STUDY DESIGNS

The evolution of tumour phylogenetics: principles and practice

Russell Schwartz[1] and Alejandro A. Schäffer[2]

Abstract | Rapid advances in high-throughput sequencing and a growing realization of the importance of evolutionary theory to cancer genomics have led to a proliferation of phylogenetic studies of tumour progression. These studies have yielded not only new insights but also a plethora of experimental approaches, sometimes reaching conflicting or poorly supported conclusions. Here, we consider this body of work in light of the key computational principles underpinning phylogenetic inference, with the goal of providing practical guidance on the design and analysis of scientifically rigorous tumour phylogeny studies. We survey the range of methods and tools available to the researcher, their key applications, and the various unsolved problems, closing with a perspective on the prospects and broader implications of this field.

Selection
An evolutionary process in which one population (or subclone, in the context of cancer) is favoured for growth or survival over another.

Cancer progression
A change of cancer from a less serious to a more serious state, typically in a manner recognizable by pathologists.

Metastasis
A progression in which cancer cells spread to a location in the body that is physically distant from the primary tumour site.

[1]*Department of Biological Sciences and Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15217, USA.*
[2]*Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20892, USA.*

Correspondence to R.S.
russells@andrew.cmu.edu

doi:10.1038/nrg.2016.170
Published online 13 Feb 2017

Cancer is a genetic disease characterized by a progressive accumulation of genomic aberrations that are sometimes augmented by predisposing germline mutations[1]. In the 1970s, Nowell[2] and others proposed that this accumulation of mutations is guided by evolutionary principles via a process of diversification and selection for mutations that promote tumour cell proliferation and survival. The idea that evolutionary mechanisms underlie cancer progression has become a guiding principle in understanding, predicting, and controlling cancer progression[3], metastasis[4], and therapeutic responses[5,6]. Models of tumour evolution have incorporated advanced evolutionary theory[7–9] and complex evolutionary mechanisms that have been revealed by modern genomic technologies[10,11]. The application of evolutionary principles to cancers has blossomed into a field in its own right, with a rich foundation of theory and methods for interpreting tumour evolution[12,13]. Here, we survey one influential thread: the use of phylogenetics — that is, evolutionary tree building — to understand tumour progression.

Although evolutionary theory has proven to be powerful for understanding cancer progression, evolutionary processes are quite different in cancers versus in species[14] in ways that are important to phylogenetic inference. These differences manifest in at least four areas: first, the types of aberration that commonly arise; second, the rates of mutation; third, the extent and intensity of selection; and fourth, the typically high heterogeneity of tumour cell subclones. One frequent feature of cancer evolution is hypermutability[15], often associated with types of mutation that are rare in species evolution. Hypermutability phenotypes include chromosome instability (CIN) phenotypes that are characteristic of p53 dysfunction[16], microsatellite instability (MIN)[17], and elevated point mutation phenotypes, such as those arising from dysregulation of the APOBEC family of deaminase proteins[17,18]. Some variant types, such as copy number variants (CNVs), may be induced by multiple mechanisms — including breakage–fusion–bridge (BFB) cycles, missegregation of chromosomes, and genome doubling — each producing distinct scales and locations of aberrations[19–22]. Other tumour-specific mutational mechanisms include the following: kataegis[23], in which single nucleotide variants (SNVs) occur at a high rate in a small chromosomal region; chromothripsis[24], in which a single chromosome shatters and reassembles in a seemingly random manner; and chromoplexy[25], a complex structural variation characterized by chains of BFB-induced chromosome rearrangements occurring in successive mitoses.

Likewise, patterns of elevated SNV accumulation can differ widely by tissue of origin or from patient to patient. Alexandrov *et al.*[26] characterized dozens of 'mutation signatures' defining the nucleotide biases exhibited in subsets of cancers, some with known environmental triggers[27], others attributable to specific sources of somatic hypermutability[18], and some of unknown cause. Mechanisms of hypermutability may vary by tumour and over time in ways that are not observed in species evolution[21,28–31]. Treatment creates another complication, as chemotherapy or radiation therapy can themselves cause double-strand breaks in the DNA[32] or other forms of hypermutation[33,34], inducing new mutation signatures[26,30]. Conversely, prophylactic therapies can suppress hypermutability[35].

The predominant mechanisms of selection in cancers also differ from those in species evolution. Most studies of tumour evolution have assumed selection for mutations that promote survival, proliferation, or other phenotypic hallmarks of cancer[36]. Selection, like diversification, can be dynamic, as cell populations adapt to or change their microenvironment[11]. However, recent work has suggested that selection often plays only a minor part in tumour evolution, in contrast to its role in Darwinian evolution of species. The repeated observation of substantial intra-tumour heterogeneity[21,37–44] runs counter to the idea that only the fittest subclones survive. Some recent studies have suggested that some tumours evolve by effectively neutral processes without selection, at least pretreatment[45–47]. It has been suggested that strong versus weak selection might be reconciled by a 'punctuated equilibrium' model[9], in which long periods of slow mutation under weak selection are interrupted by short bursts of rapid evolution under strong selection, although this model cannot explain the evidence for a lack of selection in some tumours[48].

Therapy must also be considered when modelling selection[49,50]. In contrast to the disagreement about whether tumour evolution is non-Darwinian at the pretreatment stage, there is general agreement that treatment leads to selection which can alter the dominant clones[10,14,34,51]. Single-agent treatment can lead to relapse[49,52] by selecting for non-responsive clones[29,53]. Durable targeted therapies may require the identification of driver mutations in all tumour subclones and the design of patient-specific drug combinations[8,11,54,55].

High heterogeneity is another characteristic feature of tumour evolution. Higher intra-tumour heterogeneity has been associated with poorer prognosis[8,56–58] and linked with the ability of the tumour to resist immune surveillance and therapy[3,59,60]. Progression, metastasis, and therapeutic resistance frequently proceed from clones that were rare at earlier progression stages[41,43,49,61]. Interactions among distinct clones may also drive tumour progression, for example through tumour self-seeding[4,62] and cooperation between clones[63,64].

This Review examines one important direction in which evolutionary models are shaping cancer research: the use of phylogenetic methods in interpreting genomic data from cancers. We specifically seek to provide guidance to the users of phylogenetic methods in cancer research and to those critically reading about those uses, especially those lacking formal training in phylogenetics. To accomplish that, we give a short overview of the field, we review past uses of tumour phylogenetics, and we explain some relevant principles of phylogenetic inference. We conclude with speculation about the challenges and opportunities for realizing the potential of phylogenetics in cancer research.

Overview of tumour phylogenetics

The recognition that cancer is an evolutionary phenomenon led to the insight that computational methods for reconstructing evolutionary processes — that is, phylogenetics — might prove valuable for making sense of tumour progression processes. Tsao *et al.* were among the first to suggest that variations in microsatellite markers could be used to infer a tree model of the evolution of tumour cells[65]. The idea was subsequently put into practice for bulk comparative genomic hybridization (CGH) data by Desper *et al.*[66]. After percolating for a decade within a specialist community of evolutionary and computational biologists, this type of analysis has exploded to become a new field known as tumour phylogenetics, which aims to reconstruct tumour evolution from genomic variations. In almost all cases, the goal of such work is to produce evolutionary trees, potentially allowing for uncertainty among the space of possible trees explaining a data set[21,67,68].

Within that basic framework, tumour phylogenetics encompasses diverse methods. This diversity includes various data types, referring both to the basic study design (cross-cohort studies of many tumours, single-patient studies of regional bulk genomic assays, or studies of single-cell variability in single tumours) (FIG. 1) and the type or types of genomic data profiled (initially, pre-sequencing marker types such as large-scale CGH[66] or fluorescence *in situ* hybridization (FISH)[69]; now, predominantly next-generation sequencing (NGS)-derived SNVs[70] or CNVs[71], and sometimes more exotic variant types such as gene expression, DNA methylation, or histone marks[10,14,32,72,73]). The diversity also includes variation by mathematical model; that is, the mathematical representation of the kinds of mutational processes one intends to study. The model may capture both the kind of mutations considered (for example, SNVs versus structural variants (SVs)[20,74]) and basic questions such as whether those mutations are assumed to be under selection[2,7,11,14,17,72,75] or selectively neutral[4,76–79]. Furthermore, this diversity of methods includes variation in the algorithms applied; that is, the computational instructions used to find an optimal tree or trees consistent with both the data and the model. The importance and utility of *in silico* models to study various phenomena in cancer goes far beyond tumour phylogenetics, and other kinds of models have been reviewed elsewhere[12,13]. Many of the papers cited therein take a traditional mathematical modelling approach with emphasis on the mathematics, on simulation studies, on parameter estimation, and on validating the model. As tumour phylogenetics has gained in popularity, phylogenetics now tends to show up as a small part of high-impact studies. These studies are understandably focused on data sets that were derived from human subjects and were expensive and complicated to collect. One of the main messages of this Review is that when mathematical models are used in these studies, the importance of validating the models against simulated and observed data should not be forgotten.

Most studies of tumour phylogenetics to date have adapted standard algorithms that were developed for species phylogenetics (for example, maximum parsimony[21,61], minimum evolution[73], neighbour joining[71,80], UPGMA[21], or various maximum likelihood or Bayesian probabilistic inference methods[81,82]), occasionally comparing multiple standard approaches in

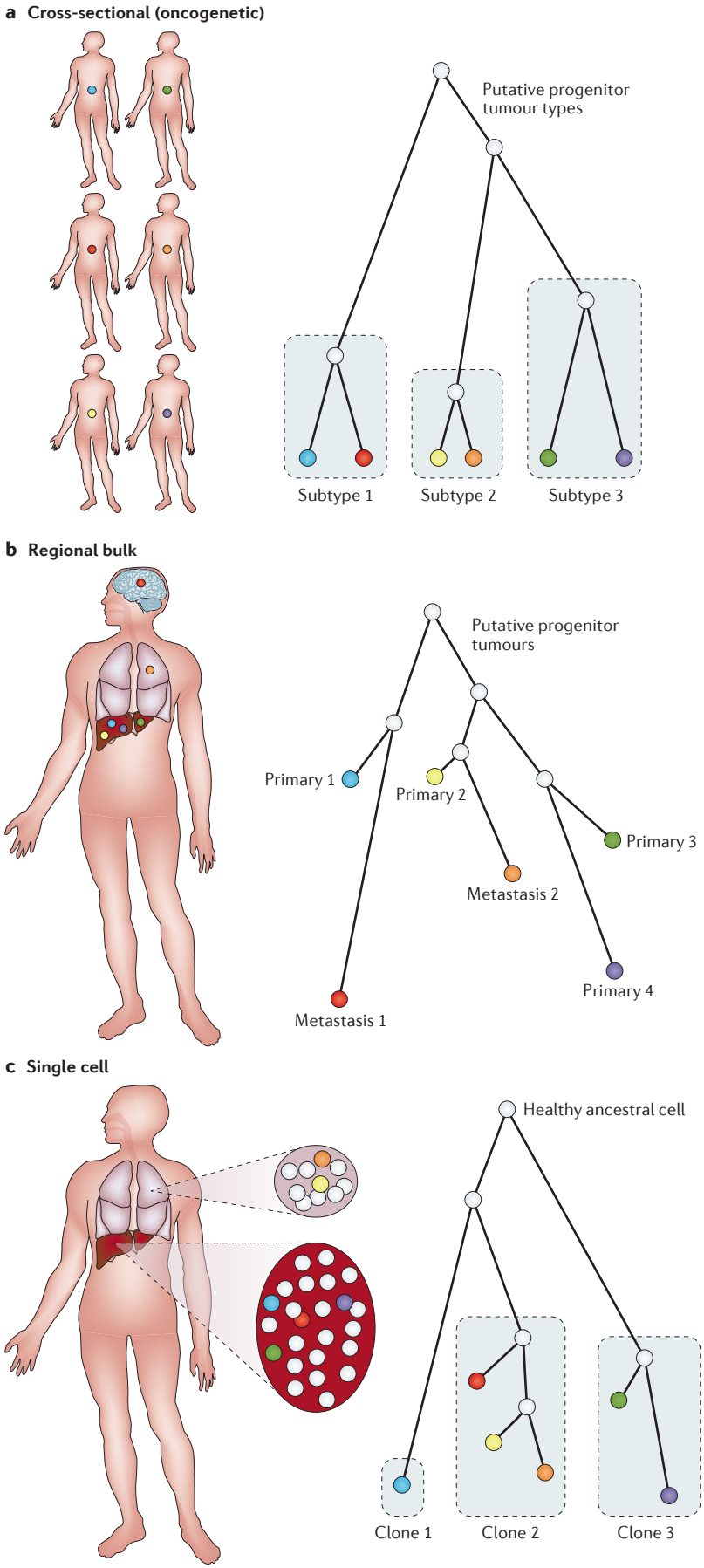
Subclones
Subpopulations of cells in a tumour; the cells in each subclone are almost or completely genetically identical for all measured cancer-related variants.

