A method for phylogenetic reconstruction of aneuploid cancers based on multiregional genotyping data

Natalie Andersson (natalie.andersson@med.lu.se)
Lund University https://orcid.org/0000-0002-3643-4404
Subhayan Chattopadhyay
Lund University
Anders Valind
Lund University
Jenny Karlsson
Lund University
David Gisselsson
Lund University https://orcid.org/0000-0002-0301-426X

Article

Keywords: tumor heterogeneity, pediatric cancer, phylogenetic reconstruction

Posted Date: January 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-140537/v1

Version of Record: A version of this preprint was published at Communications Biology on September 20th, 2021. See the published version at https://doi.org/10.1038/s42003-021-02637-6.

A method for phylogenetic reconstruction of aneuploid cancers based on multiregional genotyping data

Natalie Andersson[1], Subhayan Chattopadhyay[1], Anders Valind[1,2], Jenny Karlsson[1] & David Gisselsson[1,3,4]

[1]Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden. [2]Department of Pediatrics, Skåne University Hospital, Lund, Sweden. [3]Division of Oncology-Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden. [4]Clinical Genetics and Pathology, Laboratory Medicine, Lund University Hospital, Skåne Healthcare Region.

Correspondence and request for materials should be addressed to N.A. (email: natalie.andersson@med.lu.se).

Abstract word count: 129

Text word count: 6563

References: 37

Figures: 4

Tables: 1

Supplementary Figures: 7

Supplementary Datasets: 17

Abstract

Phylogenetic reconstruction of cancer cell populations remains challenging. There is a particular lack of tools that deconvolve clones based on copy number aberration analyses of multiple tumor biopsies separated in time and space from the same patient. This has hampered investigations of tumors rich in aneuploidy but few point mutations, as in many childhood cancers. Here, we present DEVOLUTION, an algorithm for subclonal deconvolution followed by phylogenetic reconstruction from bulk genotyping data. It integrates copy number and sequencing information across multiple tumor regions throughout the inference process, provided that the mutated clone fraction for each mutation is known. We validate DEVOLUTION on SNP-array data from 56 pediatric tumors comprising 253 tumor biopsies confirming concordance to biological mechanisms and show a robust performance on extensive simulations of bulk genotyping data.

Introduction

Neoplasms are a heterogenous group of diseases driven by Darwinian selection. Most cancers are presumed to originate from a single mutated cell, from which each mutation is conveyed to its daughter cells, that in turn can acquire additional aberrations, establishing subpopulations (subclones) of cells with diverse genetic compositions within the tumor[1]. This evolution of cancer cells is further shaped by genetic drift and selection pressure from the tumor microenvironment and oncological treatment[2-5]. Due to this process, many cancers exhibit vast intratumor heterogeneity (ITH) as well as intertumor heterogeneity between the primary tumor and its metastases[6-10]. Knowledge about how ITH emerges over time remains limited and multiple models have been proposed to explain it such as punctuated, neutral, linear, and branched evolution as well as a big bang model of tumor growth followed by neutral evolution[11-13]. By analyzing the genetic variation of the tumor spatially as well as temporally, mathematical methods can be employed in order to reconstruct its evolution, commonly in the form of a phylogenetic tree that links together distinct cancer cell subpopulations in an inferred temporal order. Such phylogenetic reconstructions can improve the understanding of tumorigenesis, progression to metastatic disease, and aid the development of novel therapeutic strategies[7,8,14].

One of the biggest challenges in phylogenetic analysis of bulk sample data from tumors is that the genetic analysis of each sample is conducted on millions of cells at once, usually constituting multiple subclones. The relative proportions of the subclones within each biopsy may also vary across the biopsied regions, stressing the need to integrate information from multiple biopsies separated in space to thoroughly assess the genetic profile of the tumor. Not addressing this may result in the prediction of illicit biological trajectories and so-called biopsy trees, not constituting true phylogenies[15]. Phylogenetic relationships should thus ideally be constructed based on the deconvolved clonal structure, i.e. one should infer which subclones are characterized by which alterations and, in addition, the ancestral order of these.

Even though single cell sequencing (SCS) has emerged as an important tool for temporal reconstruction that circumvents the issue of clonal deconvolution it is very costly to implement on the scales needed in

the clinic and usually provides limited sequence coverage[16]. A more cost-effective alternative is to perform computerized deconvolution of bulk genotyping data derived from single nucleotide polymorphism array (SNP-array), targeted deep sequencing (TDS), whole exome sequencing (WES) or whole genome sequencing (WGS). Bulk genotyping yields a set of genetic alterations present in each biopsy along with information that can be utilized to estimate the proportion of cells harboring each aberration in each biopsy, denoted the mutated clone fraction (MCF)[17]. However, most tools developed for computerized deconvolution of bulk genotyping data focuses solely on somatic point mutations, presumes a diploid background, and lacks specific pipelines to handle intratumoral heterogeneity of copy number alterations (CNAs) of chromosomal segments or whole chromosomes. In addition, they do not provide the possibility to infer phylogenetic trees solely on copy number aberration data from multiple biopsies separated in time and space[18-21]. Since most cancers are aneuploid to some degree[22], this is a serious shortcoming, especially for cancer types where aneuploidy is a common feature such as high-grade adult carcinomas and many childhood cancers[8,23]. Consequently, there is a particular need for tools capable to infer phylogenetic trees based on multiregional copy number data.

To fill this methodological gap, we introduce DEVOLUTION, an algorithm for subclonal deconvolution followed by phylogenetic reconstruction from bulk genome profiles including high-resolution copy number data (e.g. from SNP-array, WES or WGS) and sequencing information (e.g. from WES, WGS or TDS) separately or in unison. The deconvolution is based on à priori MCF-estimation of the individual aberrations in each sample and the algorithm systematically combines information from all available biopsies throughout the inference process to reconcile the most probable temporal evolution of the tumor by inferring an event matrix that is used to reconstruct phylogenetic trees. Importantly it can deduce evolutionary trajectories based on copy number data alone. In addition, predictions of the subclonal size and compositions across biopsies are visualized directly in the phylogenetic tree. DEVOLUTION provides an objective framework for creating event matrices and phylogenetic trees from bulk genotyping data, avoiding subjective bias compromising the validity of tree-to-tree comparisons (S. Table 1).

To demonstrate DEVOLUTION's utility, the algorithm was evaluated using SNP-array data from 253 tumor regions from 56 pediatric cancers including neuroblastoma (NB), Wilms tumor (WT), and rhabdomyosarcoma (RMS), comprising the most common extracranial, solid tumors in children. Additionally, the algorithm showed a robust performance on simulated multiregional bulk genotyping data. DEVOLUTION holds the potential to facilitate further insights into the development, progression, and response to treatment, particularly in tumors with high burden of chromosomal copy number alterations.

Results

Overview of the algorithm workflow

The algorithm operates on multiregional sampling data analyzed using whole genome profiling followed by MCF-computation (**Figure 1a-c**). The input file is an $u \times v$ dimensional matrix, containing information about the u genetic alterations detected in a tumor. For each alteration there are v columns indicating the genetic position, alteration type as well as the proportion of cells in the biopsy harboring the alteration (**Supplementary Figure 1**). For copy number aberrations, the matrix is subjected to an algorithm identifying all unique events across the samples while considering the uncertainty in the aberration breakpoint measurement (**Supplementary Figure 2**). The DBSCAN (density-based spatial clustering of applications with noise) algorithm is then used to identify clusters of genetic alterations having similar cellular proportions across multiple samples, indicating that they might be reflecting a group of cells having an identical genetic profile (**Figure 1d**). By identifying these clusters of genetic alterations, the computational load can be decreased and the unfolding of the subclonal composition aided.

