-wise Pearson's correlation coefficients between all the cells in the lineage, using gene-level copy number profiles inferred by *inferCNV*. We treat the mean correlation coefficient as the GHL of the lineage. Then average the GHLs across the lineages to obtain an overall GHL of the MEDALT.

Similarly, we calculate a GHL for a Monocle trajectory by averaging cluster-level GHLs estimated from cell clusters defined by the trajectory.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02291-5.

Additional file 1: Fig. S1. Methodology of the framework. **a.** Illustration of minimal event distance (MED) calculation. **b.** Average lineage partitioning accuracy (LPA) on 100 simulation datasets without noise. **c**. Estimating lineage specific cumulative fold level (CFL). **d.** Estimating significance of CFL in an individual sample. **e.** AUC of non-random fitness-associated alterations (FAAs) detection based on LSA, permutated SCCN matrix rather than reconstructing tree, GISTIC test and one-side Wilcoxon signed-rank test on 100 simulation datasets without noise. **f.** Identification of non-random fitness-associated CNAs in a cohort of samples. **g.** Identification of parallel evolution CNAs in an individual sample. **Fig. S2.** The efficiency of MEDALT based on $9 \times 3 \times 20$ simulation datasets with the population size from 400 to 2000, genome size from 100 to 1000. **Fig. S3.** Simulation and evaluation of CNA evolution model. **a.** Illustration of simulated genomic structural rearrangements in the evolution of a tumor. K represents the number of CNAs during Δt period. r represents the number of adjacent regions which are affected by a CNA. TD: tandem duplication. TER: terminal deletion. DEL: interstitial deletion. BFB: breakage fusion bridge. **b**. Simulated and inferred copy number evolution distance between two genomes. Compared with MED are commonly used distance metrics Hamming, Euclidean and Manhattan. **c.** The AUC for identifying FAAs based on different combinations of models. Wilcox represents one-side Wilcoxon signed-rank test. **d.** The effects of noise on FAAs detection. **Fig. S4.** SCCN profile of TNBC patient KTN102. Each row represents a cell from pre-, mid-, or post-treatment. **Fig. S5.** Average distance between root node and cells from pre-, mid- or post-treatment based on MEDALT, maximal parsimony (MP), neighbor-joining (NJ) and maximum likelihood tree. FC refers to the fold changes between the average distance to root of the mid−/post- cells and that of the pre-treatment cells. **Fig. S6.** Stratified average CNA rates and fractions of DDR genes loss among lineages (distinguished by colors) in 6 primary TNBC samples. **Fig. S7.** Gene set enrichment analysis (GSEA) for genes identified by LSA in patient t1. Colors correspond to branches. **Fig. S8.** Significant genes identified through cohort LSA from the TNBC scDNA-seq data. **a.** Venn diagram of the genes identified by the MEDALT, MP and GISTIC but not reported in oncoKB, COSMIC and intOGen. **b.** Overall survival (OS) analysis of breast cancer patients in TCGA. **c**. Progression free survival (PFS) analysis of breast cancer patients in TCGA. **d**. Overall survival analysis of breast cancer patients in the METABRIC. **e.** The fraction of cancer genes overlapping with events which were significant in single lineage (#Lineage = 1), multiple lineages (#Lineage > 1), parallel evolution test ((#Lineage > 1& PLSA < 0.001) and cohort LSA (inter-tumor recurrent). **Fig. S9.** Results of multiple myeloma patient 60,359. **a.** Inferred MEDALT and heatmap based on Pearson's correlation of the inferCNV profiles between cells ordered by lineages in MEDALT. **b.** Inferred trajectory from Monocle and heatmap of Pearson's correlation of the inferCNV profiles between cells ordered by states defined by Monocle. **Table S1**. The algorithm for minimal event distance (MED) inference. **Table S2.** The algorithm for rooted directed minimal spanning tree reconstruction. **Table S3** Information on the TNBC data**. Table S4**. Annotation of the broad CNAs identified in TNBCs based on literature. **Table S5**. Information on the scRNA-seq data.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Additional file 6.

Additional file 7. Review history.

Acknowledgements

Not applicable.

Peer review information

Andrew Cosgrove was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 7.