Hypermutability
An elevated mitotic mutation rate, relative to that in healthy cells; this is often specific to a given mutation type (for example, a single nucleotide variant or a copy number variant).

Intra-tumour heterogeneity
Variation in the genomes of different cells in the same tumour.

Tumour self-seeding
A process by which descendants of cells that escaped the primary tumour re-enter circulation and return to the primary site.

Mathematical model
A formal mathematical abstraction of a physical or biological process, such as a set of evolutionary mechanisms.



a single study[21,83] (TABLES 1,2). Only recently have new phylogeny algorithms emerged to deal with the peculiarities of tumour versus species evolution[84–88]. In the next section, we survey the diversity of methods available, with particular focus on those suited to modern sequencing technologies.

This variety of phylogeny methods has corresponded to a variety of applications. Tumour evolutionary trees, which were once merely conceptual models[2], are now central in the results of many studies[11]. Early uses of phylogeny methods often focused on applying the new tool of tumour phylogenetics to old problems, such as using evidence of evolutionary selection to separate driver mutations from passenger mutations[29,50], or using novel algorithms to find the order and timing of driver mutations[89–91] or to determine how these driver mutations associate with progression stages[92]. Other key results have emerged organically, for example from studies addressing the still controversial question of whether tumour evolution follows the expectations of classical clonal evolution theory[93–95] in producing predominantly linear phylogenies[54,76,96,97], whether it exhibits predominantly branched evolution exemplified by the early divergence of subclones[30,33,40,42,49,73,83,98–100], or whether it occupies some continuum encompassing both extremes in different tumours[34,101]. Researchers continue to find new applications for phylogeny models, such as the use of phylogenies prognostically to predict the likely future progression of a tumour[43,58,85,92,102]; such applications are an evolution of older approaches that have been used to predict progression from simpler measures of tumour heterogeneity[38,58,59,102–105].

One worrisome trend among these studies is their seemingly conflicting conclusions about the evolutionary trajectories of cancers, such as on the questions of linear versus branched evolution or Darwinian selection versus no selection. The distinctions may be traced to differences in the application of phylogenetics, such as looking at distinct marker types (for example, SNVs versus CNVs) or using distinct evolutionary models or phylogeny algorithms. For example, the studies that

Figure 1 | **Classification of tumour phylogeny methods by study design. a** | Cross-sectional tumour phylogeny methods model distinct tumours (coloured circles) sampled from multiple patients as though they are species. These methods infer phylogenies (also known as oncogenetic trees) in which tumours are grouped approximately into subtypes, with tree edges corresponding to common recurring mutations that identify a subtype. **b** | Regional bulk tumour phylogeny methods are applied to bulk genomic samples from a single patient, typically subregions of a tumour or distinct tumour sites (coloured circles). Trees provide a coarse model of the major cell lineages developing over the course of progression in the single patient. **c** | Single-cell tumour phylogeny methods build phylogenetic trees using variations between single cells (coloured circles) in one or more tumour sites. Trees group cells into major clonal subgroups and infer shared ancestry and mutation events at the level of single clones.

Table 1 | **Software tools available for tumour phylogenetics**

Tool*	Data type	Model type	Algorithm type	Refs
Cross-sectional data				
TO-DAG	Bulk, presence/absence of any aberrations	Probabilistic	Combinatorial optimization	91
ct-cbn	CNVs	Probabilistic (non-standard)	Specialized	113
NAM	CNVs	Probabilistic (non-standard)	Maximum likelihood (EM)	112
N/A	CNVs	Distance-based (various)	Several off-the-shelf	125
N/A	CNVs	Maximum parsimony	Combinatorial optimization	177
RESIC (generalized from genes to pathways)	DNAseq-based SNVs and CNVs, and gene expression	Probabilistic (pathway generalization of RESIC)	Specialized (simulation-based)	118
RESIC	DNAseq-based SNVs and CNVs or gene expression	Probabilistic	Specialized (simulation-based)	117‡,§
N/A	Gain/loss events	Probabilistic (non-standard)	Statistical hypothesis testing and PCA	109
METREX	Gene expression	Distance-based (various)	WLS via Fitch (from Phylip), neighbour joining, and FASTME	123
N/A	Gene expression	Distance-based (WLS minimum evolution)	Fitch (from Phylip)	124
unmix	Gene expression	Distance-based (minimum spanning tree)	Combinatorial optimization (with deconvolution)	139‡,§
Rtreemix	Generalized binary mutation array, cross-sectional	Probabilistic (mixture model)	Maximum likelihood (EM)	115
Mtreemix	Generalized mutation array	Probabilistic (mixture model)	Maximum likelihood (EM)	111§
oncotrees	Large CNVs or cytogenetic breaks	Statistical (non-standard)	Combinatorial optimization	66§
oncotrees	Large CNVs or cytogenetic breaks	Distance-based (non-standard)	Combinatorial optimization	122
DiProg	Large CNVs or cytogenetic breaks	Probabilistic	Combinatorial optimization (ILP)	120
oncomodel	Large CNVs or cytogenetic breaks	Probabilistic	Maximum likelihood	108
N/A	Large CNVs or cytogenetic breaks	Statistical (non-standard)	Custom heuristic optimization	178
BML	Mutational array	Probabilistic	Bayesian sampling (MCMC)	90
CAPRI, TRONCO, and PiCnic	Mutational array	Specialized probabilistic (PiCnic is a general pipeline)	Custom heuristic optimization	121,179, 180
Single-patient, bulk data				
PhyloSub	SNVs	Probabilistic	Bayesian sampling (MCMC) and maximum likelihood (EM)	119§
BitPhylogeny	Methylation, WGS	Probabilistic	Bayesian sampling (MCMC)	86§
GRAFT	DNAseq-based SNVs, CNVs, and rearrangements	Specialized	Combinatorial optimization	134
Single-patient, multiple-site, bulk data				
cITUP	DNAseq-based SNV VAFs	Probabilistic (joint deconvolution and phylogenetics)	Combinatorial optimization (quadratic programming)	144
MEDICC	DNAseq- or CGH-based CNVs	Minimum evolution	Combinatorial optimization	85§
TuMult	CNVs (large-scale)	Maximum parsimony	Combinatorial optimization	129§
Clomial	DNAseq-based SNV VAFs	Probabilistic	Maximum likelihood (EM)	142‡

Table 1 (cont.) | **Software tools available for tumour phylogenetics**

Tool*	Data type	Model type	Algorithm type	Refs
Single-patient, multiple-site, bulk data (cont.)				
PhyloWGS	DNAseq-based SNV and CNV VAFs	Probabilistic	Bayesian sampling (MCMC)	135§
Canopy	DNAseq-based SNV and CNV VAFs	Probabilistic	Bayesian sampling (MCMC)	137
SPRUCE	DNAseq-based SNV and CNV VAFs	Specialized (joint deconvolution and phylogenetics)	Combinatorial enumeration	136
SubcloneSeeker	Any variant with a VAF	Specialized (joint deconvolution and phylogenetics	Combinatorial enumeration	138
AncesTree	SNVs	Weighted parsimony	Combinatorial optimization (ILP)	131
rec-BTP	SNVs	Specialized (joint deconvolution and phylogenetics)	Combinatorial optimization	130
LICHeE	SNVs	Specialized (joint deconvolution and phylogenetics)	Combinatorial optimization	132
SCHISM	Output of a clone prediction program such as PyClone or SciClone	Probabilistic	Maximum likelihood (genetic algorithm)	143
Single-patient, single-cell data				
N/A	FISH	Probabilistic	Maximum likelihood (EM)	181
N/A	FISH	Probabilistic	Maximum likelihood (EM)	69
N/A	FISH	Weighted maximum parsimony (with constraint satisfaction)	Combinatorial optimization (ILP)	169
FISHtrees	FISH	Maximum parsimony (with several different formulations of the optimization problem)	Combinatorial optimization	84§,149§, 151§,152§
N/A	FISH	Maximum parsimony (rectilinear)	Combinatorial optimization	153
N/A	qPCR and FISH	Maximum parsimony	Combinatorial optimization (PAUP)	182
OncoNEM	scSeq-based SNVs	Probabilistic	Maximum likelihood (specialized heuristic)	154§
SCITE	scSeq-based SNVs	Probabilistic	Bayesian sampling (MCMC)	87
muttree	SNVs	Probabilistic	Maximum likelihood (specialized optimization)	89

CGH, comparative genomic hybridization; CNV, copy number variant; DNAseq, bulk DNA sequencing; EM, expectation maximization; FISH, fluorescence *in situ* hybridization; ILP, integer linear programming; MCMC, Markov chain Monte Carlo; MST, minimum spanning tree; N/A, not applicable; PCA, principal components analysis; qPCR, quantitative PCR; scSeq, single-cell sequencing; SNV, single nucleotide variant; VAF, variant allele frequency; WGS, whole-genome sequencing; WLS, weighted least squares. *Additional related tools, including tools that identify subclones by deconvolution, are listed in Supplementary information S1 (table), which also contains more information, including the URLs, for the tools listed here. For consistency with the text, the order of tools is sorted primarily by study type and secondarily by data type. Supplementary information S1 (table) is provided in Excel and includes an explicit study type column to allow the reader to sort the rows in the same way or in other ways. ‡These studies have some phylogenetic aspects, but do not produce phylogenies as their primary output. §These studies use some of the more important or innovative software packages.

concluded that there was little selection in some tumours looked mostly at SNVs and CNVs, but perhaps there is selection in those tumours via evolutionary mechanisms that would be apparent only when looking at other marker types, such as karyotypes or methylation patterns. Few studies have tested whether the phylogenetic inferences made are robust to a change of methods, with notable exceptions[68,106].