DBSCAN prepares the data set to be subjected to the subclonal deconvolution algorithm, which is employed to elucidate the temporal order of the clusters of mutations. Information from multiple

biopsies is integrated throughout this process to minimize the occurrence of parallel evolution (PLC) and back mutation contradictions (BMC). PLC in copy number data means that the same type of genetic alteration with the same genomic start- and endpoints in the chromosome appears independently in different cells within the tumor. This is, in most cases, unlikely from a biological standpoint, unless there is a copy number alteration including an entire chromosome. BMC incorporates genetic alterations that are gained and then lost further down the evolutionary history in the tree, which may be feasible scenarios for some types of genetic alterations such as a gain of a whole chromosome that is later lost but are less likely to occur for structural chromosomal aberrations and point mutations, and should never be occur for loss of heterozygosity events[24].

In addition, the user can provide a matrix containing information about illicit orders of genetic events, that can be taken into consideration during the deconvolution (**Figure 1e**). The deconvolution culminates in a suggestion of the most likely temporal order of all genetic aberrations, constituting the basis for the creation of an event matrix, illustrating the distribution of genetic alterations across subclones (**Figure 1f**). Using this event matrix, the biological distance between the subclones is calculated using the Hamming distance[25] and phylogenetic trees are reconstructed using the maximum likelihood and parsimony methods. In addition, the algorithm provides the distribution and size of the clusters across the samples, resulting in an overview of the dynamics, spatial distribution, and dissemination of the tumor (**Figure 1g**).

Validation on pediatric tumors confirms concordance to biological mechanisms

DEVOLUTION was applied to a previously reported dataset of 56 pediatric cancers and phylogenetic trees were generated based on copy number data for 22 neuroblastomas, 20 Wilms tumors and 8 rhabdomyosarcomas comprising a total of 253 biopsies (**Figure 2, Supplementary Figure 3, Supplementary File 1**). Event matrices but not phylogenetic trees could be reconstructed from six patients (NB1, NB24, WT1, WT2, WT3 and WT5) in which all samples from the same patient had identical genomic profiles.

The phylogenetic trees of the remaining 50 tumors all represented plausible biological scenarios and often illustrated key events in tumor evolution. For example, in NB5 (**Figure 2a**), where a primary tumor and a metastasis presented at the same time, the metastasis was demonstrated to originate from a population of cells having the genetic alterations of the stem as well as one group of cells encompassing a subclone also present in the primary tumor. This indicates polyclonal seeding. A more complex pattern of polyclonal seeding was observed in NB22 (**Figure 2b**), a patient with progressive tumor growth across multiple metastatic locations. Here there was more subclonal variation among the lymph node metastasis than among metastases to distant organs. The metastases to distant organs often presented as solitary branches, exemplified by the same subclone colonizing both the lung and skull. This might indicate that the threshold for tumor cells to escape the primary tumor and colonize the lymph nodes is lower than for colonization of distant organs, displayed as a wider variety of different subclones across lymph node locations compared to an extensive selection for a certain subclone in distant loci. WT11 and WT19 (**Figure 2c and d**) both showed subclones that were distributed across several locations within the primary tumors, a phenomenon which has previously been demonstrated to be common in Wilms tumors[17]. RMS8, an alveolar rhabdomyosarcoma (**Figure 2e**), displayed an intricate evolutionary pattern with many genetic alterations. Here the primary tumour's subclones form a cluster at the root of the tree, while the cell populations from a metastasis and a local relapse share a branch having a vast amount of additional genetic alterations. Hence, phylogenetic trees produced from copy number profiles by DEVOLUTION, can provide biological insights that might aid the understanding of how cancer develops and progresses in individual patients.

Contradictions are rarely seen in the phylogenetic trees

In 80 % of all tumors analyzed, the maximum likelihood (ML) and parsimony (MP) methods resulted in identical phylogenetic trees (19/22 NB, 17/20 WT, 4/8 RMS) (**Figure 3a-d**). When the ML and MP tree for the same case did differ from one another, the differences in the branching structure were minor (**Supplementary Figure 3**). We identified the positions of the genetic copy number alterations in each

tree to identify contradictions based on prior knowledge about how genetic aberrations occur in cancer cells. More specifically, we analyzed instances of PLC and BMC.

When the ML and MP trees differed from one another, this was always due to a PLC/BMC in the phylogeny, often with both contradictions together in the same tree. PLC and/or BMC were found in 14/50 ML-trees and 14/50 MP-trees (5 NB, 5 WT, 4 RMS), hence 28/100 trees in total. Of these, 3/5 NB, 3/5 WT and 1/4 RMS trees contained only one single contradiction located among the leaves of the trees i.e. it did not have any significant impact on the tree structure. In addition, there did not seem to be any apparent difference between the frequency of the types of contradictions between ML and MP trees (**Figure 3e**). Excluding the cases with PLC and BMC of whole chromosomes, which are plausible events, only 8/50 ML-trees and 8/50 MP-trees (0 NB, 4 WT and 4 RMS) exhibited contradictions in the tree structure. In these eight cases, merely a few genetic alterations caused the PLC and/or BMC. They were caused by aberrations altering clone size compared to another event across samples or similar aberrations that still fell outside the breakpoint cutoff for similarity causing them to be considered separate events. These situations may be resolvable by critically reviewing the original data (**Supplementary Figure 4**). Alternating clone sizes was particularly common in the RMS trees (**Supplementary Figure 5**). The RMS tumors had a significantly higher total branch length than NB and WT (**Figure 3b**), indicating a more complex genomic profile. They also had a mean number of genetic alterations per biopsy of 20, most of them present in > 50 % of cells in a single biopsy thus allowing just one single solution of the temporal evolution. To prompt review of original data when pertinent, the software will warn the user that there is a contradiction in the data set and the tree might therefore not be entirely biologically accurate.

Evaluation using simulated data

The reliability of the algorithm was further evaluated using simulated bulk sampling data. In the patient data set, the median number of unique subclonal alterations in total across all biopsies were 6 for NB, 5 for WT and 13 for RMS. To accommodate this large variation, the simulation was conducted for three

different mutation frequencies resulting in 15, 50 and 100 subclonal genetic alterations distributed across 40000 virtual tumor cells (**Figure 4a**). Virtual biopsies were sampled randomly from the set of cells while varying the number of biopsies from 1 to 10, generating a segment file along with a list of true unique subclones across biopsies for each mutational frequency. Hence, analysis could be performed using DEVOLUTION while having the true subclonal composition at hand.

As expected, when increasing the number of biopsies or the mutation frequency more genetic aberrations were identified. In addition, a higher number of subclones were correctly allocated. (**Figure 4b**). A higher number of overall mutations will provide the software with more information, which is why the number of correctly allocated genetic alterations increase with mutation frequency for the same number of biopsies. Specifically, the number of positions in the temporal sequence at which genetic alterations can be allocated decreases with the number of genetic alterations. The number of unique subclones increases while the proportion of cells in each biopsy representing each subclone decreases. However, when more alterations are correctly allocated in absolute numbers, the proportion of correctly allocated alterations will not. Throughout the different mutational frequencies 93 $\pm$ 5.5 % of the genetic alterations were correctly allocated in the event matrix (**Figure 4c**). Thus, when increasing the number of biopsies, the absolute number of correctly allocated genetic alterations increases, but the proportion of correctly allocated alterations does not change significantly. The reason for this is that sampling additional locations will also increase the chance of finding an area with a late genetic alteration that is hard to correctly place in the phylogeny because of its low spatial dissemination. We further dissected why not all subclones were correctly allocated. It was hypothesized to occur due to low spatial dissemination, resulting in the presence of certain genetic alterations in a subset of biopsies. Excluding all genetic alterations found in only one single biopsy in fact resulted in the correct allocation if 99.5 $\pm$ 0.8 % of the genetic alterations in this mixture of entities (**Figure 4d-e**).