Authors' contributions

KC conceived the experiments. FW developed the statistical model LSA and performed all data analysis with the help from VM, SL, JD, JH, DM, and RG. QW developed MEDALT. LD and NN contributed to data and evaluation. KC, FW, and QW wrote the manuscript with input from all authors. KC supervised the work. The authors read and approved the final manuscript.

Authors' information

Twitter handles: @kchenken (Ken Chen)

Funding

This work was supported in part by the NIH [R01CA172652, U01CA211006, U01CA247760], the CPRIT [RP180248, RP180684], the MD Anderson Cancer Center Sheikh Khalifa Ben Zayed Al Nahyan Institute of Personalized Cancer Therapy grant [U54CA112970], and the NCI Cancer Center Support Grant [P30 CA016672]. This work was also supported by the Human Breast Cell Atlas Seed Network Grant (CZF2019-002432) from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

Availability of data and materials

The source code of MEDALT and LSA is available at GitHub (https://github.com/KChen-lab/MEDALT) [75], and the latest release is hosted by Zenodo [76], under the MIT License. Single-cell DNA sequencing data of triple-negative breast cancer (TNBC, $N = 20$) were downloaded from NCBI Sequence Read Archive (SRA) under accession number SRP064210 [77] and SRP114962 [78]. Single-cell RNA sequencing data of paired primary and metastatic/relapse head and neck cancer (HNSCC, $N = 6$), oral squamous cell carcinomas (OSCC, $N = 2$), and ovarian cancer (OV, $N = 4$) were downloaded from Gene Expression Omnibus (GEO) with accession number GSE103322 [79], GSE117872 [80], and GSE118828 [81].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, 1400 Pressler St, Houston, TX, USA. [2]Present Address: Precision Medicine Institute, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China. [3]Department of Computer Science, Rice University, Houston, USA. [4]Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, USA. [5]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, USA. [6]Department of Cardiovascular Sciences, Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, USA. [7]Department of Medicine, McDonnell Genome Institute Washington University School of Medicine, St. Louis, USA.

Received: 25 August 2020 Accepted: 9 February 2021
Published online: 23 February 2021

References

1. McConnell MJ, Moran JV, Abyzov A, Akbarian S, Bae T, Cortes-Ciriano I, Erwin JA, Fasching L, Flasch DA, Freed D, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science. 2017;356:eaal1641.
2. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. 2009;10:451–81.
3. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016; 17:224–38.
4. Zhang CZ, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes Dev. 2013;27:2513–30.
5. Potapova TA, Zhu J, Li R. Aneuploidy and chromosomal instability: a vicious cycle driving cellular evolution and cancer genome chaos. Cancer Metastasis Rev. 2013;32:377–89.
6. Brastianos PK, Carter SL, Santagata S, Cahill DP, Taylor-Weiner A, Jones RT, Van Allen EM, Lawrence MS, Horowitz PM, Cibulskis K, et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. Cancer Discov. 2015;5:1164–77.
7. Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet. 2012;44:642–50.
8. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, Cagan A, Murai K, Mahbubani K, Stratton MR, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018;362:911–7.
9. Turajlic S, Xu H, Litchfield K, Rowan A, Chambers T, Lopez JI, Nicol D, O'Brien T, Larkin J, Horswell S, et al. Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal. Cell. 2018;173:581–94 e512.
10. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, et al. Tracking the evolution of non-small-cell lung Cancer. N Engl J Med. 2017;376:2109–21.
11. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, Schumacher SE, Wang C, Hu H, Liu J, et al. Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell. 2018;33:676–89 e673.
12. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Rosebrock D, Mitchell TJ, Rubanova Y, Anur P, Yu K, et al. The evolutionary history of 2,658 cancers. Nature. 2020;578:122–8.
13. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nat Rev Genet. 2020;21:44–62.
14. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194:23–8.
15. Jiang D, Zhang X, Pang Y, Zhang J, Wang J, Huang Y. Terminal transfer amplification and sequencing for high-efficiency and low-bias copy number profiling of fragmented DNA samples. Protein Cell. 2019;10:229–33.
16. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. Cell. 2018;173:879–93 e813.
17. Casasent AK, Schalck A, Gao R, Sei E, Long A, Pangburn W, Casasent T, Meric-Bernstam F, Edgerton ME, Navin NE. Multiclonal invasion in breast tumors identified by topographic single cell sequencing. Cell. 2018;172:205–17 e212.
18. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai PC, Casasent A, Waters J, Zhang H, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48:1119–30.
19. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, Chen K, Scheet P, Vattathil S, Liang H, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.

20. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell. 2017;171:1611–24 e1624.
21. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.
22. Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. 2019;37:925–36.
23. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biol. 2020;21:208.
24. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC. Interactive analysis and assessment of single-cell copy-number variations. Nat Methods. 2015;12:1058–60.
25. Wang R, Lin DY, Jiang Y. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. Cell Syst. 2020;10:445–52 e446.
26. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, Richman AR, Silverbush D, Shaw ML, Hebert CM, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019;178:835–49 e821.
27. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306–13.
28. Schwartz R, Schaffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017;18:213–29.
29. Yang ZH. Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol. 1996;42:294–307.
30. Greenman CD, Pleasance ED, Newman S, Yang F, Fu B, Nik-Zainal S, Jones D, Lau KW, Carter N, Edwards PA, et al. Estimation of rearrangement phylogeny for cancer genomes. Genome Res. 2012;22:346–61.
31. Beerenwinkel N, Schwarz RF, Gerstung M, Markowetz F. Cancer evolution: mathematical models and computational inference. Syst Biol. 2015;64:e1–25.
32. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput Biol. 2014;10:e1003535.
33. El-Kebir M, Raphael BJ, Shamir R, Sharan R, Zaccaria S, Zehavi M, Zeira R. Complexity and algorithms for copy-number evolution problems. Algorithms Mol Biol. 2017;12:13.
34. Zeira R, Zehavi M, Shamir R. A linear-time algorithm for the copy number transformation problem. J Comput Biol. 2017; 24:1179–94.
35. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, Aas T, Alexandrov LB, Larsimont D, Davies H, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med. 2015;21:751–9.
36. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499:214–8.
37. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007;104:20007–12.
38. Stern DL. The genetic causes of convergent evolution. Nat Rev Genet. 2013;14:751–64.
39. Charlebois DA, Balazsi G. Modeling cell population dynamics. In Silico Biol. 2019;13:21–39.
40. Baslan T, Kendall J, Rodgers L, Cox H, Riggs M, Stepansky A, Troge J, Ravi K, Esposito D, Lakshmi B, et al. Genome-wide copy number analysis of single cells. Nat Protoc. 2012;7:1024–41.
41. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, Greene CS, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. Cell Rep. 2018;23:239–54 e236.
42. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. Cell. 2018;173:321–37 e310.
43. Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017;49:1779–84.
44. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol. 2017;2017:1–16.
45. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A, Lopez-Bigas N. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013;10:1081–2.
46. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47:D941–7.
47. Pereira B, Chin SF, Rueda OM, Vollan HK, Provenzano E, Bardwell HA, Pugh M, Jones L, Russell R, Sammut SJ, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
48. Li S, Chen PL, Subramanian T, Chinnadurai G, Tomlinson G, Osborne CK, Sharp ZD, Lee WH. Binding of CtIP to the BRCT repeats of BRCA1 involved in the transcription regulation of p21 is disrupted upon DNA damage. J Biol Chem. 1999;274: 11334–8.
49. Lips EH, Mulder L, Oonk A, van der Kolk LE, Hogervorst FB, Imholz AL, Wesseling J, Rodenhuis S, Nederlof PM. Triple-negative breast cancer: BRCAness and concordance of clinical features with BRCA1-mutation carriers. Br J Cancer. 2013; 108:2172–7.
50. Lips EH, Laddach N, Savola SP, Vollebergh MA, Oonk AM, Imholz AL, Wessels LF, Wesseling J, Nederlof PM, Rodenhuis S. Quantitative copy number analysis by Multiplex Ligation-dependent Probe Amplification (MLPA) of BRCA1-associated breast cancer regions identifies BRCAness. Breast Cancer Res. 2011;13:R107.
51. Sharma A, Cao EY, Kumar V, Zhang X, Leong HS, Wong AML, Ramakrishnan N, Hakimullah M, Teo HMV, Chong FT, et al. Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. Nat Commun. 2018;9:4931.
52. Shih AJ, Menzin A, Whyte J, Lovecchio J, Liew A, Khalili H, Bhuiya T, Gregersen PK, Lee AT. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. PLoS One. 2018;13:e0206785.

53. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV. Accessed 8 Feb 2020.
54. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32:381–6.
55. Watanabe K, Umicevic Mirkov M, de Leeuw CA, van den Heuvel MP, Posthuma D. Genetic mapping of cell type specificity for complex traits. Nat Commun. 2019;10:3222.
56. Forment JV, Kaidi A, Jackson SP. Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer. 2012;12:663–70.
57. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. Cell. 2017;168:613–28.
58. Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome Res. 2017;27:1885–94.
59. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet. 2020;21:410–27.
60. Baron CS, van Oudenaarden A. Unravelling cellular relationships during development and regeneration using genetic lineage tracing. Nat Rev Mol Cell Biol. 2019;20:753–65.
61. Durante MA, Rodriguez DA, Kurtenbach S, Kuznetsov JN, Sanchez MI, Decatur CL, Snyder H, Feun LG, Livingstone AS, Harbour JW. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. Nat Commun. 2020;11:496.
62. Novozhilov AS, Karev GP, Koonin EV. Biological applications of the theory of birth-and-death processes. Brief Bioinform. 2006;7:70–85.
63. Allen LJS. An introduction to stochastic processes with applications to biology. 2nd ed. Boca Raton: Chapman & Hall/CRC; 2011.
64. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10:551–64.
65. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature. 2009;462:1005–10.
66. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al. Diverse mechanisms of somatic structural variations in human cancer genomes. Cell. 2013;153:919–29.
67. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.
68. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010;463:899–905.
69. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, Calderaro J, Bioulac-Sage P, Letexier M, Degos F, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012;44:694–8.
70. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.
71. Gower JC, Ross GJS. Minimum spanning trees and single linkage cluster analysis. Royal Stat Soc Ser C-Appl Stat. 1969;18:54.
72. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A, Levy D, Esposito D, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011;472:90–4.
73. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557–72.
74. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics. 2005;21:4084–91.
75. Wang F, Wang Q, Mohanty V, Liang S, Dou J, Han J, Conterno Minussi D, Gao R, Ding L, Navin N, Chen K. MEDALT. Github. 2020. https://github.com/KChen-lab/MEDALT. Accessed 26 Jan 2021.
76. Wang F, Wang Q. Inference of minimal event distance aneuploidy lineage tree based on single cell copy number profile. Zenodo. https://doi.org/10.5281/zenodo.4468537.
77. Gao RL, Davis A, McDonald TO, Sei E, Shi XQ, Wang Y, Tsai PC, Casasent A, Waters J, Zhang H, et al. Single cell sequencing identifies clonal stasis and punctuated copy number evolution in triple negative breast cancer patients. Datasets, Sequence Read Archive. 2016. https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP064210. Accessed 1 Nov 2019.
78. Kim C, Gao RL, Sei E, Brandt R, Hartman J, Hatschek T, Crosetto N, Foukakis T, Navin NE. Adaptive and acquired evolution in response to chemotherapy in triple-negative breast cancer. Datasets, Sequence Read Archive. 2018. https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP114962. Accessed 1 Nov 2019.
79. Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, et al. Single cell RNA-seq analysis of head and neck cancer. Datasets, Gene Expression Omnibus. 2017. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103322. Accessed 8 Feb 2020.
80. Sharma A, Cao EY, Kumar V, Zhang X, Leong HS, Wong AML, Ramakrishnan N, Hakimullah M, Teo HMV, Chong FT, et al. Single cell RNA-seq profiles of primary, metastatic, drug-resistant and drug-holiday cells. Datasets, Gene Experssion Omnibus. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117872. Accessed 8 Feb 2020.
81. Shih AJ, Menzin A, Whyte J, Lovecchio J, Liew A, Khalili H, Bhuiya T, Gregersen PK, Lee AT. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. Datasets, Gene Expression Omnibus. 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118828. Accessed 8 Feb 2020.

Publisher's Note