Variations on tumour phylogenetics

Recent years have seen a rapid proliferation of methods for tumour phylogenetics. In this section, we categorize some of the seminal advances made. We can roughly distinguish three classes of method, based on the kind of phylogeny study for which they are designed: cross-sectional methods, which use data on many tumours to build trees describing the common progression pathways

Table 2 | **Case studies using tumour phylogenetics**

Data type*	Method	Refs
WGS	Bayesian	183
WES and FISH cytogenetics	Bayesian	101
Binary SGAs	BEAST and PAUP	35‡
WES	Broad Institute custom heuristic (parsimony, branch sibling model, and grafting at tips)	98‡
WES	Broad Institute custom heuristic (parsimony, branch sibling model, and grafting at tips)	147
WES-based SNVs and SNP-based CNVs	Broad Institute custom heuristic (parsimony, branch sibling model, and grafting at tips)	22
DNAseq	REFS 23,53	67‡
WGS and targeted deep sequencing	Custom heuristic in three stages	50
WES	Custom heuristic in three stages	148
Ultra-deep multi-region DNAseq-based SNVs	Custom heuristic in three stages	28
FISH	FISHtrees	41
FISH	FISHtrees	102
FISH	FISHtrees	58
DNAseq, methylation, and CNVs	Minimum evolution and third-party tool for some CNVs	73‡
CNVs	MEDICC	43‡
DNAseq, aCGH, and FISH	MEDICC	47‡
SNP-based CNVs, and LOH	MEDICC	184
Cross-sectional DNAseq	Mtreemix	110
Karyotyping	Mtreemix and REF. 114	92
aCGH	Neighbour joining	56
Regional aCGH	Neighbour joining	145
scSeq-based CNVs	Neighbour joining	71‡
WGA scSeq	Neighbour joining	156
scSeq	Neighbour joining	80
scSeq	Neighbour joining	185
CNVs, RNA expression, and methylation	Neighbour joining and RESIC	186
Single-cell microsatellite	Neighbour joining using L1 distance between alleles	78‡
Targeted methylation	Neighbour joining using methylation Hamming distance and ABC parameter inference	79‡
Targeted methylation	Neighbour joining using methylation Hamming distance and ABC parameter inference	187
Targeted DNAseq	Neighbour joining from ape with clones from PyVCF	100
Single-cell microsatellite	Neighbour joining with Manhattan distance	76
Separate SNVs and CNVs with blood normal	Neighbour joining, maximum likelihood, and ultrametric	83‡
20 poly(G) tracts, regional	Neighbour joining with bootstrapping, and MrBayes	106
Deep DNAseq	Not documented	54
CNVs	Not documented	188
WES-based SNVs and CNVs	Not documented	64

across a population; regional bulk methods, which build trees for single patients through bulk genomic assays of distinct tumour sites or regions; and single-cell methods, which build trees from the cell-to-cell variations in single tumours (FIG. 1). Not all methods fit neatly within one category, but the categories provide a crude organization for the description of methods below.

Within these high-level categories, we see a diversity of genomic data types (TABLE 3), evolutionary models, and phylogeny algorithms. Below, we consider a subset of methods that were of particular importance in introducing new techniques to the field or were of unique value to likely users. TABLE 1 and the extended version, Supplementary information S1 (table), provide a more comprehensive list of important methods. TABLE 2 and the extended version, Supplementary information S2 (table), list important studies that have made use of tumour phylogeny methods.

Cross-sectional tumour phylogenetics. Key ideas behind cross-sectional tumour phylogenetics originate in the pre-phylogenetic work of Fearon and Vogelstein, who proposed that bulk analysis of collections of tumours from multiple patients could allow one to infer the likely orders of aberrations and stages of progression (for example, from adenoma to carcinoma) so that each aberration is associated with progression to a specific stage[93]. They proposed a linear (event 2 follows event 1 follows event 0) model for the progression of colorectal cancer. This Fearon–Vogelstein model, although a simplification[107], has been highly influential on thinking about tumour evolution. Phylogenetic methods were first brought to the reconstruction of tumour progression pathways by Desper *et al.*, who generalized the Fearon–Vogelstein linear progression model to allow branching in the form of a tree, sometimes called an oncogenetic tree[66]. FIGURE 1a provides an illustration of the oncogenetic tree model for interpreting cross-sectional data that has come from multiple patients. In the original oncogenetic tree model, each tree edge corresponds to a possible aberration with an associated probability of occurrence. Paths in the tree correspond to possible sequences of accumulating aberrations.

Many methods have since applied this basic strategy of inferring trees or graphs of possible progression sequences from combinations of mutations observed across a patient cohort. We refer the reader to a general phylogenetics text[82] for more background on the basic classes of phylogenetic models and algorithms summarized in TABLE 4 and their trade-offs. The original Desper *et al.* method[66] was a character-based phylogeny method, meaning that it modelled evolution from a discrete set of phylogenetic markers (variant loci), and it was specifically a kind of maximum parsimony method, meaning that it was a combinatorial optimization method that sought to explain a data set with the smallest number of distinct mutations possible. Character-based methods tend to be most informative for reconstructing the sequence of mutations and unobserved ancestral states, but they become computationally infeasible on large marker sets. Parsimony methods are the most computationally efficient of the character-based

Table 2 (cont.) | **Case studies using tumour phylogenetics**

Data type*	Method	Refs
Ultra-deep multi-region DNAseq	Not documented	40‡
WES, and genotyping for SNVs	Not documented	46‡
Regional WES-based VAFs, and aneuploidy	Maximum parsimony	146
Microsatellites	Maximum parsimony (Camin–Sokal) from Phylip (using MIX)	172
DNAseq-based SNVs, CNVs, and gene-fusion VAFs	Maximum parsimony (manual)	158
WES	Maximum parsimony (Wagner) from Phylip	61
WES	Maximum parsimony (Wagner) from Phylip	99
SNVs and CNVs	Maximum parsimony from Phylip (using B&B)	42
Deep DNAseq-based SNVs and indels	Maximum parsimony from Phylip (using Dnapars)	49
Microsatellites	Maximum parsimony from Phylip (using penny)	96
Regional WES-based SNVs	Maximum parsimony in phangorn (in Bioconductor)	30
scSeq-based CNVs	Maximum parsimony in phangorn (in Bioconductor)	157
WES, with verification by ultra-deep NGS	Maximum parsimony, with a third-party max-mini tool	167
WES	Maximum parsimony, maximum likelihood, and Bayesian	68‡
Targeted deep sequencing	Maximum parsimony from Phylip with bootstrapping	189
WGS and WES	Maximum parsimony and UPGMA using MEGA5	21‡
WES, FISH, and targeted deep sequencing	PyClone and neighbour joining on three FISH loci	70
WGS-based SNVs and CNVs, and a custom targeted method for clonality	PyClone, EXPANDS and BloNJ	53
WES	PyClone and customized maximum likelihood	190
WGS- or array-based CNVs, and scSeq	TITAN and PyClone for bulk data, and MrBayes for scSeq data	191
CNVs (and other data for non-phylogenetic analyses)	TuMult	59‡
WES	TEDGs	34
Microsatellites	Statistical analysis of allele sizes	107‡
X chromosome microsatellites in males	Statistical analysis of allele sizes	165

See Supplementary information S2 (table) for more information about each of these studies, and also for some comparable studies that did not use tumour phylogenetics but addressed similar problems by non-phylogenetic methods. ABC, approximate Bayesian computation; aCGH, array comparative genomic hybridization; CNV, copy number variant; FISH, fluorescence *in situ* hybridization; DNAseq, bulk DNA sequencing; indel, small insertions or deletions; LOH, loss of heterozygosity; NGS, next-generation sequencing (typically a mixture of WES and WGS); scSeq, single-cell sequencing; SGAs, somatic genetic abnormalities; SNV, single nucleotide variant; TEDGs, tumor evolutionary directed graphs; VAFs, variant allele frequencies; WES, whole-exome sequencing; WGA, whole-genome amplification; WGS, whole-genome sequencing. *DNA samples are bulk samples unless they are explicitly noted as coming from single cells. ‡These studies are, in our opinion, the most important, innovative or controversial.

methods, but they depend on the assumption that mutations are rather rare, which is a questionable assumption for tumours. The field later moved largely towards more sophisticated probabilistic character-based methods[108,109], which seek either the most probable tree (the maximum likelihood method) or some measure of the space of possible trees and tree parameters (Bayesian sampling). Compared with the earlier approaches, such models better handle high mutation rates, noisy data, and uncertainty in tree inferences, but can be more computationally demanding than parsimony methods. Beerenwinkel *et al.*[110,111] introduced an important class of probabilistic model that enables the joint inference of several possible trees for binary mutation data, via the Mtreemix tool, an approach that was later generalized to CNV data[112,113] and became the basis of the newer Rtreemix package[114–116]. More recent approaches include making better use of the detailed information specifically offered by DNA sequencing (for example, as in the RESIC[117] approach and a later pathway-level variant[118]). Algorithmically, most such methods rely on comparatively faster maximum likelihood techniques[113]. However, more advanced Bayesian models commonly use variants of Markov chain Monte Carlo (MCMC) sampling, which is a statistical technique for exploring the ranges of possible tree models and evolutionary parameters but at a much greater computational cost than maximum likelihood methods[90,119]. The recurring theme of trade-offs between more realistic and more computationally tractable models has inspired a great deal of research into more exotic algorithmic techniques in this domain[91,120,121].

The major alternative to character-based methods are distance-based methods, which use mutation data to estimate evolutionary distances between samples, and these distances then serve as the basis for tree inference. Such methods can handle much larger marker sets at the cost of losing the fine-scale modelling of mutational events achieved by character-based methods. Desper *et al.* extended their approach to distance-based methods[122] and later extended those from DNA to RNA expression data[123]. Riester *et al.*[124] developed a similar approach specifically for RNA sequencing data using minimum evolution phylogenies, which is a distance-based analogue of parsimony methods. Liu *et al.*[125] applied cross-sectional distance-based methods to CNVs using several off-the-shelf distance-based phylogeny tools.

Oncogenetic tree methods in recent years have been primarily used to analyse DNA sequencing-derived SNV or CNV data[43,70,85], but they have also been used for methylation data[73,79]. They have proven to be valuable primarily for the original purpose of identifying combinations and orders of recurring driver mutations. In hindsight, the cross-sectional tumour phylogeny methods are domain-specific clustering methods that happen to use phylogenetics tools on the assumption that distinct tumours can share common evolutionary trajectories. However, this was not clear until sequencing studies revealed both inter-tumour and intra-tumour heterogeneity, and this finding is part of the 'evolution' of tumour phylogenetics alluded to in the title of this Review.

Given the diversity of methods available, though, one should be aware that simulation studies[126,127] have shown that qualitative results may depend considerably on the model used to generate the data. Furthermore, most methods for cross-sectional data were developed before

the extent of intra-tumour heterogeneity in observed data was appreciated[37], and tree inferences from cross-sectional data can be unreliable in the presence of intra-tumour heterogeneity[128]. These latter observations help motivate the trend towards phylogenetic studies of single tumours, discussed below.