Relationship to existing work

Much progress has been made in the field of clonal deconvolution. Many methods are although limited to integrating information from a single biopsy, such as TITAN and THetA in addition to only accepting sequence data[26,27]. PyClone and SciClone are, however, employable on multiple biopsies, but assume that all detected CNAs are clonal i.e., present in all cells, do not infer the evolutionary relationship between the identified clusters and sequence data is required for them to operate[19,20]. In addition, SciClone focuses exclusively on SNVs in copy number neutral and loss of heterozygosity (LOH) free portions of the genome. Both PyClone and SciClone mainly operate as clustering algorithms and do not infer the order of genetic alterations, hence their output could be used as an input to DEVOLUTION. Hence, they fulfill different purposes and is not meant to exclude one other. PhyloWGS on the other hand can use both SNAs and CNAs to infer a phylogeny and is employable on multiple samples. The algorithm does although not integrate information between samples during the inference procedure, representing a loss valuable information, and is limited to WGS data[18]. Also, other methods such as Clomial, LiCHeE and SCHISM are specifically designed for SNAs and cannot solely include CNAs to infer a phylogeny[28-30]. REVOLVER also requires sequencing data, cannot use CNA-data alone and is specifically designed to integrate phylogenies from large cohorts of patients in order to infer common trajectories of repeated evolution. SPRUCE requires input of sequencing data to infer phylogenetic trees. Hence, most methods available focus exclusively on sequencing data and there is currently no available tool that can reconstruct a phylogenetic tree based on multiregional SNP-array data alone, which is a commonly used genotyping method in the clinic. Additionally, many methods focused on SNA only outputs clusters of genetic alterations and their MCF. Event matrices are subsequently often constructed manually from these MCF estimates, posing a risk for unintentional subjective bias in the deconvolution process, especially when integrating information from multiple biopsies conjointly. These methodological gaps are filled by DEVOLUTION, which provides a subjective method to infer phylogenies based on a priori MCF estimations based on preset rules that are employed equally across all patient data, hence providing a standardized framework for inferring phylogenies from bulk genotyping data, thus allowing tree-to-tree comparison without the risk of bias from subjective curation.

Discussion

A single biopsy from a tumor can consist of multiple distinct subclones and their prevalence may vary across biopsied areas. Not addressing this fact when studying cancer cell evolution can be deleterious and result in incorrect phylogenies[15], stressing the need for multispatial and temporal sampling to unfold the genomic landscape. DEVOLUTION thoroughly assesses the problem by combining information obtained across multiple biopsied regions throughout the entire subclonal deconvolution effectively deconvolving subclones transversing clonal territories, with the potential to concomitantly include both point mutations and copy number alteration data. In contrast to other methods DEVOLUTION allows phylogenetic trees to be constructed using copy number information alone and integrating information from multiple biopsied areas throughout the inference procedure. The algorithm can combine data from SNP-array, TDS, WES and WGS, provided the MCF for each genetic alteration is known, which can be computed based on the log2 ratio for copy number alterations or variant allele frequencies (VAF) for point mutations as described extensively elsewhere[17,19,20]. Since aneuploidy is a common feature across many adult carcinomas and a majority of childhood cancers, the integration of copy number aberrations in the phylogeny holds the potential to increase the understanding of the evolution of these diseases. If needed, there is also a possibility for user curation, for example if it is known that certain genetic alterations cannot co-exist.

Evaluating the software using high-throughput SNP-array data from 253 multitemporal and spatial regions from 56 pediatric tumors produced phylogenetic trees that were well in concordance with prior knowledge of how chromosomal aberrations occur in cancer cells. Surprisingly, generating phylogenetic trees using ML and MP predominantly yielded identical tree structures. When thoroughly examining these trees, contradictions such as PLC and BMC were identified in the cases where the ML/MP trees differed, which were found to always be due to disagreements in the original data set. The user is

therefore encouraged to reevaluate this genetic alteration in the input segment file, since it may be particularly subjected to noise or keep the tree if the phylogenetic situation is considered biologically plausible. The extensive evaluation of the two methods did not indicate that the choice of mathematical method favors a certain type of error in the phylogenetic tree.

Clustering genetic alterations using DBSCANs were sufficient for analyzing the pediatric tumors incorporated in this study and the simulated data sets. The size of ϵ can be changed by the user to increase the flexibility in the clustering, thus optimizing it further for the data set at hand to account for noise. Note that the purpose of the ad hoc clustering mainly is to reduce the computational complexity. If the clustering is too strict, the DEVOLUTION algorithm compensates by thoroughly integrating the information across samples. The clustering algorithm can easily be changed.

In summary, we have seen how DEVOLUTION can be used to analyze the intratumoral heterogeneity as well as the intertumoral heterogeneity between the primary tumor and metastasis through evaluation on a dataset of pediatric tumors harboring extensive aneuploidy. By analyzing a cancer's phylogenetic tree, an overview of its heterogeneity and temporal order of genetic alterations can be assessed, which can be used to follow the tumor's evolutionary response to treatment[31,32] and may aid the identification of subclones posing a risk of metastasizing, relapsing or being resistant to therapy. It may also make it possible to identify genetic alterations that seems to appear early in the tumor's development, posing attractive targets for therapy since they are present in a large proportion or all cells in the tumor.

Conclusions

DEVOLUTION produces phylogenetic trees from multiregional cancer genomics data, integrating copy number aberrations and sequence mutations separately or in unison, provided the MCF is known. The algorithm reliably deconvolves subclones and infers the evolutionary history in well in concordance to prior knowledge of how genetic aberrations occur in cancer cells. It performs robustly across real data from multiple aneuploid tumor types and in simulated data across a wide variation in mutation rates and sample numbers. DEVOLUTION holds the potential to facilitate insights into the development,

progression, and response to treatment, particularly in tumors with high burden of chromosomal copy number alterations.

Methods

The software

The major structure of the software can be divided into five steps

1. Preprocessing of the data.
2. Clustering of genetic alterations based on information from multiregional sampling from the same patient.
3. Subclonal deconvolution based on information from multiregional sampling from the same patient.
4. Construction of an event matrix.
5. Usage of a mathematical model to reconstruct the phylogenetic trees, in this case
    a. Maximum likelihood
    b. Maximum parsimony

Preprocessing of the data

The input data consist of an $u \times v$ dimensional matrix containing information about the u detected genetic alterations present in each biopsy. The matrix should also specify the genetic location of each alteration, its type (gain, loss, cnni etc.) as well as the proportion of cells harboring the alteration in that particular biopsy (the mutated clone fraction, MCF), represented by the *v* columns (**Supplementary Figure 1**). How the MCF can be computed from log2 ratios as well as VAFs is described extensively elsewhere[17]. Also, methods such as PyClone and SciClone could be used to infer clusters of somatic mutations and their corresponding VAFs across samples, which could then be used to calculate the MCF.

If allelic copy number alterations are considered, the user is advised to choose a cutoff for the detected genetic alterations in the segment file to be considered separate events, reflecting the measurement uncertainty regarding the start and end positions of the genetic alterations. The default cutoff is 1 Mbp. The user can also choose which data types to include in the analysis. In this way e.g. SNP array and sequencing data can be analyzed separately for comparison or in unison without having to separate the matrix manually.

The algorithm scans the MCFs for missing values, indicating that the MCF has not been able to be determined. If the event is considered to belong to the stem, based on biological knowledge or additional data, the missing value is replaced by 100 %. Amplicon accumulation is an example of such a case when it is not possible to determine the fraction of cells harboring it since the number is hypervariable. A stem event is defined as the presence of the alteration in ≥ 90 % of the cells in all samples. The alterations containing missing values for MCF are removed entirely if part of a subclone to not overestimate genetic variation within the tumor.

The clustering algorithm was constructed to localize all unique genetic alterations throughout the tumor samples. The program loops through the rows of the data file representing the genetic aberrations. For each row it compares the genetic alterations and their position on the chromosome to all the other rows, representing other detected genetic aberrations throughout the samples. If the events' start or end positions differ by a certain cutoff, set by the user based on the measurement uncertainty of the data set, and/or they are different aberration types, they are considered as two separate events, else they are considered as the same event (**Supplementary Figure 2**). Thus, *all* conditions stated below must be met for the algorithm to consider two alterations detected in the same patient to be the same.