Regional bulk tumour phylogenetics. A major step forwards was the recognition that one could produce phylogenies for single patients, initially through sampling multiple regions or tumour sites. One treats each site sequenced as if it were a species and infers a tree connecting those species. FIGURE 1b provides an illustration of a regional bulk phylogeny built from samples of multiple tumour sites and multiple regions within a tumour site for a single patient. The earliest such tools used data types that predate NGS, such as large-scale CNVs used by TuMult, a parsimony-based combinatorial algorithm[129]. Similar ideas have since been brought to DNA sequencing-derived data types, including SNVs (as used in rec-BTP[130], AncesTree[131], and LICHeE[132]) and CNVs (as used in TITAN[133] and MEDICC[85]). Given the variations in the rates and mechanisms of SNV versus CNV evolution, some methods have found particular power in combining data types, as is done by GRAFT[134], PhyloWGS[135], SPRUCE[136], and Canopy[137]. The available methods also cover a range of models and algorithmic techniques, including various combinatorial (parsimony-like) character-based methods[130,131,134,138], probabilistic character-based methods[133,135], and distance-based minimum evolution[85].

An important variation on regional bulk tumour phylogenetics is the combination of phylogenetics with clonal deconvolution from bulk sequence[139]. Here, deconvolution means the inference of clonal subpopulations from one or more bulk genomic samples. Numerous tools are now available for clonal deconvolution (for example, SciClone[140], PyClone[141], and Clomial[142]). Some tumour phylogeny methods listed in TABLE 1 explicitly depend on clonal deconvolution either as a preprocessing step or integrated into the phylogenetic inference strategy. These include some early approaches to deconvolution that were motivated explicitly by the application to tumour phylogenetics[139], tumour phylogeny methods such as SCHISM[143], which require a third-party clonal deconvolution program to generate their input data, and tools such as cITUP[144] and LICHeE[132], which fully integrate deconvolution and phylogenetics into a single inference.

Regional bulk phylogenetics has been used in several seminal studies, building on earlier work on multi-region progression without explicit phylogenetics[104]. Early, pre-NGS examples of true multi-region tumour phylogenetics include the use of microsatellite markers by Khalique *et al.*[96] and of array CGH (aCGH) by Navin *et al.*[145]. Regional bulk tumour phylogeny methods using DNA sequencing-derived markers have since seen extensive use (TABLE 2). Many studies that apply regional bulk phylogenetics approaches have relied on standard methods or phylogeny programs derived from species evolution[42,61,73,79,146]. Others have developed custom heuristic phylogeny approaches[28,50,98,147,148] or relied on manual phylogeny-like inferences[33,52]. Only recently have mature third-party tools become available (TABLE 1) and begun to appear in case studies (TABLE 2). Examples include the studies by Schwarz *et al.*[43], which applied the MEDICC software to ovarian tumours to demonstrate that relapse tumours typically show early divergence from the primary tumour, by Sottoriva *et al.*[47], which also used MEDICC applied to colorectal tumours, resulting in the 'big bang model' of evolution

Table 3 | **Experimental technologies and data types for tumour phylogenetics**

Technology and data type	Comments	Refs
Pre-NGS technologies		
Large-scale cytogenetic abnormalities	Convenient before sequencing became ubiquitous, but superseded by more comprehensive genomic studies	66
Microsatellite markers	Rapidly evolving, usually neutral markers	78
FISH	Useful for probing small numbers of CNVs in single cells; largely displaced by scSeq, but still important owing to its practicality for much larger numbers of single cells	69
aCGH	Early high-throughput method for bulk CNV profiling; still in use, although being displaced by DNAseq	145
Expression microarrays	Convenient for high-throughput before RNAseq became widely available; not commonly used for phylogenetics, as it provides only a noisy and indirect measure of genetic evolution	123
SNP chips	Designed initially for genotyping and association studies, but also used in many cancer studies to infer copy number profiles along the genome and to infer CNVs	
Bulk sequence technologies		
DNAseq SNVs	Perhaps the most commonly used marker type, it provides whole-exome or whole-genome profiles of evolution by point mutations	117
DNAseq CNVs	CNVs can be inferred by local changes in sequence coverage, instead of using aCGH or SNP arrays	117
RNAseq expression	More precise and accurate replacement for expression microarrays; nonetheless remains a niche technology for phylogeny studies	124
DNA methylation	Measured by bisulfite sequencing, provides unique information on the evolution of the cell state that is not apparent from conventional DNAseq methods; some methylation markers are neutral, others evolve to select for gene expression	79
scSeq technologies		
DNAseq SNVs	Uniquely powerful method for identifying large numbers of phylogenetic markers at the single-cell level; only recently making inroads as the technology has matured and data quality has improved	87,154
DNAseq CNVs	Perhaps the dominant technology for single-cell tumour phylogenetics, offering coarse-grained profiles of evolution by copy number change; robust to data quality issues in emerging scSeq technologies	71
Single-cell microsatellites	Not a widely used technique but one important to early tumour phylogeny studies; offers important advantages in profiling a putatively selectively neutral marker type	76

aCGH, array comparative genomic hybridization; CNVs, copy number variants; DNAseq, bulk DNA sequencing; FISH, fluorescence *in situ* hybridization; NGS, next-generation sequencing; RNAseq, RNA sequencing; scSeq, single-cell sequencing; SNVs, single nucleotide variants.

Table 4 | **Phylogeny models and algorithms**

Model or algorithm name	Description	Refs
Evolutionary models and objective functions		
Maximum parsimony (MP)	Simplest phylogeny model; assumes that mutations are rare and so the tree with the fewest mutations is the most plausible	96
Minimum evolution (ME)	Distance-based analogue of maximum parsimony; assumes that the tree with the least amount of evolution is the most plausible	85
Probabilistic	Broad class of models well suited to complicated evolutionary scenarios, noisy data, and sampling over unknown evolutionary parameters; generally divided into maximum likelihood (ML), used to find one best-fitting tree for the data and model, and Bayesian, used to identify the space of plausible trees and parameters consistent with the model and data	86, 110, 111, 135, 137
Weighted least-squares (WLS)	Distance-based model defining the most plausible tree as that most closely approximating an input set of distances between taxa by a mean-square measure	121
Phylogeny algorithms and algorithmic techniques		
Combinatorial	Broad class of methods frequently used for character-based phylogenies to optimize over a discrete set of possible topologies; generally the most efficient methods, but suitable only for simpler models; examples include B&B, in which one exhaustively searches a space of all possible solutions while avoiding provably unproductive subspaces, and integer linear programming (ILP) or quadratic programming (QP), in which one converts the problem to a special class of mathematical optimization for which efficient solver programs are available	89, 131, 144
Heuristic search	Broad class of algorithms designed to approximately search a space of trees that are based on empirical effectiveness but are not proven to find the best possible trees; also used when solving for phylogeny models for which efficient, exact methods are unknown; a common generic heuristic is a genetic algorithm (for example, REF. 143), in which one generates a pool of possible solutions and 'evolves' them under a model of mutation and mating; many phylogeny-specific heuristics have also been developed (for example, REF. 82)	121, 143, 176
Neighbour joining (NJ)	Fast method for phylogenetics by successively refining subtrees, approximating a minimum-evolution tree while allowing a possibility of temporally impossible scenarios	76
Unweighted pair group with arithmetic mean (UPGMA)	Method for hierarchically constructing a tree by successively joining subtrees, yielding fast tree reconstruction but being dependent on the molecular clock hypothesis (that variation accumulates at equal rates in all tree branches)	21
Markov chain Monte Carlo (MCMC)	Class of algorithms that is suitable to many forms of probabilistic model and allows one to explore parameter ranges and uncertainty in assignments, but is generally too computationally costly to use on trees of more than a small number of nodes	87

without apparent selection, and by Sottoriva *et al.*[59], which utilized TuMult[129] to help demonstrate the role of intra-tumour heterogeneity in promoting resistance in glioblastomas.

Single-cell tumour phylogenetics. The advance that most raised awareness of tumour phylogenetics among non-computational cancer researchers was its application to single-cell data, allowing the generation of a phylogenetic tree based on individual tumour cells extracted from a single patient (FIG. 1c). Single-cell tumour phylogenetics predates single-cell sequencing (scSeq), as it was applied through various older methods offering more limited profiling of single cells via microsatellite[76] or FISH[69] markers; such approaches remain valuable owing to their ability to examine much larger numbers of cells than scSeq[39,84,149] (TABLE 3). Nevertheless, the introduction of scSeq to tumour phylogenetics by Navin *et al.*[71] deserves much of the credit for bringing tumour phylogenetics into the mainstream of cancer research. Since that work, methods for and applications of scSeq in tumour evolution have proliferated, along with related analyses on the data needs of robust scSeq-based phylogenetic analysis[150].

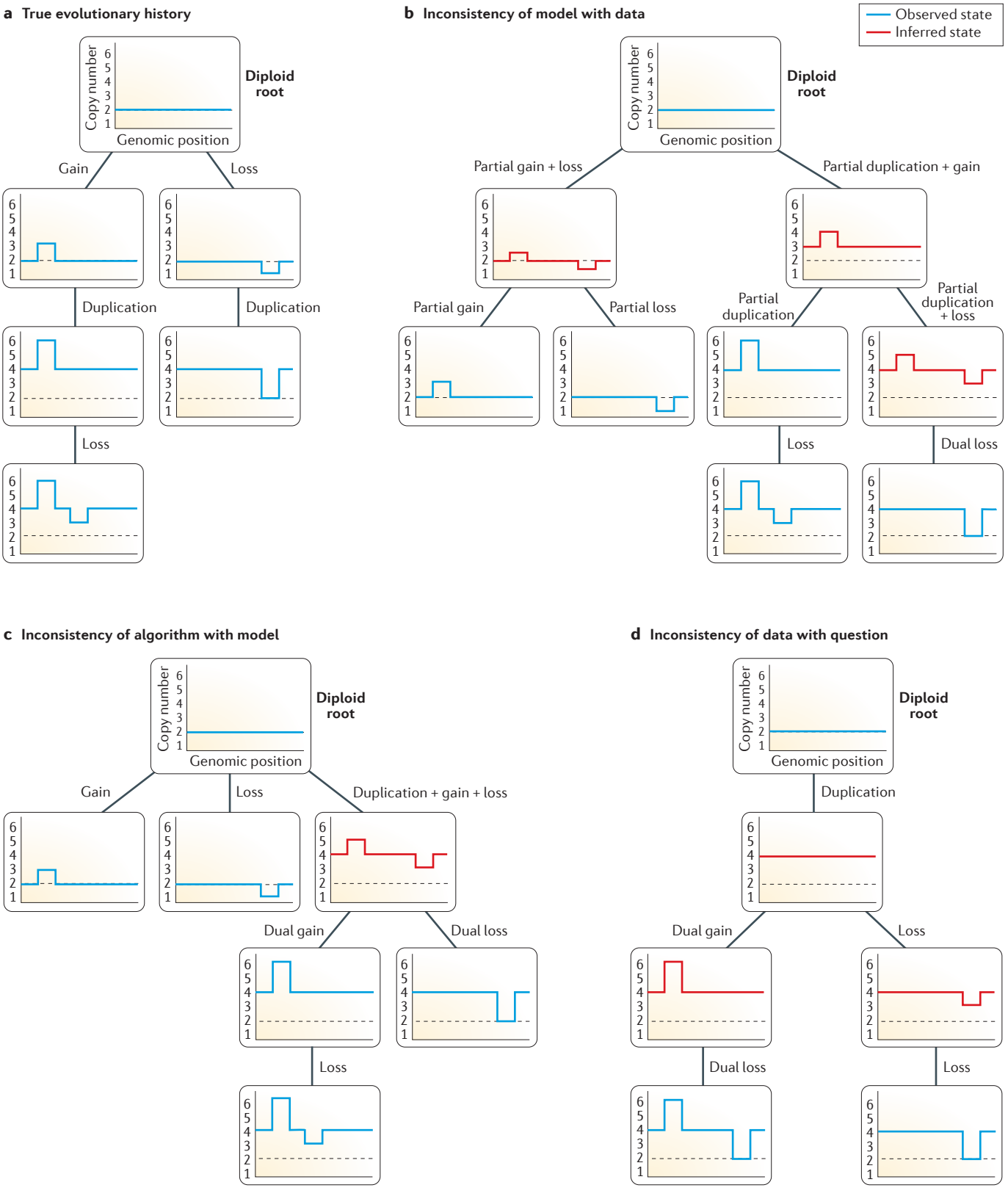
The majority of published tools for single-cell phylogenetics are still based on pre-scSeq technologies[84,148,151–153], with just a handful having been developed specifically for scSeq. Kim and Simon[89] introduced the muttree program, which uses a custom combinatorial inference to find trees optimized for a specialized probabilistic model that differs from the models used by other tools which accept the same input. Ross and Markowetz[154] and Jahn *et al.*[87] developed sophisticated Bayesian probabilistic models for scSeq-derived SNVs, and these models were implemented in OncoNEM and SCITE, respectively.

Most applications of scSeq phylogenetics to date have thus relied on tools for general species phylogenetics or on phylogenies that have been manually constructed without an explicit model or algorithm (for example, see REF. 39). Navin *et al.*[71] relied on neighbour joining[155], which had earlier been used by Frumkin *et al.*[76] with microsatellite data, to infer phylogenies from

scSeq-derived CNVs. Neighbour joining was also used by Xu *et al.*[80] for application to renal cancers and by Wang *et al.*[156] for what was, until recently[157], the largest scSeq study of tumour evolution.

An illustrative tutorial

Data, models, algorithms. As more cancer research groups embrace phylogenetic methods, it becomes important to understand what goes into a phylogenetic



analysis and how to interpret its results. Phylogenetics is a complicated subject for which tools can easily be misused. In this section, we provide guidance to aspiring users of tumour phylogenetics and those who want to read such work critically.

Our primary take-away message is this: there is no such thing as a generically 'correct' approach to phylogenetics. Phylogenetic inference, like many problems in bioinformatics, depends on a model representing the biological processes we seek to explain, a data source that we seek to explain in terms of that model, and an algorithm to fit the data to the model. Effective use of phylogenetics involves making appropriate choices of model, data, and algorithm so that all three are mutually consistent and suited to the question at hand.

To frame this discussion, we present it in terms of a hypothetical research study. Let us suppose we have a general scientific question: what are the common recurring sequences and timings of CNVs over the progression from healthy breast tissue to breast cancer? We further suppose that we have gathered data to address that question: whole-genome DNA sequencing at 50× coverage from 200 single cells from a tumour and matched normal control. We then imagine that we built a phylogeny using an off-the-shelf neighbour joining phylogeny program[152], as was done in several prominent studies[71,76,80,97,145,158]. This is a qualitatively similar plan to the pioneering work of Navin *et al.*[71,157]. How can we evaluate, and perhaps improve upon, this initial plan?

Is our model consistent with our data? If we carry out the study described above, it will yield a phylogenetic tree, and we can expect that tree to be qualitatively similar to those of Navin *et al.*[71] and Wang *et al.*[156]: an early split of clones into ploidy classes (diploid, pseudo-diploid, and tetraploid) followed by later separation by more fine-scale CNVs. This may indeed be the true evolutionary history of the tumour, but we also need to consider that it may be an artefact of the approach. In fact, the study as described will yield this outcome regardless of the actual evolutionary history of the tumour for reasons implicit in the model of evolution that our strategy assumed.

The described approach uses a phylogeny model designed primarily for SNV data; such a model assumes that evolution occurs by mutations independently accumulating one at a time, with roughly equal rates. This happens to be a reasonable simplification for species evolution[82], and probably even for tumour evolution — provided that we are tracking evolution in which SNVs accumulate largely without selection[8,46–48]. However, it is a questionable model for CNVs, as CNVs violate the model assumption that changes in distinct variant regions accumulate independently; instead, CNVs accumulate at multiple scales, from localized gene-scale variants to variation at the scale of large chromosome segments, whole chromosomes or even whole-genome ploidy[17,74,159]. The mismatch between model and data can lead to discrepancies between evolutionary distance measures. For example, a 3 billion bp change induced by a whole-genome duplication will yield the same estimated evolutionary distance as 300,000 independent 10 kb changes, even though genome duplications are common events[21,22,147,160] and far more likely to occur than 300,000 independent 10 kb changes. That discrepancy will lead to large-scale changes being misinterpreted as being older than they actually are relative to localized changes, which could radically skew our trees.

If we recognize this issue, it would be logical to propose that we fix the model. There are models for representing the more complex nature of evolution by CNVs versus evolution by SNVs[51], and there are some custom-designed phylogeny tools for specific variants of CNV evolution[43,69,84,85,151,152]. A Bayesian probabilistic model, as has been used in many tumour phylogeny approaches[47,79,86,135], can handle arbitrarily complex evolutionary scenarios and is well suited to learning the complicated lineage-specific rate parameters one would need for such a model[78,101,152], given the diversity of CNV hypermutability mechanisms any given tumour might exhibit[20,25,94,161]. Let us suppose, then, that we replace the Euclidean distance model with a Bayesian probabilistic model that captures the multi-scale nature of CNV evolution, thus bringing our model in line with our data. Are we now finished?

Aligning algorithm and model. Unfortunately, the change to a Bayesian model is insufficient because we cannot change the model without also changing the algorithm. In principle, one could use neighbour joining with a more nuanced probabilistic model of evolutionary distances. However, a distance-based method such as neighbour joining will work poorly if we lack large numbers of mutations of each type to average out uncertainty over the mutation frequencies and relative orders, and will therefore be likely to fail for important but rarer CNVs, such as ploidy changes.

◀ Figure 2 | **Some challenges in synchronizing data, models and algorithms when applying tumour phylogenetics to a scientific question.** An illustration of a hypothetical scenario described in the main text, in which we seek to infer a phylogenetic history of copy number variant (CNV) events in the progression of a single tumour. Each tree shows the potential evolution of genomic copy number profiles for a set of observed clones (blue lines) and computationally inferred intermediate states (red lines) for a single tumour. **a** | The hypothetical 'true' tree describing the evolution of a set of clones from a diploid root via a series of CNVs: gain or loss of copy number in a localized region, as well as whole-genome duplication, leading to a doubled copy number genome-wide. **b** | Incorrect inference due to the use of a model designed for single-base changes, leading to a substantially incorrect phylogeny involving various biologically 'impossible' evolutionary events, such as partial (non-integer) gain, loss, or whole-genome duplications, leading to fractional copy numbers. **c** | Improved but still inaccurate inference after correcting to an evolutionary model cognizant of the type of variation occurring with CNVs; this eliminates impossible events and leads to a more accurate tree topology, but still fails to identify the correct tree because the analysis is using an algorithm that identifies biologically plausible but still sub-optimal phylogenies for this kind of evolutionary model. **d** | Still inaccurate inference after changing to a more sophisticated model and algorithm that are well suited to CNV evolution but make it impractical to use single-cell sequence data; this forces a change to a bulk genomic data type, leading to inadequate sampling of extant clones to capture the rapid mutation process typical of CNV-driven evolution and observed in the true tree.

Recognizing that problem, we can adopt a more appropriate algorithm for a probabilistic model, such as the MCMC approach of the BitPhylogeny program[86]. Although other classes of algorithm can be used with such probabilistic models[108,112,113,143,162,163], MCMC sampling is the standard for accurately fitting a complicated probabilistic model for which we do not yet have a specialized body of theory[87]. Let us suppose, then, that we have replaced the neighbour joining algorithm with an MCMC Bayesian sampler, making our algorithm consistent with our model. Having synchronized algorithm to model, and model to data, are we finished?

Aligning model and data. Unfortunately, the algorithm change is insufficient, because in synchronizing our algorithm to our model, we selected an algorithm that is not appropriate to our data. Algorithms, like models, carry assumptions and limitations. One limitation of MCMC is computational cost[82]: the number of steps for which one must run an MCMC algorithm to get accurate results generally grows exponentially with the number of species (or cells) in the data. This limitation is perilous to the novice user because an MCMC algorithm can still generate a tree as an output, even if it has not run for long enough to identify the right tree. MCMC phylogeny algorithms were therefore traditionally used only for the order of 10–20 species, although somewhat larger numbers are now possible[85]. State-of-the-art Bayesian methods in tumour phylogenetics are commonly accelerated with a technique called approximate Bayesian computation (ABC)[164], as used in two recent investigations[47,79]; with this technique, one accelerates sampling by collapsing sets of solutions that appear to be similar by one or more summary statistics. However, the curse of exponential blow-up (in which the number of possible trees an algorithm must consider grows exponentially with the number of 'species' they contain) is that better algorithms will allow us only a few more species, not the order-of-magnitude increase we need to handle 200 cells. In short, the algorithm we have chosen is probably unsuitable for our single-cell data.

A logical next step, then, would be to ask whether we might use a different kind of data more appropriate to our approach. There are other marker types that we could consider — such as SNVs[87,154], expression[123], methylation[73], or microsatellites[165]. However, as we are interested in evolution by CNVs, we must keep the marker type unchanged and instead change only the study design. We might propose to use a regional bulk method, replacing our 200 single cells with bulk sequencing of 10 regions from each of 20 tumours. Each of the twenty resulting trees is on a scale an MCMC sampler can handle. Similar regional MCMC strategies for regional bulk sequencing have yielded important insights into tumour evolution in prior studies[47,59,79] and have been used successfully for CNV data[86,129]. We will then have a model appropriate to our data, an algorithm appropriate to our model, and a data set appropriate to our algorithm, harmonizing the three components of our method. Are we now finished?

Aligning method and questions. We are unfortunately still not finished because by changing the data collection strategy to smaller sets of 'species' per tree, we have ended up with data sets that are too small to resolve the fine-scale trajectories of CNV evolution. Most solid tumours have chromosome replication defects that lead to rapid accumulation of CNVs[20], and progression can happen via clones that are minor or rare in the earlier tumour stages[11,41,50,53,54,166,167] and that may lay dormant through much of the clinical progression[168]. We can expect that there are too many CNVs among ten tumour regions to have hope of resolving the orders or timings of CNV events[157]. Our study design might be fine for other questions about CNV evolution, but not for the question we are asking. We still have not managed to find a model, algorithm, and data source that are consistent with one another and with the question we are asking.