1. Alteration 1 and 2 are localized on the same chromosome.
2. Alteration 1 and 2 harbor the same type of alteration.
3. Neither alteration 1 nor 2 should belong to the stem.
    a. Alterations belonging to the stem are always considered as separate events.
4. $X_1 = \left\| A_{1_{start}} - A_{2_{start}} \right\| \leq co_{ev}$

335 **5.** $X_2 = \left\| A_{1_{end}} - A_{2_{end}} \right\| \leq co_{ev}$

336 In the present study, considering allelic copy number aberrations, the cutoff (co_{ev}) for measurement
337 uncertainty in the start and end position of the events was set to 1 Mbp, also constituting the default for
338 DEVOLUTION. Since the chromosome sizes ranges from 48-250 Mbp this cutoff constitutes a start and
339 end point deviation of 0.4-2 % of the chromosome length.

340

341 **Clustering of genetic alterations incorporating information from multiregional and temporal**
342 **sampling from the same patient**

343 In our model a tumor is proposed to consist of multiple subpopulations of cells that harbor different sets
344 of genetic alterations. Each individual alteration is part of a mutation space $m_i \in \{m_1, m_2 \dots m_\theta\}$
345 comprising all mutations present in the tumor where $i, \theta \in \mathbb{N}^+$ and θ is the total number of mutations.
346 The mutational profile obtained from the biopsies thus represent a subset of the total mutation space and
347 is the information at hand to describe the evolutionary trajectory of the tumor. For this purpose, all
348 detected mutations are combined with their respective MCF-values into a matrix representing their
349 distribution across samples (**Supplementary Figure 6**). For a particular tumor, this results in a matrix
350 T_{MxB} with the dimensions M × B, where M is the total number of unique genetic alterations and B is the
351 total number of biopsies available. Hence m_δ indicates a certain genetic alteration δ, and b_ω represent
352 a biopsy ω. The value $t_{\delta\omega}$ consequently corresponds to the MCF for an alteration, m_δ, in a sample, b_ω
353 where $t_{\delta\omega} \in [0,100]$ i.e. it is bound between 0 and 100 %. This can be written as

354 $$T_{MxB} = \begin{bmatrix} t_{1,1} & t_{1,2} & \dots & t_{1,B} \\ t_{2,1} & t_{2,2} & \dots & t_{2,B} \\ \vdots & \vdots & \ddots & \vdots \\ t_{M,1} & t_{M,2} & \dots & t_{MB} \end{bmatrix}$$

355 $$\text{where } t_{\delta\omega} \in [0,100], \delta \in \{1, \dots, M\}, \omega \in \{1, \dots, B\} \text{ and } \delta, \omega \in \mathbb{N}^+$$

356 In order to generate phylogenetic trees illustrating the relationship between the subclones present in the
357 tumor, which aberrations reside in the same cells as well as which subpopulations of cells the tumor
358 consist of must be determined. To solve this, the idea is that a true subclone of cells should form a cluster

of unique genetic alterations that persist. They should remain grouped irrespective of inclusion of new data from an additional region of the primary tumor or metastasis. Alterations that seem to follow each other are more likely to be in the same cells. The first step is thus to yield a clustering to identify groups of genetic alterations, uniquely identifying a certain subclone. The subsequent step is to determine the temporal order of the alterations in question, since each subclone will represent a linear combination of the clusters identified. Note that for DEVOLUTION the clustering is only used to reduce the computational complexity for the upcoming subclonal deconvolution algorithm. Mostly alterations are clustered that show similar MCF in all available biopsies. This is not to be confused with the more intricate clustering methods used in for example SciClone.

Density based clustering techniques such as DBSCAN[33] are superior at unsupervised clustering of non-uniform clusters. Furthermore, the number of clusters does not have to be specified beforehand, which you have to do with many other established clustering algorithms. In addition, it does only have two hyperparameters named *minPts*, which is the minimal number of points that is allowed in a cluster, and ϵ representing the radius in which points i.e. the genetic alterations' position in the B-dimensional space, are included, where B is the total number of biopsies. If ϵ is chosen too small a large part of the data might not be clustered and choosing it too big will put all alterations in the same cluster. The choice of ϵ can be aided by using a k-distance-graph which illustrates the distance to the minPts-1 = k nearest neighbor. The value to choose is when this plot shows an elbow which can be obtained by visual inspection. Another method would be to create a vector between the start- and endpoint of the graph. Then create a vector perpendicular to this having its start position at this line and its end position at our data curve. The elbow can be computed by finding the vector with the largest magnitude. The clustering method is not integrated directly in the subclonal deconvolution algorithm described below and is therefore easy to replace or alter if needed for noisier data[33].

The algorithm provides a matrix containing all clusters of genetic alterations. Let C_{KxN} be the matrix representing the clusters of genetic alterations. It has the dimensions K x N where K is the number of genetic alterations in the cluster and N is the cluster number. All matrix positions $c_{kn} \neq 0$ are unique i.e. the same genetic alteration cannot belong to multiple clusters.

386 $$C_{KxN} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,N} \\ c_{2,1} & c_{2,2} & \cdots & c_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{K,1} & c_{K,2} & \cdots & c_{KN} \end{bmatrix}$$

387 where $c_{kn} \in m_\delta \wedge c_{kn} \neq c_{ed}, (\forall\, k, e \in \{1, \dots, K\}\ \&\ n, d \in \{1, \dots, N\} \wedge c_{kn} \neq 0)$

388 A matrix representing the clusters present in each biopsy and their size determined by the mean of the
389 aberrations in the cluster is also constructed.

390 $$Z_{CxB} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,B} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,B} \\ \vdots & \vdots & \ddots & \vdots \\ z_{C,1} & z_{C,2} & \cdots & z_{CB} \end{bmatrix} \text{ where } z_{cb} \in [0,100]$$

391 Where c is a specific cluster of aberrations, b the biopsy and z_{cb} the size of the cluster c in sample b.

392

393 **Subclonal deconvolution based on information from multiple samples from the same patient**

394 The space of a single biopsy is 100 % and the space of all biopsies can thus be represented by a matrix
395 where p is the partitioning of the available space in the biopsy and s_{pb} is the space available in a specific
396 partitioning p in biopsy b. Initially $s_{1,b} = 100 \wedge s_{p\neq 1,b} = 0$.

397 $$S_{PxB} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,B} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,B} \\ \vdots & \vdots & \ddots & \vdots \\ s_{P,1} & s_{P,2} & \cdots & s_{PB} \end{bmatrix} \text{ where } s_{pb} \in [0,100] \text{ and } \sum_{p=1}^{P} s_{p,b} = 100 \wedge b \in \{1, \dots, B\} \in \mathbb{N}^+$$

398 The clusters of aberrations in each biopsy, as supplied by Z_{CxB}, are allocated to the space in decreasing
399 order, altering the magnitude of the spaces in S_{PxB} based on the MCF of the clusters allocated to it.

400 The allocation iteration algorithm is initially conducted considering each sample individually, resulting
401 in a matrix encompassing all possible allocations of the clusters in every biopsy. Subsequently, all
402 possible allocations throughout the samples are addressed to minimize the occurrence of parallel
403 evolution. The algorithm hence tries to produce one uniform solution of the temporal order of events
404 that does not contradict any information provided in the biopsies. The solution should be in concordance
405 to every biopsy provided. If not possible, parallel evolution or back mutations will occur in the final

phylogenetic tree and the user is advised to reconsider the original data set, since it may be biologically unlikely. It may be possible to allocate a cluster to multiple positions without producing contradicting temporal orders in any of the samples, for which the largest available space assumption is employed to make an objective decision, based on the presumption that the mutational frequency is equal in all cells within the biopsy no matter how many mutations they have. The cluster will consequently be placed as a descendant to the cluster constituting the largest proportion of the biopsy, in the absence of further biological information steering it elsewhere. Clusters presenting with only one possible allocation in a biopsy provides especially valuable information concerning the temporal order of events, e.g. having a group of alterations that are all present in all cells in a biopsy clearly indicates that there exists a group of cells in the tumor harboring all of these alterations, aiding the temporal allocation in other samples where these alterations may present themselves as subclonal. DEVOLUTION does allow some overlap in cellular frequency in the allocation algorithm taking into consideration the measurement uncertainty of MCF. This iterative computation results in the subclones present in the biopsies along with an estimation of their size and distribution across the samples.