FIGURE 2 provides a simplified overview of the pitfalls in this process, as we seek to infer a true tree (FIG. 2a) but struggle with erroneous inferences induced by a mismatch between the evolutionary model and data type (FIG. 2b), between the algorithm and the model (FIG. 2c), and between the data type and the research question (FIG. 2d).

That does not mean that we are out of options. We could try a wholly different approach, perhaps reverting to our original scSeq study design but using a parsimony model with a faster algorithm that might be better able to handle the scale of data. That is essentially what was done in a recent scSeq study of breast cancer[157]. TABLES 1,2 give examples of available methods and prior studies that may be helpful for finding an existing tool or strategy that has been successfully applied to similar questions. We could try to bring in more exotic algorithms (for example, integer linear programming (ILP)[120,131,144,169]) or completely different marker types (for example, FISH[69,84,151,152]). However, we should be aware that we might run through every existing option for a model, an algorithm, and the data type and still fail to find a combination that is mutually consistent and appropriate to our questions. What, then, is the cancer researcher to do?

A final important point is that posing a computational problem is not the same thing as solving it, even if we have perfect data and a perfect model of the relevant biological mechanisms. Many reasonable phylogenetics problems are classified as 'intractable' (REF. 170), which informally means that for even moderately large data sets, we may not have any method for finding a good tree efficiently. Often, we will need to develop new computational theory to find an adequate explanation of the data within the models of evolution that we believe describe them. The answer to the question above — what, then, is the cancer researcher to do? — is often to recognize that there is no standard off-the-shelf technique available for many important questions and that developing one is a research problem which will require time and significant expertise in computational biology. BOX 1 provides a few examples of important unsolved methodological problems in tumour phylogenetics.

Box 1 | **Outstanding problems in tumour phylogenetics**

Novel or heterogeneous data sources
There are many tumour phylogeny methods for common forms of DNA variation (such as single nucleotide variants (SNVs) and copy number variants (CNVs)), some methods for other genomic data (for example, karyotypes, expression data or methylation data), and a few methods integrating multiple forms of data[29,73]. The field is just beginning to make sense of other sources of information, such as spatial distributions of cell populations[70,79,83], and to make productive use of heterogeneous data[10,134].

Comprehensive evolutionary models
We are currently lacking the quantitative models that are required for phylogenetic inference of complex evolutionary events, particularly for recently discovered mechanisms such as chromothripsis or kataegis. However, models are also lacking even for SNVs, which exhibit nuanced combinations of mutational signatures[26,27] and selective pressures[8] that vary by tumour type, patient, and time[9,21,48,146,175] and may require extensive sampling to estimate[46].

Tumour-specific phylogeny algorithms
Most tumour phylogeny work to date has relied on a handful of conventional phylogeny algorithms (neighbour joining, maximum parsimony, maximum likelihood or Markov chain Monte Carlo), and it remains unclear which, if any, are sufficiently accurate for tumour evolution for any given data type and tumour type. Phylogenies including more exotic tumour-specific mechanisms will require new classes of algorithms, which is a largely unexplored topic.

Beyond 'species' trees
Models drawn from species trees themselves may be inadequate descriptions of clonal evolution of tumours[10] for such reasons as cooperation between clones, seeding of metastases by multiple clones, reseeding of primary tumours[8], or co-evolution with the microenvironment[176]. More specialized tree models, generalizations to non-tree-based evolution, and methods informed by more sophisticated population genetics and ecological models are just beginning to emerge[47].

Statistical analysis, study design and reproducibility
Few studies examine enough subjects to draw statistically sound conclusions in the presence of extensive inter-tumour heterogeneity[14], particularly for single-cell sequencing studies, which usually involve at most a few tumours[150]. Questions that depend on finding reproducible features across many tumours — for example, whether tumour evolution is linear or branched, or whether it branches early or late — have largely been addressed anecdotally rather than by adequately powered analyses. There are currently no accepted methods to judge whether a phylogenetic tree provides a well-supported fit to a single tumour[86,89]. The field has barely explored the problem of how to plan a study to ensure that informative and robust phylogenetic tree building will even be possible.

Conclusions and discussion

The use of phylogenetic techniques in cancer research is growing, as is evidenced by the large body of work completed in the past 3 years and referred to herein. Studies of cancer phylogenetics have advanced far beyond the theoretical evolutionary model of Nowell[2] to reveal the enormous complexity of the actual processes of tumour evolution[14,20,26,171], and to uncover the heterogeneity of those processes both patient to patient[101,110,172] and lineage to lineage in a single patient[3,8,21,40,73,145]. Such studies have revealed mechanisms underlying this heterogeneity[21,85,145,173], the dynamics by which these mechanisms themselves evolve over tumour progression[47], and possibilities for novel prognostic indicators[43,56,57,105]. As our knowledge of tumour evolution has expanded, tumour phylogenetics has itself evolved from a new tool for asking old questions, such as distinguishing driver from passenger mutations[13,174], to a source of new questions on topics such as how the evolutionary landscapes of tumours are shaped by environmental factors[16,21,26,61] and treatment[32] and how they can reveal the past and possibly the future of the progression of a tumour[27–30,47,59] in ways that are tangential to the specific driver mutations dominant in a given tumour at some time. In this Review, we have sought to survey key methods used and results obtained to date and to provide insight into how best to harness phylogenetic tools for new applications. We conclude by considering where tumour phylogenetics might go next.

Most uses of tumour phylogenetics to date have been in retrospective studies; a major opportunity is moving to prospective studies in research clinics. Looking ahead to this opportunity, we consider what happened in clinical cancer research with the advent of gene expression microarrays and NGS. Both technologies were expected to have an impact on cancer diagnosis and treatment, but they have had different outcomes. Gene expression microarrays were shown to have prognostic value in hundreds of research studies, but are not currently widely used in the clinic. By contrast, NGS is being used in the clinic and has led to the phenomenon of 'tumour boards' formed by multidisciplinary scientists and clinicians who study mutation profiles, determine which mutations are 'actionable' on the basis of approved drugs, and develop 'precision medicine' treatment plans. We hypothesize that the difference is primarily due to the type of output that these two technologies produce. NGS generates lists of discrete mutations that can be validated and evaluated individually. By contrast, microarrays yield patterns of expression changes, sometimes called gene expression signatures, that are conceptually similar to the nucleotide variant mutational signatures of cancer genomes[26]. Similarly, so far the prognostic value from tumour phylogenetics has come from analysing the patterns of evolution[43,58], not from analysing discrete events.

Tumour phylogenetics is far from achieving the reproducibility that is needed for clinical work. The field will need to overcome resistance to complex data and dynamic analysis, and must develop principled, robust methods of analysis implemented in software that is used in many laboratories. Tumour phylogenetics is itself evolving, but not as quickly as biotechnologies to measure aberrations in tumours. In the future, new phylogeny methods will need to be tested on more data sets and compared head to head. The head-to-head comparisons will be very complicated if new methods address one of the key unsolved problems in the field: the integration of different data types beyond CNVs and SNVs (BOX 1). The analysis methods that are deemed most fit will be selected by more research groups and gain wider usage. After some specific methods are used in hundreds of studies, we hope that the results will be sufficiently robust and interpretable to aid in patient prognosis and treatment planning. Similarly, clinicians interpreting phylogenetic analyses should insist that different methods be tried and that results are actionable only when different methods of analysis lead to the same qualitative understanding of a patient's tumour[68,106].

Finally, one cannot comprehensively discuss the future prospects of tumour phylogenetics without considering the education of cancer researchers, or biomedical researchers in general. We have provided

guidance on how someone new to tumour phylogenetics might evaluate and carry out research in this domain, but such basic principles can only go so far. As we have shown here, effective use of even well-developed tools requires some understanding of their mathematical and algorithmic underpinnings. Finding or developing the appropriate phylogeny tools for a given application will often involve difficult problems of model selection (as reported by Yuan *et al.*[86]) and algorithm design that lie far beyond what we can discuss here. Furthermore, by the time one has identified a study cohort and planned data collection, the questions it is possible to ask with these tools are already constrained. In addition, there are limits to what one can ask at all with the available computational tools. Some important questions require new tool development or theoretical advances before they can be answered. It is crucial to involve computational biologists early in the study design phase, to ensure that it will be possible, in principle, for the study to resolve the questions that motivate it. More specifically, these specialists can determine that analysis tools appropriate to the data, evolutionary models, and questions do currently exist or that there is a plausible path to developing appropriate tools. Even the casual user must be able to recognize these situations. Actually posing and solving new data-driven questions, within the constraints of the limits of biotechnology and human cohorts, are demanding skills that will be needed by the leaders of future research efforts in cancer evolution. Few life scientists today are adequately trained in the fundamentals of computational thinking to handle these questions, and not many computational scientists are adequately trained in the challenges of genomic data and research involving human subjects. If we are to realize the full potential of cancer phylogenetics, we will require a sea change in the training of cancer researchers to inculcate a sophisticated understanding of how to reason about data-driven research. The required changes in education practice are likely to face institutional obstacles, but resolving them is as important to the future progress of cancer research as any purely scientific question considered here.

1. Hanks, S. *et al.* Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in (BUB1B). *Nat. Genet.* **36**, 1159–1161 (2004).
2. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
This is a seminal paper proposing that solid tumours evolve clonally while accumulating mutations from one mitosis to the next via a process of selection of mutant subpopulations from a common progenitor cell.
3. Polyak, K. Is breast tumor progression really linear? *Clin. Cancer Res.***14**, 339–341 (2008).
4. Naxerova, K. & Jain, R. K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **12**, 258–272 (2015).
5. Foo, J. & Michor, F. Evolution of acquired resistance to anti-cancer therapy. *J. Theor. Biol.* **355**, 10–20 (2014).
6. Enriquez-Naxas, P. M. *et al.* Exploiting evolutionary principles to prolong tumor control in preclinical models of breast cancer. *Sci. Transl Med.* **8**, 327ra24 (2016).
7. Merlo, L. M. F., Pepper, J. W., Ried, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer* **6**, 924–935 (2006).
8. Burrell, R. A. & Swanton, C. Re-evaluating clonal dominance in cancer evolution. *Trends Cancer* **2**, 263–276 (2016).
9. Cross, W. C., Graham, T. A. & Wright, N. A. New paradigms in clonal evolution: punctuated equilibrium in cancer. *J. Pathol.* **240**, 126–136 (2016).
10. Podlaha, O., Riester, M., De, S. & Michor, F. Evolution of the cancer genome. *Trends Genet.* **28**, 155–163 (2012).
11. Ding, L., Raphael, B. J., Chen, F. & Wendl, M. C. Advances for studying clonal evolution in cancer. *Cancer Lett.* **340**, 212–219 (2013).
12. Altrock, P. M., Liu, L. L. & Michor, F. The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer* **15**, 730–745 (2015).
13. Beerenwinkel, N., Schwarz, R. F., Gerstung, M. & Markowetz, F. Cancer evolution: mathematical models and computational inference. *Syst. Biol.* **64**, e1–e25 (2015).
This is an in-depth review of applications of mathematical models of evolution to many problems in cancer research, including examples of various techniques drawn from phylogenetics, population genetics, stochastic processes, and game theory and related areas.
14. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
15. Loeb, L. A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075–3079 (1991).
16. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **54**, 4855–4878 (1994).
17. Cahill, D. P., Kinzler, K. W., Vogelstein, B. & Lengauer, C. Genetic instability and darwinian selection in tumours. *Trends Cell Biol.* **9**, M57–M60 (1999).
18. Harris, R., Petersen-Mahrt, S. & Neuberger, M. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10**, 1247–1253 (2002).
19. Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
20. Heng, H. H. *et al.* Chromosome instability (CIN): what it is and why it is crucial to cancer evolution. *Cancer Metastasis Rev.* **32**, 325–340 (2013).
21. de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
This is a particularly instructive study for critically evaluating the application of phylogenetics to bulk tumour samples, in part because it considers multiple phylogenetic methods and recognizes that some samples yield multiple optimal tree topologies.
22. Gibson, W. J. *et al.* The genomic landscape and evolution of endometrial carcinoma progression and abdominopelvic metastasis. *Nat. Genet.* **48**, 848–855 (2016).
23. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
24. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
25. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
26. Alexandrov, L. *et al.* Signatures of mutation processes in human cancers. *Nature* **500**, 415–421 (2013).
27. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Devel.* **24**, 52–60 (2014).
28. Hong, M. *et al.* Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat. Commun.* **6**, 6605 (2015).
29. McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl Med.* **7**, 283ra54 (2015).
30. Murugaesu, N. *et al.* Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov.* **5**, 821–831 (2015).
31. Kim, Y. A., Madan, S. & Przytycka, T. M. WeSME: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics* http://dx.doi.org/10.1093/bioinformatics/btw242 (2016).
32. Allan, J. M. & Travis, L. B. Mechanisms of therapy-related carcinogenesis. *Nat. Rev. Cancer* **5**, 943–955 (2005).
33. Johnson, B. E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
34. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
35. Kostadinov, R. L. *et al.* NSAIDs modulate clonal evolution in Barrett's esophagus. *PLoS Genet.* **9**, e100353 (2013).
This is an important investigation for demonstrating the ability of treatment to shape the pre-cancer evolutionary landscape. It provides evidence of a more than order-of-magnitude decrease in mutation rates for patients with Barrett oesophagus who took non-steroidal anti-inflammatory drugs (NSAIDs) versus those who did not.
36. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
37. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim. Biophys. Acta* **1805**, 105–117 (2010).
38. Park, S. Y., Gönen, M., Kim, H. J., Michor, F. & Polyak, K. Cellular and genetic diversity in the progression of *in situ* human breast carcinomas to an invasive phenotype. *J. Clin. Invest.* **120**, 636–644 (2010).
39. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).
40. Gerlinger, M.*et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
41. Heselmeyer-Haddad, K. *et al.* Single-cell genetic analysis of ductal carcinoma *in situ* and invasive breast cancer reveals enormous tumor heterogeneity, yet conserved genomic imbalances and gain of *MYC* during progression. *Am. J. Pathol.* **181**, 1807–1822 (2012).
42. Kim, T. M. *et al.* Subclonal genomic architectures of primary and metastatic colorectal cancer based on intratumoral genetic heterogeneity. *Clin. Cancer Res.* **21**, 4461–4472 (2015).
43. Schwarz, R. F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med.* **12**, e1001789 (2015).
44. Turajlic, S., McGranahan, N. & Swanton, C. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochim. Biophys. Acta* **1855**, 264–275 (2015).

45. Hong, W. S., Shpak, M. & Townsend, J. P. Inferring the origin of metastases from cancer phylogenies. *Cancer Res.* **75**, 4021–4025 (2015).
46. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl Acad. Sci. USA* **112**, E6496–E6505 (2015).
47. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
This study on the evolution of colorectal cancer illustrates the importance of deep evolutionary theory in interpreting genomic data from tumours. It provides evidence that largely selectively neutral mutations can occur, in contrast to one of the two evolutionary principles of Nowell (reference 2) and others: that cancer evolves by a gradual series of genomic aberrations and that there is strong selection for those aberrations that are more favourable to tumour progression.
48. 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 (2016).
49. Shi, H. *et al.* Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy. *Cancer Discov.* **4**, 80–93 (2014).
50. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
51. Andor, N., Harness, J. V., Müller, S., Mewes, H. W. & Petritsch, C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* **30**, 50–60 (2014).
52. Juric, D. *et al.* Convergent loss of PTEN leads to clinical resistance to a PI(3)Kα inhibitor. *Nature* **518**, 240–244 (2015).
53. Morrissy, A. S. *et al.* Divergent clonal selection dominates medulloblastoma at recurrence. *Nature* **529**, 351–357 (2016).
54. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
55. Fisher, R., Pusztai, L. & Swanton, C. Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer* **108**, 479–485 (2013).
56. Cooke, S. L. *et al.* Intra-tumour genetic heterogeneity and poor chemoradiotherapy response in cervical cancer. *Br. J. Cancer* **104**, 361–368 (2011).
57. Almendro, V. *et al.* Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep.* **6**, 514–527 (2014).
58. Wangsa, D. *et al.* Phylogenetic analysis of multiple FISH markers in oral tongue squamous cell carcinoma suggests that a diverse distribution of copy number changes is associated with poor prognosis. *Int. J. Cancer* **138**, 98–109 (2016).
59. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl Acad. Sci. USA* **110**, 4009–4014 (2013).
60. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
61. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256–259 (2014).
62. Comen, E., Norton, L. & Massagué, J. Clinical implications of cancer self-seeding. *Nat. Rev. Clin. Oncol.* **8**, 369–377 (2011).
63. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54–58 (2014).
64. Sanborn, J. Z. *et al.* Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc. Natl Acad. Sci. USA* **112**, 10995–11000 (2015).
65. Tsao, J. *et al.* Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. *Am. J. Pathol.* **153**, 1189–1200 (1998).
66. Desper, R. *et al.* Inferring tree models of oncogenesis from comparative genomic hybridization data. *J. Comput. Biol.* **6**, 37–51 (1999).
This is the first report to suggest that there might be difficulties with modelling tumour progression as a tree construction problem in phylogenetics.
67. Papaemmanuil, E. *et al.* Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* **122**, 3616–3627 (2013).
68. Zhao, Z. *et al.* Early and multiple origins of metastatic lineages within primary tumors. *Proc. Natl Acad. Sci. USA* **113**, 2140–2145 (2016).
This study is an important advance over prior bulk sequencing studies for at least three reasons: it compares different phylogenetic methods and draws inferences only when the methods agree on the tree topology; it combines the SNVs in a manner that does not require the use of variant allele frequencies to infer subclones; and it provides clear evidence that some metastases branch early and in parallel, whereas others have a single late origin, reconciling contradictory conclusions reached by earlier studies.
69. Pennington, G., Smith, C. A., Shackney, S. & Schwartz, R. Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol.* **5**, 407–427 (2007).
70. Bashashati, A. *et al.* Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J. Pathol.* **231**, 21–34 (2013).
71. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
This is the seminal paper in developing and demonstrating the biotechnology to perform scSeq in tumours and apply it to phylogenetic inferences of single tumours.
72. Klein, C. A. Selection and adaptation during metastatic cancer progression. *Nature* **501**, 365–372 (2013).
73. Brocks, D. *et al.* Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.* **8**, 798–806 (2014).
This bulk tumour phylogeny study is of interest because the authors combine CNV data and DNA methylation data, showing a high correlation of inferred inter-sample evolutionary distances between inferences derived from genetic and from epigenetic data.
74. Hastings, P., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10**, 551–564 (2009).
75. Attolini, C. S. O. & Michor, F. Evolutionary theory of cancer. *Ann. NY Acad. Sci.* **1168**, 23–51 (2009).
76. Frumkin, D. *et al.* Cell lineage analysis of a mouse tumor. *Cancer Res.* **68**, 5924–5931 (2008).
77. Salk, J. J., Horwitz, M. S. & Risques, R. A. Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. *Semin. Cancer Biol.* **20**, 294–303 (2010).
78. Shlush, L. I. *et al.* Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability. *Blood* **120**, 603–612 (2012).
79. Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C. & Tavaré, S. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer Res.* **73**, 41–49 (2013).
80. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
81. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314 (2001).
82. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, Inc.,2004).
83. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).
84. Chowdhury, S. A. *et al.* Phylogenetic analysis of multiprobe fluorescence *in situ* hybridization data from tumor cell populations. *Bioinformatics* **29**, i189–i198 (2013).
85. Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**, e1003535 (2014).
This work provides an important example of a robust cross-platform computational tool for tumour-specific phylogenetic inference, MEDICC, which carries out phylogenetic analysis of multiple samples from a tumour by quantifying intra-tumour heterogeneity while taking into account dependencies between genomic changes.
86. Yuan, K., Sakoparnig, T., Markowetz, F. & Beerenwinkel, N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* **16**, 36 (2015).
This study represents an exciting advance in the development and implementation of tumour phylogeny methods for third-party use, developing a full Bayesian model that can be applied to both bulk sequencing data and single-cell data.
87. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol.* **17**, 96 (2016).
88. Nicoloau, M., Levine, A. J. & Carlsson, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc. Natl Acad. Sci. USA* **108**, 7265–7270 (2011).
89. Kim, K. I. & Simon, R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* **15**, 27 (2014).
90. Misra, N., Szczurek, E. & Vingron, M. Inferring the paths of somatic evolution in cancer. *Bioinformatics* **30**, 2456–2463 (2014).
91. Lecca, P., Casiraghi, N. & Demichelis, F. Defining order and timing of mutations during cancer progression: the TO-DAG probabilistic graphical model. *Front. Genet.* **6**, 309 (2015).
92. Urbschat, S. *et al.* Clonal cytogenetic progression within intratumorally heterogeneous meningiomas predicts tumor recurrence. *Int. J. Oncol.* **39**, 1601–1608 (2011).
93. Fearon, E. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
94. Tomlinson, I. & Bodmer, W. Selection, the mutation rate and cancer: ensuring that the tail does not wag the dog. *Nat. Med.* **5**, 11–12 (1999).
95. Nowak, M. A. *et al.* The role of chromosomal instability in tumor initiation. *Proc. Natl Acad. Sci. USA* **99**, 16226–16231 (2002).
96. Khalique, L. *et al.* The clonal evolution of metastases from primary serous epithelial ovarian cancers. *Int. J. Cancer* **124**, 1579–1586 (2009).
97. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
98. Brastianos, P. K. *et al.* Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* **5**, 1164–1177 (2015).
This is one of the most interesting and instructive bulk tumour phylogeny studies to date because the study design was to investigate the evolution of metastasis across many tumour types, and the sample size (86) is among the largest for the bulk tumour studies carried out thus far.
99. Joung, J. G. *et al.* Nonlinear tumor evolution from dysplastic nodules to hepatocellular carcinoma. *Oncotarget* http://dx.doi.org/10.18632/oncotarget.10502 (2016).
100. Paracchini, L. *et al.* Regional and temporal heterogeneity of epithelial ovarian cancer tumor biopsies: implications for therapeutic strategies. *Oncotarget* http://dx.doi.org/10.18632/oncotarget.10505 (2016).
101. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
102. Heselmeyer-Haddad, K. *et al.* Single-cell genetic analysis reveals insights into clonal development of prostate cancers and indicates loss of *PTEN* as a marker of poor prognosis. *Am. J. Pathol.* **184**, 2671–2686 (2014).
103. Janocko, L. E. *et al.* Distinctive patterns of Her-2/neu, c-myc, and cyclin D1 gene amplification by fluorescence *in situ* hybridization in primary breast cancers. *Cytometry* **46**, 136–149 (2001).
104. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).
105. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
106. Naxerova, K. *et al.* Hypermutable DNA chronicles the evolution of human colon cancer. *Proc. Natl Acad. Sci. USA* **111**, E1889–E1898 (2014).
107. Tsao, J. *et al.* Colorectal adenoma and cancer divergence: evidence of multilineage progression. *Am. J. Pathol.* **154**, 815–1824 (1999).
108. von Heydebreck, A., Gunawan, B. & Füzesi, L. Maximum likelihood estimation of oncogenetic tree models. *Biostatistics* **5**, 545–556 (2004).