Incorporating user-controlled rules for avoiding imposition of illicit biological trajectories

Some genetic aberrations present in the data set might be known to never occur in the same cell for some well-known biological reason. Such constraints should optimally be supplied to the algorithm to ensure biologically plausible solutions. The user can therefore provide the DEVOLUTION algorithm with a matrix indicating which genetic aberrations in the data set that cannot be placed after one another. The first column represents a mother genetic alteration that the daughter alteration specified in the second column, cannot have (**Supplementary Figure 7**). The subclonal deconvolution algorithm extracts a list for each genetic alteration containing information about in how many of the samples it can be allocated after a certain cluster. There might be multiple possible solutions, equally prevalent. In this instance the matrix containing information about illicit biological orders can aid the program in taking a decision regarding which of these allocations are less likely, subsequently discarding them. These rules will thus only be employed if the data set allows the genetic alterations to be placed in any other way. If the only

possible way for the events to be allocated is to place them as descendants, the user will be advised to revise the original data set. No such rules were integrated in the analysis of the 56 pediatric tumors in the present study.

Construction of an event matrix

Based on the estimated subclonal composition, an event matrix $E = [\hat{a}_1, \hat{a}_2 \ldots \hat{a}_k]$ was constructed illustrating the distribution of genetic alterations across the identified subclones. Each $\hat{a}_i$ is binary vector belonging to subclone *i*. Each row represents a genetic alteration indicated with a 1 if present or 0 if absent in the subclone. The event matrix is used as the foundation for phylogenetic tree generation, illustrating the relationship between the subclones within the tumor.

Reconstruction of phylogenetic trees

In order to generate the phylogenetic trees, the genetic distances between all the subclones must be computed. Here, the Hamming distance was used to assess a distance matrix displaying the genetic distance between each of the subclones. It computes the distance between two vectors by adding all positions in which they differ from one another. Note that it does not consider the magnitude of the entity in the bins compared, which does not matter in this case since they are binary, hence the branch lengths in the phylogenetic trees will be in units of number of aberrations. No bias in the estimation of branch lengths were seen for ML in this study.

Using the function *stem,* a column is added in which the tree will be rooted. The stem can be chosen as a cell containing all mutations shared between the subclones or a cell containing no alterations thus representing a normal cell. The event matrix is then transformed into phyDat format using the function *phydatevent*. This is the data class needed for phylogenetic analysis using the R package phangorn [34].

In the next step the maximum likelihood and maximum parsimony algorithms were used in order to reconstruct phylogenetic trees based on the event matrices. The maximum likelihood trees were

458 reconstructed using the pml algorithm in the package phangorn[34]. First a Hamming distance matrix was
459 calculated from the event matrix which was used to obtain an initial tree given by the neighbor joining
460 method. The initial tree as well as the initial event matrix was used as input variables in the pml
461 algorithm. This function returns an object containing the tree parameters, the data as well as the
462 likelihood for that phylogenetic tree. In order to optimize the tree parameters further, the function
463 optim.pml was used in combination with the Jukes Cantor model. The tree was subsequently rooted in
464 a constructed cell having all the events shared between all subclones. Since the model used is time
465 reversible the choice of the root does not influence the computed likelihood[35]. The tree was visualized
466 using the ggtree package in R[36].

467 The maximum parsimony trees were constructed using the parsimony ratchet algorithm (pratchet) in the
468 R package phangorn with the Fitch algorithm[37]. Using the acctran algorithm the branch lengths and the
469 ancestral character probability distributions were obtained. The trees were rooted in a cell containing no
470 alterations.

471

472 **Performance testing using simulated bulk sampling data**

473 In order to further assess the reliability of the algorithm, it was evaluated using simulated bulk
474 genotyping data using a basic 3D-lattice based, stochastic model of tumor growth. The simulation is
475 initiated imagining a cell having one single genetic aberration, representing a stem event, which will be
476 conveyed to all cells comprising the virtual tumor. In each time unit one cell can proliferate. When a
477 cell has been chosen for proliferation a certain inherent mutation frequency determines whether the cell
478 will mutate or not. If not, two cells identical to the mother cell are obtained, otherwise a stochastic
479 genetic copy number aberration algorithm is used to randomly select a genetic alteration (copy number
480 alteration), conjointly considering the chromosome boundaries and sizes. Making use of a random
481 number generator, a chromosome is randomly selected, then a start and end position and finally the type
482 of event. The result is one cell identical to the mother and one cell harboring one additional genetic
483 aberration. The spatial orientation of the cells is also considered where each cell is assumed to have 26

neighbors and two cells cannot occupy the same position. The position of the second daughter cell is randomly selected among the available neighbor positions. The simulation was conducted generating 40,000 cells giving a simple 3D lattice structure illustrating the spatial intratumoral heterogeneity for three different mutation frequencies (**Figure 4a**). Note that only the fractions of cells harboring a certain genetic alteration is of importance in this model and not any absolute numbers. The goal is not to correctly simulate tumor growth but to obtain a random mixture of entities to be demixed using DEVOLUTION. The mutation frequency chosen, 10/40,000, 50/40,000 and 100/40,000, is not biologically accurate but chosen such that a certain number of mutations will appear during the simulation, resulting in segment files resembling the MCF-distributions seen in the patient cases.

Virtual biopsies were drawn from the set of simulated entities while varying the number of biopsies from 1 to 10. The biopsies were drawn randomly from different parts of the simulated entities. Values for x, y and the z coordinates were randomly chosen while fulfilling

$$x \in [x_{min}, x_{max}], y \in [y_{min}, y_{max}], z \in [z_{min}, z_{max}]$$

$$\sqrt{x^2 + y^2 + z^2} \in [\frac{r_{mean}}{2}, r_{mean}] \in \mathbb{R}$$

where

$$r_{mean} = \frac{x_{max} + x_{min} + y_{max} + y_{min} + z_{max} + z_{min}}{6}$$

Based on the position chosen, the entities within a radius of 2 units were extracted and the fraction of cells harboring each of the alterations found was calculated. The data from each cell can be used to create artificial bulk sampling with MCFs as well as single cell data. Hence, the segment files can be analyzed while having the true subclones at hand for comparison. DEVOLUTION was evaluated using 1-10 biopsies for the three tumors.

Statistics and reproducibility

For phylogenetic tree characteristics, significance was tested using the Mann-Whitney U-test (two-tailed). The stem lengths, total branch lengths and number of branches between individual patients are assumed to be independent of each other and not normal distributed.

All simulation points were repeated at least three times and the standard deviation are illustrated with bars in Figure 4 c and d.

Data availability

All data generated or analyzed during this study are included in this article as supplementary files 1-4.

Code availability

The code is freely available and relies on R 4.0.2 or later. Setup instructions and dependencies can be found on github.

https://github.com/NatalieKAndersson/DEVOLUTION

Author contributions

N.A. and D.G. conceived and designed the project, N.A., D.G. and S.C. developed the methodology, A.V., J.K. and D.G. did tumor biopsy data acquisition, N.A. and D.G. analyzed and interpreted the data, N.A., D.G., S.C., A.V. and J.K. contributed towards the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

We would like to thank the Swegene Centre for Integrative Biology at Lund University (SCIBLU) for assistance.

References

1 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28, doi:10.1126/science.959840 (1976).

2 Quail, D. F. & Joyce, J. A. Microenvironmental regulation of tumor progression and metastasis. *Nat Med* **19**, 1423-1437, doi:10.1038/nm.3394 (2013).

3 Komarova, N. L., Burger, J. A. & Wodarz, D. Evolution of ibrutinib resistance in chronic lymphocytic leukemia (CLL). *Proc Natl Acad Sci U S A* **111**, 13906-13911, doi:10.1073/pnas.1409362111 (2014).

4 Misale, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486**, 532-536, doi:10.1038/nature11156 (2012).

5 Leder, K. *et al.* Fitness conferred by BCR-ABL kinase domain mutations determines the risk of pre-existing resistance in chronic myeloid leukemia. *PloS one* **6**, e27682, doi:10.1371/journal.pone.0027682 (2011).

6 Cresswell, G. D. *et al.* Intra-Tumor Genetic Heterogeneity in Wilms Tumor: Clonal Evolution and Clinical Implications. *EBioMedicine* **9**, 120-129, doi:10.1016/j.ebiom.2016.05.029 (2016).

7 Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* **22**, 105-113, doi:10.1038/nm.3984 (2016).

8 Mengelbier, L. H. *et al.* Intratumoral genome diversity parallels progression and predicts outcome in pediatric cancer. *Nat Commun* **6**, 6125, doi:10.1038/ncomms7125 (2015).

9 Martelotto, L. G., Ng, C. K. Y., Piscuoglio, S., Weigelt, B. & Reis-Filho, J. S. Breast cancer intra-tumor heterogeneity. *Breast Cancer Research : BCR* **16**, 210, doi:10.1186/bcr3658 (2014).

10 Villamon, E. *et al.* Genetic instability and intratumoral heterogeneity in neuroblastoma with MYCN amplification plus 11q deletion. *PloS one* **8**, e53740, doi:10.1371/journal.pone.0053740 (2013).

11 Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat Genet* **48**, 238-244, doi:10.1038/ng.3489 (2016).

12 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

13 Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat Genet* **47**, 209-216, doi:10.1038/ng.3214 (2015).

14 Andersson, N. *et al.* Extensive Clonal Branching Shapes the Evolutionary History of High-Risk Pediatric Cancers. *Cancer research* **80**, 1512-1523, doi:10.1158/0008-5472.CAN-19-3468 (2020).

15 Alves, J. M., Prieto, T. & Posada, D. Multiregional Tumor Trees Are Not Phylogenies. *Trends Cancer* **3**, 546-550, doi:10.1016/j.trecan.2017.06.004 (2017).

16 Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome biology* **21**, 208, doi:10.1186/s13059-020-02119-8 (2020).

17 Karlsson, J. *et al.* Four evolutionary trajectories underlie genetic intratumoral variation in childhood cancer. *Nat Genet* **50**, 944-950, doi:10.1038/s41588-018-0131-y (2018).

18 Deshwar, A. G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology* **16**, 35, doi:10.1186/s13059-015-0602-8 (2015).

19 Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-398, doi:10.1038/nmeth.2883 (2014).

20 Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* **10**, e1003665, doi:10.1371/journal.pcbi.1003665 (2014).

21 Strino, F., Parisi, F., Micsinai, M. & Kluger, Y. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res* **41**, e165, doi:10.1093/nar/gkt641 (2013).

22 Ben-David, U. & Amon, A. Context is everything: aneuploidy in cancer. *Nat Rev Genet* **21**, 44-62, doi:10.1038/s41576-019-0171-x (2020).

23 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

24 Heim, S. & Mitelman, F. *Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells*. (John Wiley & Sons, 2015).

25 Hamming, R. W. Error detecting and error correcting codes. *Bell Labs Technical Journal* **29**, 147-160 (1950).

26 Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881-1893, doi:10.1101/gr.180281.114 (2014).

27 Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome biology* **14**, R80, doi:10.1186/gb-2013-14-7-r80 (2013).

28 Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome biology* **16**, 91, doi:10.1186/s13059-015-0647-8 (2015).

29 Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q. & Karchin, R. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. *PLoS Comput Biol* **11**, e1004416, doi:10.1371/journal.pcbi.1004416 (2015).

30 Zare, H. *et al.* Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol* **10**, e1003703, doi:10.1371/journal.pcbi.1003703 (2014).

31 Maley, C. C. *et al.* Classifying the evolutionary and ecological features of neoplasms. *Nat Rev Cancer* **17**, 605-619, doi:10.1038/nrc.2017.69 (2017).

32 Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat Rev Genet* **20**, 404-416, doi:10.1038/s41576-019-0114-6 (2019).

33 Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (AAAI Press, Portland, Oregon, 1996).

34 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).

35 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-376, doi:10.1007/BF01734359 (1981).

36 Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28-36 (2017).

37 Nixon, K. C. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407-414 (1999).

Supplementary datasets

S. Data 1 Childhood cancer data set. All segment files for NB, WT and RMS along with the clustering obtained by the software. Segments are annotated similar to Supplementary Figure 1.

S. Data 2 The event matrices. The files named "before" illustrates the allocation within each biopsy while the files denoted "after" are the final event matrices.

S. Data 3 Simulation results. The segment file for each simulation, the biopsy positions randomly chosen, which unique cells are present in these biopsies, the event matrix and evaluation of which alterations are correctly allocated using the software.

S. Data 4 Summary of all simulations. Tables illustrating the number of alterations in total, alterations with a clone size > 10 % and the number of alterations correctly allocated.

Figures

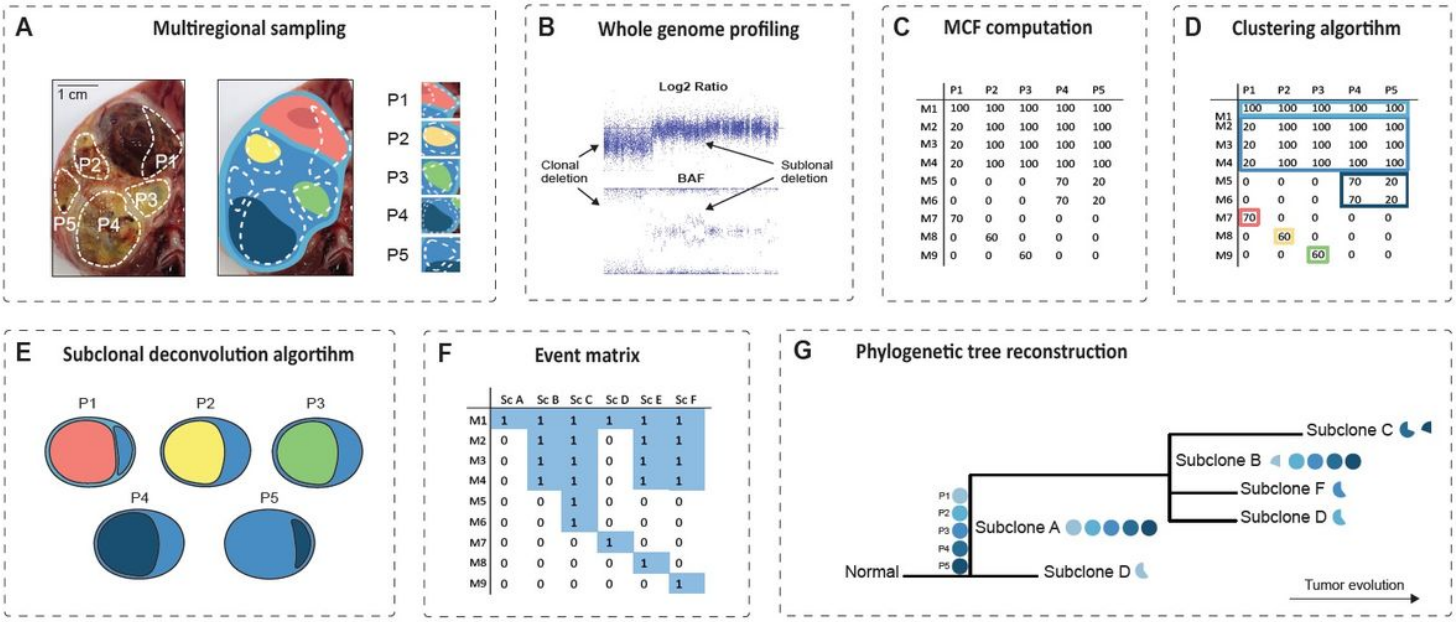

Figure 1

Overview of the methodological outline a) An example of multiregional sampling to obtain biopsies P1-P5 from a Wilms tumor. The tumor is composed of several subclones with distinct genomic profiles, exemplified by the schematic genomic landscape in the rightmost panel where each color unifies cells with identical sets of genetic alterations. The photograph is adapted from a previous publication14 and the colors do not represent true sets of genetic alterations. b) Acquisition of genomic tumor data for each sample, exemplified by copy number analysis by SNP-array. c) The whole genome profiling data can be used to compute the mutated clone fraction (MCF) illustrating the proportion of cells in each biopsy harboring a certain genetic alteration (M1-9 in the left column). d) A clustering algorithm is employed to identify genetic alterations that seem to follow each other in size across samples. e) A subclonal deconvolution algorithm determines the temporal order of these clusters by considering the information obtained throughout all samples while minimizing the occurrence of parallel evolution and back mutations. f) The proposed solution for the temporal order of subclones (Sc) is integrated into an event matrix. g) This event matrix can be used to generate phylogenetic trees with either the maximum likelihood or the parsimony method. At the stem, all available biopsies for the patient are visualized as filled circles with the biopsy name (P1-P5). At the branches of the tree the subclones can be seen along with pie charts illustrating in which biopsies, and in in what fraction of the tumor cells they appear.

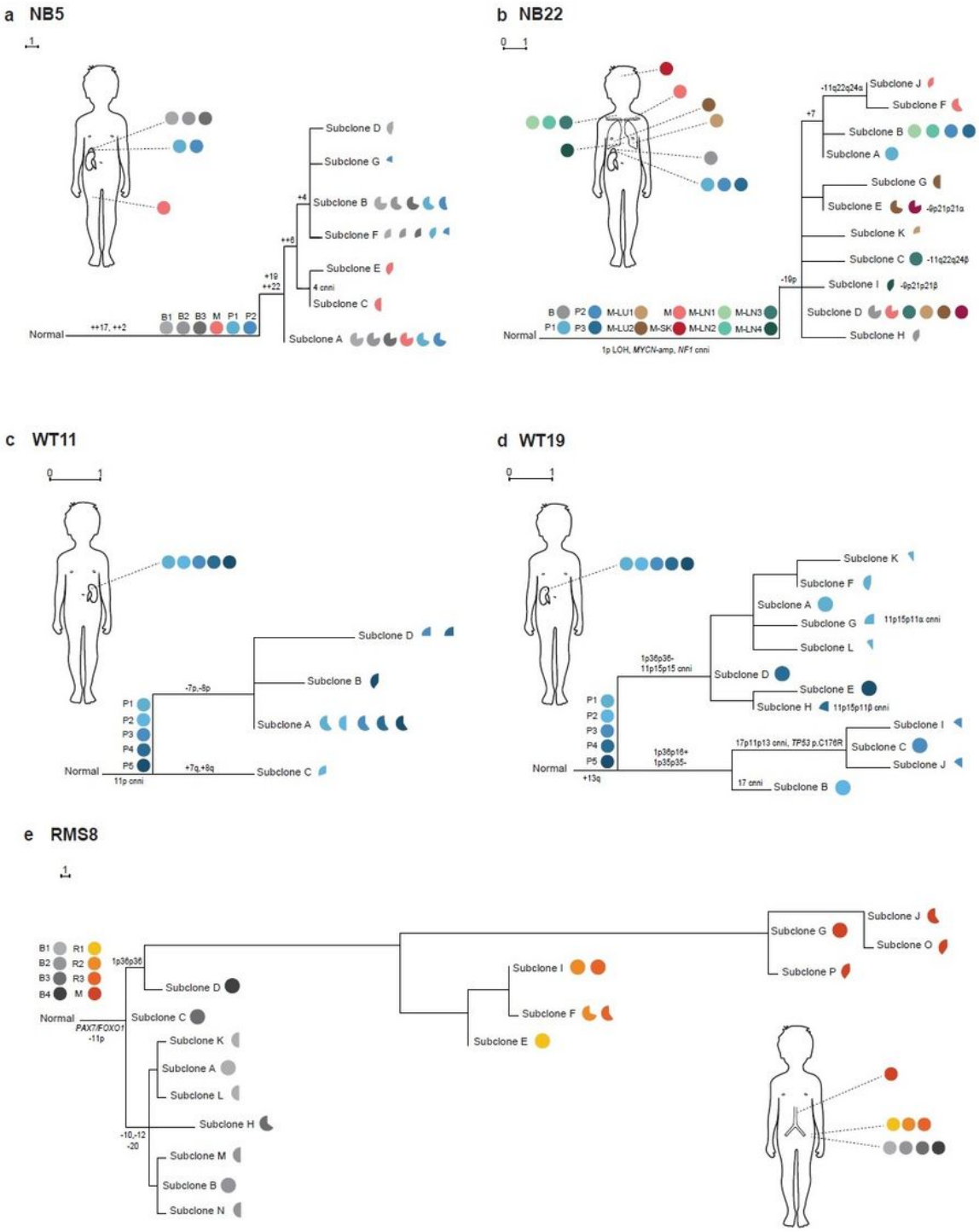

Figure 2

Phylogenetic trees of childhood cancers. At the stem of the maximum parsimony trees illustrated here, the biopsies available from each patient are denoted. The genetic alterations belonging to the stem are present in all cells in all samples, indicated by filled pies. The endpoints represent cell populations harboring distinct genomic profiles (subclones), whose fractions per sample are visualized by pie charts. The scale bar indicates the distance corresponding to one genetic aberration. Gains and losses of

chromosomes or segments of chromosomes that are characteristic of each tumor subtype are indicated by + and – signs. a) In NB5 samples are available from the primary tumor before treatment (B1-B3), a synchronous metastasis (M), and the primary tumor post treatment (P1-P2). The metastasis must have originated from a subclone harboring the stem events only and another subclone with the copy number profile seen in subclone A, indicating polyclonal seebding. The metastasis also has a private copy number neutral imbalance (cnni) of chromosome 4. b) NB22 also shows evidence of polyclonal seeding. Samples are from the primary tumor before treatment (B), the primary tumor post treatment (P1-P3), metastases to the lung (M-LU1-2), to the lymph nodes (M-LN1-4), the skull (M-SK) and from the area around the clavicle (M). The stem harbors a 1p cnni, MYCN-amplification and a NF1 deletion. Greek letters denote different structural alterations targeting partly overlapping regions. c) WT11, shows subclones present across multiple primary tumor areas (P1-P5). d) WT19 display a similar distribution of subclones as WT11 across the post treatment

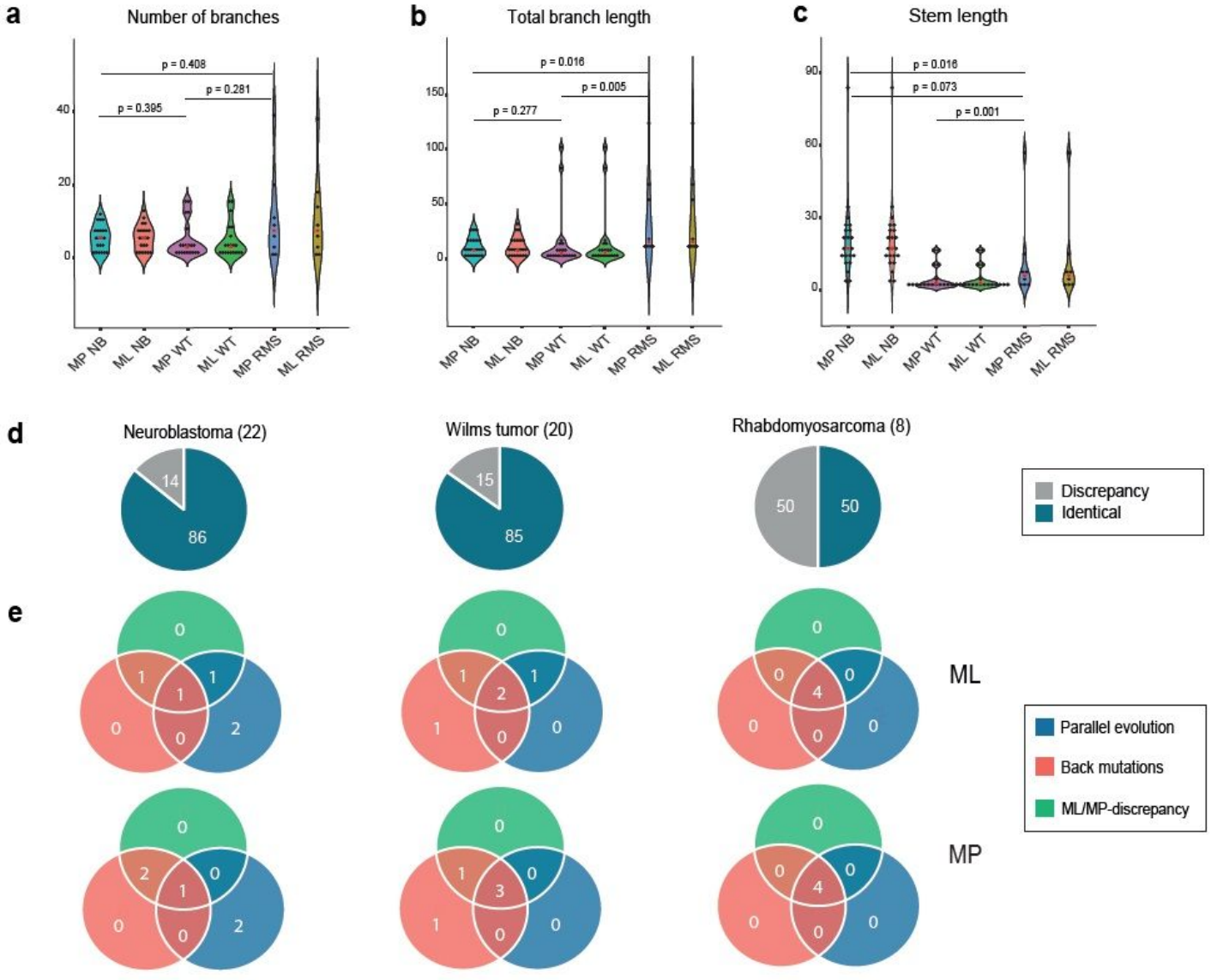

Figure 3

Structural properties of the generated phylogenetic trees. Violin plots of a) the number of branches, b) total branch lengths and c) total stem lengths for neuroblastoma (NB), Wilms tumor (WT) and rhabdomyosarcoma (RMS) using either the maximum likelihood (ML) or parsimony (MP) method for phylogenetic reconstruction. Significance represented by P-vales were calculated using the two-sided Mann-Whitney U-test. d) In 14% of neuroblastomas, 15% of Wilms tumors and 50% of the 8 rhabdomyosarcomas analyzed, the phylogenetic trees obtained using the ML and MP methods differed from one another. NB1, NB24, WT1, WT2, WT3 and WT5 are not included in this calculation since they did not display any private genetic alterations. Hence only event matrices could be generated but not phylogenetic trees. If including these cases, the proportions would be 12.8% for NB and 12.5% for WT. e) Venn diagrams of how often discrepancies between ML/MP trees, back mutations, and parallel evolution occurred together in the same case/tree. Numbers indicate the number of tumors where a particular contradiction or combination of contradictions occurred.

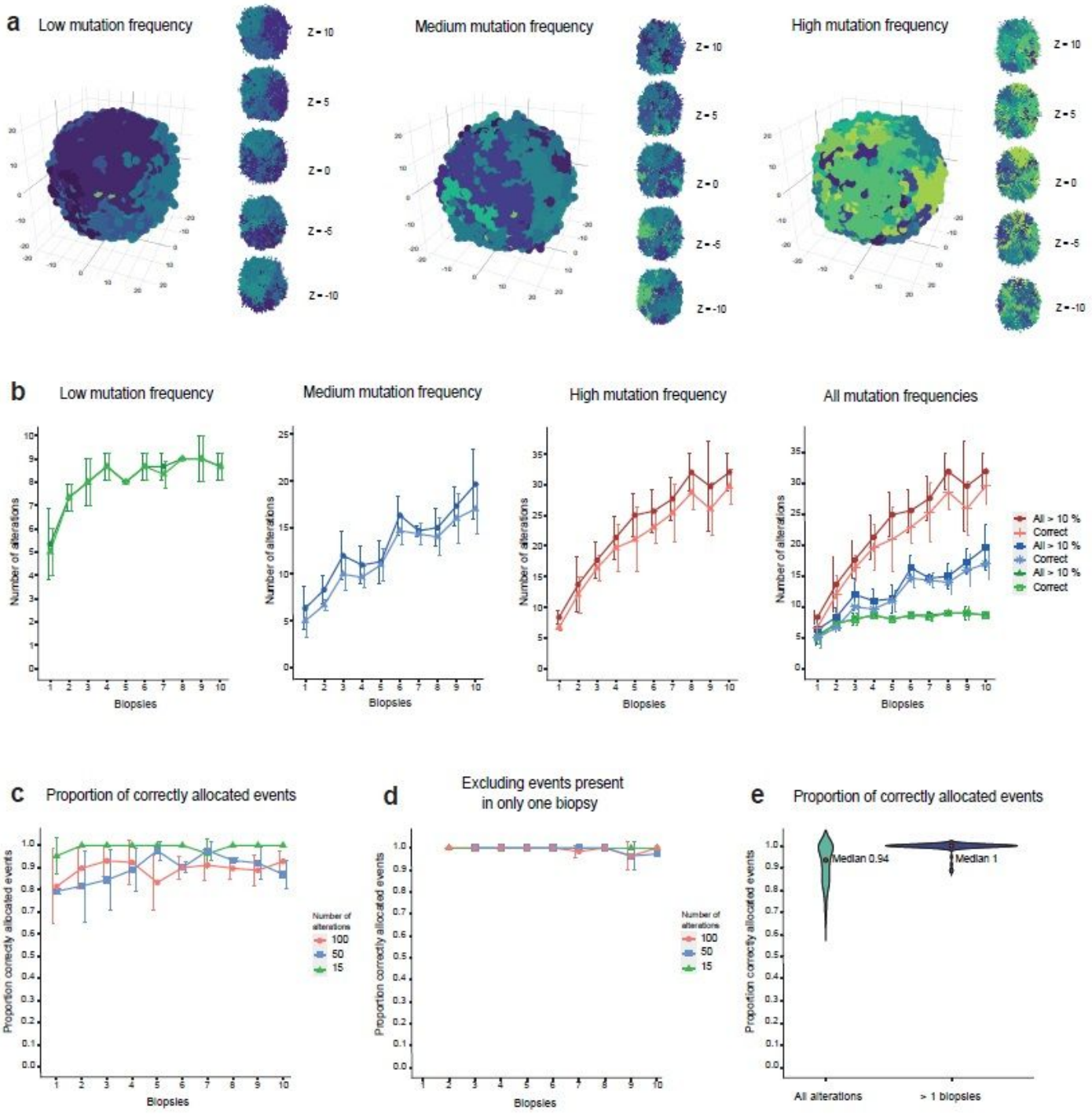

Figure 4

Properties of the simulated bulk genotyping data set and performance measure of DEVOLUTION. a) Visualization of three simulated tumors with increasing mutation frequency. To the right of each tumor are cross-sections at five positions (z). Each color represents a subclone harboring a unique genetic profile. b) When increasing the number of biopsies, more genetic alterations and hence subclones, are identified. In addition, the algorithm is also able to identify more subclones correctly. Each point in the graph is the mean of three consecutive measurements and the error bars consequently the standard deviation. c) The proportion of genetic alterations correctly allocated ± SD when increasing the number of

biopsies from 1-10 for three different mutation frequencies resulting in 10 (green), 50 (blue) and 100 (red) genetic alterations present in the virtual tumor. Each point is represents the mean of three consecutive measurements. d) The proportion of correctly allocated genetic alterations when excluding genetic alterations that were only found in a single biopsy. e) Violin plot showing the spread of the proportion of correctly allocated genetic alterations when including all alterations and when excluding the genetic alterations only found in one biopsy.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S.Figure12.pdf
- S.Figure3.pdf
- S.Figure3text.pdf
- S.Figure4.pdf
- S.Figure4text.pdf
- S.Figure5.pdf
- S.Figure6.pdf
- S.Figure7.pdf