109. Bilke, S. *et al.* Inferring a tumor progression model for neuroblastoma from genomic data. *J. Clin. Oncol.* **23**, 7322–7331 (2005).
110. Beerenwinkel, N. *et al.* Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.* **12**, 584–598 (2005).
111. Beerenwinkel, N. *et al.* Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics* **21**, 2106–2107 (2005).
112. Hjelm, M., Höglund, M. & Lagergren, J. New probabilistic network models and algorithms for oncogenesis. *J. Comput. Biol.* **13**, 853–865 (2006).
113. Gerstung, M., Baudis, M., Moch, H. & Beerenwinkel, N. Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics* **25**, 2809–2815 (2009).
114. Rahnenführer, J. *et al.* Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics* **21**, 2438–2446 (2005).
115. Bogojeska, J., Alexa, A., Altmann, A., Lengauer, T. & Rahnenführer, J. Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics* **24**, 2391–2392 (2008).
116. Bogojeska, J., Lengauer, T. & Rahnenführer, J. Stability analysis of mixtures of mutagenetic trees. *BMC Bioinformatics* **9**, 165 (2008).
117. Attolini, C. S. *et al.* A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl Acad. Sci. USA* **107**, 17604–17609 (2010).
118. Cheng, Y. *et al.* A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Comput. Biol.* **8**, e1002337 (2012).
119. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
120. Shahrabi Farahani, H. & Lagergren, J. Learning oncogenetic networks by reducing to mixed integer linear programming. *PLoS ONE* **8**, e65773 (2013).
121. Ramazzotti, D. *et al.* CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31**, 3016–3026 (2015).
122. Desper, R. *et al.* Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.***7**, 789–803 (2000).
123. Desper, R., Khan, J. & Schäffer, A. A. Tumor classification using phylogenetic methods on expression data. *J. Theor. Biol.* **228**, 477–496 (2004).
124. Riester, M., Attolini, C., Downey, R. J., Singer, S. & Michor, F. A differentiation-based phylogeny of cancer subtypes. *PLoS Comput. Biol.* **6**, e1000777 (2010).
125. Liu, J., Bandyopadhyay, N., Ranka, S., Baudis, M. &Kahveci, T. Inferring progression models for CGH data. *Bioinformatics* **25**, 2208–2215 (2009).
126. Hainke, K., Rahnenführer, J. & Fried, R. Cumulative disease progression models for cross-sectional data: a review and comparison. *Biom. J.* **54**, 617–640 (2012).
127. Diaz-Uriarte, R. Identifying restrictions in the order of accumulation of mutations during tumor progression: effects of passengers, evolutionary models, and sampling. *BMC Bioinformatics* **16**, 41 (2015).
128. Sprouffske, K., Pepper, J. W. & Maley, C. C. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prev. Res.* **4**, 1135–1144 (2011).
129. Letouzé, E., Allory, Y., Bollet, M. A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol.* **11**, R76 (2010).
130. Hajirasouliha, I., Mahmoody, A. & Raphael, B. J. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics* **30**, i78–i86 (2014).
131. El-Kebir, M., Oesper, L., Acheson-Field, H. & Raphael, B. J. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* **31**, i62–i70 (2015).
132. Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* **16**, 91 (2015).
133. 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 (2014).
134. Greenman, C. D. *et al.* Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* **22**, 346–361 (2012).
135. Deshwar, A. G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
136. El-Kebir, M., Satas, G., Oesper, L. & Raphael, B. J. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.* **3**, 43–53 (2016).
137. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **113**, E5528–E5537 (2016).
138. Qiao, Y. *et al.* SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol.* **15**, 443 (2014).
139. Schwartz, R. & Shackney, S. E. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics* **11**, 42 (2010).
140. Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).
141. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
142. Zare, H. *et al.* Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10**, e1003703 (2014).
143. 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 (2015).
144. Malikic, S., McPherson, A. A., Donmez, N. & Sahinalp, C. S. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* **31**, 1349–1356 (2015).
145. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res.* **20**, 68–80 (2010).
146. Newburger, D. E. *et al.* Genome evolution during progression to breast cancer. *Genome Res.* **23**, 1097–1106 (2013).
147. Stachler, M. D. *et al.* Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat. Genet.* **47**, 1047–1055 (2015).
148. Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
149. Gertz, E. M. *et al.* FISHtrees 3.0: tumor phylogenetics using a ploidy probe. *PLoS ONE* **11**, e0158569 (2016).
150. Spiro, A. & Shapiro, E. Accuracy of answers to cell lineage questions depends on single-cell genomics data quality and quantity. *PLoS Comput. Biol.* **12**, e1004963 (2016).
151. Chowdhury, S. A. *et al.* Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput. Biol.* **10**, e1003740 (2014).
152. Chowdhury, S. A. *et al.* Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics* **31**, i258–i267 (2015).
153. Zhou, J. Lin, Y., Rajan, V., Hoskins, W. & Tang, J. in *Proc. 15th Int. Workshop on Algorithms in Bioinformatics. WABI 2015. Lecture Notes in Computer Science* Vol. 9289 (eds Pop, M. & Touzet, H.) 108–120 (Springer, 2015).
154. Ross, E. M. & Markowetz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **17**, 69 (2016).
155. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
156. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
157. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
158. Tao, Y. *et al.* Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proc. Natl Acad. Sci. USA* **108**, 12042–12047 (2011).
159. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
160. Dewhurst, S. M. *et al.* Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discov.* **4**, 175–185 (2014).
161. Aguilera, A. & Gómez-González, B. Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.* **9**, 204–217 (2008).
162. Youn, A. & Simon, R. Estimating the order of mutations during tumorigenesis from tumor genome sequencing data. *Bioinformatics* **28**, 1555–1561 (2012).
163. Purdom, E. *et al.* Methods and challenges in timing chromosomal abnormalities within cancer samples. *Bioinformatics* **29**, 3113–3120 (2013).
164. Beaumont, M. A. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41**, 379–406 (2010).
165. Tsao, J. *et al.* Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl, Acad. Sci. USA* **97**, 1236–1241 (2000).
166. Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999–1005 (2010).
167. Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).
168. Eyles, J.*et al.* Tumor cells disseminate early, but immunosurveillance limits metastatic outgrowth, in a mouse model of melanoma. *J. Clin. Invest.* **120**, 2030–2039 (2010).
169. Catanzaro, D., Shackney, S., Schäffer, A. A. & Schwartz, R. Classifying the progression of ductal carcinoma from single-cell sampled data via integer linear programming: a case study. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **13**, 643–655 (2016).
170. Garey, M. R. & Johnson, D. S. *Computers and Intractability* (WH Freeman New York, 2002).
171. Salk, J. J. *et al.* Clonal expansions in ulcerative colitis identify patients with neoplasia. *Proc. Natl Acad. Sci. USA* **106**, 20871–20876 (2009).
172. McGlynn, K. A. *et al.* A phylogenetic analysis identifies heterogeneity among hepatocellular carcinomas. *Hepatology* **36**, 1341–1348 (2002).
173. Oesper, L., Mahmoody, A. & Raphael, B. J. THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14**, R80 (2013).
174. Greenman, C. *et al.* Patterns of somatic mutations in human cancer genomes. *Nature* **446**, 153–158 (2007).
175. Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352**, 169–175 (2016).
176. Martincorea, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
177. Subramanian, A., Shackney, S. & Schwartz, R. Inference of tumor phylogenies from genomic assays on heterogeneous samples. *J. Biomed. Biotechnol.* **2012**, 797812 (2012).
178. Szabo, A. & Boucher, K. Estimating an oncogenetic tree when false negatives and positives are present. *Math. Biosci.* **176**, 219–236 (2002).
179. De Sano, L. *et al.* TRONCO: an R package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics* **32**, 1911–1913 (2016).
180. Caravagna, G. *et al.* Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc. Natl Acad. Sci. USA* **113**, E4025–E4034 (2016).
181. Pennington, G., Smith, C. A., Shackney, S. & Schwartz, R. Expectation-maximization method for reconstructing tumor phylogenies from single-cell data. *Comput. Syst. Bioinformatics Conf.* **2006**, 371–380 (2006).
182. Potter, N. E. *et al.* Single cell mutational profiling and clonal phylogeny in cancer. *Genome Res.* **23**, 2115–2125 (2013).
183. Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat. Genet.* **47**, 367–372 (2015).
184. Cresswell, G. D. *et al.* Intra-tumor genetic heterogeneity in Wilms tumor: clonal evolution and clinical implications. *EBioMedicine* **9**, 120–129 (2016).

185. Yang, Z. *et al.* Single-cell sequencing reveals variants in ARID1A, GPRC5A and MLL2 driving self-renewal of human bladder cancer stem cells. *Eur. Oncol.* **71**, 8–12 (2017).
186. Ozawa, T. *et al.* Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like recursor glioma. *Cancer Cell* **26**, 288–300 (2014).
187. Eskilsson, E. *et al.* EGFRvIII mutations can emerge as late and heterogenous events in glioblastoma development and promote angiogenesis through Src activation. *Neuro Oncol.* **18**, 1644–1655 (2016).
188. Notta, F. *et al.* Evolution of human BCR–ABL1 lymphoblastic leukaemia-initiating cells. *Nature* **469**, 362–367 (2011).
189. Campbell, P. J. *et al.* Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA* **105**, 13081–13086 (2008).
190. Lamy, P. *et al.* Paired exome analysis reveals clonal evolution and potential therapeutic targets in urothelial carcinoma. *Cancer Res.* **76**, 5894–5906 (2016).
191. Eirew, P. *et al.* Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* **518**, 422–426 (2015).

Acknowledgements
This research was supported in part by the Intramural Research Program of the National Library of Medicine (part of the US National Institutes of Health) and by a grant from the Pennsylvania Department of Health (grant number 4100070287). The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

Competing interests statement
The authors declare competing interests: see Web version for details.

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF