Review

Giulio Caravagna*

Measuring evolutionary cancer dynamics from genome sequencing, one patient at a time

https://doi.org/10.1515/sagmb-2020-0075
Received December 5, 2020; accepted December 6, 2020; published online December 21, 2020

Abstract: Cancers progress through the accumulation of somatic mutations which accrue during tumour evolution, allowing some cells to proliferate in an uncontrolled fashion. This growth process is intimately related to latent evolutionary forces moulding the genetic and epigenetic composition of tumour subpopulations. Understanding cancer requires therefore the understanding of these selective pressures. The adoption of widespread next-generation sequencing technologies opens up for the possibility of measuring molecular profiles of cancers at multiple resolutions, across one or multiple patients. In this review we discuss how cancer genome sequencing data from a single tumour can be used to understand these evolutionary forces, overviewing mathematical models and inferential methods adopted in field of Cancer Evolution.

Keywords: bulk genome sequencing; clonal evolution; subclonal deconvolution.

1 Introduction

Cancer, with roughly 10 million deaths in 2018, is the second leading cause of death globally. A lot of efforts are at play to understand its aetiology, its relation to the genetic and epigenetic background of the individual patient and the associated disease trajectory. From a precision medicine point of view, we are interested in optimising treatment for each and every patient, leveraging on the most sophisticated data generation technologies, and the most powerful computational methods for data interpretation.

In this review we discuss mathematical models and statistical challenges to measure cancer evolutionary dynamics from genome sequencing of a single patient (we plan to discuss the joint analysis of multiple patients in a follow up manuscript). This topic is at the core of the emerging field of Cancer Evolution (CE), where we investigate tumours from an evolutionary perspective. This area of research relies on some of the most advanced technologies for Next Generation Sequencing (NGS), and is a very strong testbed for the application of many ideas from Machine Learning and Artificial Intelligence (Bi et al. 2019; Topol 2019).

Tumours are made of subpopulations that evolve from a single cell and compete for survival, following an evolutionary process where *positive*, *neutral* and *negative* selection modulates clone dynamics (see "A primer on tumour evolution", Section 2). The process outcome depends on the complex interplay of these forces, and the past tumour history is recorded in the cancer molecules (Greaves and Maley 2012; Greenman et al. 2007; Nik-Zainal et al. 2012). In particular, the signal of somatic evolution can be detected from multiple molecules that we can readout by NGS assays. The one we refer to in this review is DNA, where the signal are somatic mutations (in a broad sense), as we explain later. Of equal importance are RNA and other molecules (e.g., chromatin), which we however do not discuss in this review. The *clonal evolution model*, originally

***Corresponding author: Giulio Caravagna**, Department of Mathematics and Geosciences, University of Trieste, Via Valerio 12/1, 34127, Trieste, Italy, E-mail: gcaravagna@units.it. https://orcid.org/0000-0003-4240-3265

postulated by Peter Nowell in 1976, is a key component to study tumour evolution, both with and without therapy (Nowell 1976). In particular, "response" to therapy can be modelled as shifts in selective forces and summarised by evolutionary trajectories (Turajlic et al. 2019). The CE paradigm uses cancer NGS data to bring a "temporal dimension" in cancer analysis, using computational models to infer the evolutionary history of a neoplasm. From the practical point of view, many computational methods popularised in CE implement some form of feature selection or clustering for different types of NGS data. We apply these technologies either to single tumours, or to cohorts; in both case, data are strongly affected by sequencing noise, sampling bias and other confounders that depend on the specific analysis.

In a broad sense, the aim of CE is to quantify clonal selection from spatio-temporal patterns of somatic genetic and epigenetic changes (McGranahan and Swanton 2017; Shackleton et al. 2009; Turajlic et al. 2019). From the measurements obtained from sequencing we can unravel the tumour architecture and its "evolutionary signature". We can apply this to primary tumours, metastasis or post-treatment relapse samples. With a model of the tumour and data we can address precise quantitative questions spanning from basic tumour biology to advanced, controllable, tumour evolutionary dynamics (Gatenby et al. 2009). With data from multiple patients, we can extend these patterns across tumour patients and identify prognostic subtypes, along with their evolutionary biomarkers (Caravagna et al. 2016, 2018; Turajlic et al. 2018a, b).

This review regards the former type of problem, working with data of one patient. The concepts that we highlight can be extend to work with longitudinal data of a single patient, but are not covered in this review. The actual implementation of these ideas requires also to use bioinformatic tools to generate the somatic calls used for the analysis. This is a key step, and all we discuss here holds under the assumption that we can generate "good" calls to begin with (Househam et al., In preparation 2021). However, for the sake of brevity we focus this review on the evolutionary aspects of the analysis. The interested reader can find many other reviews that cover these topics and can help getting a broader perspective on these data analysis problems (see e.g. Dentro et al. 2017).

2 A primer on tumour evolution

Cancer growth is fuelled by various genetic and epigenetic lesions that accrue across generations of cancer cells, which in the following we just shall call "mutations". Depending on cancer type, these can be from few hundreds to several hundred thousands (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020). In loose terms, these can drive proliferation and therefore competition between cancer subclones and non-cancer cells. Among all mutations, the most prevalent ones are Single Nucleotide Variants (SNVs), substitutions of a single DNA nucleotide; less frequent events are insertion and deletions that involve multiple nucleotides, up to larger chromosomal structural rearrangements such as Copy Number Alterations (CNAs) and genomic fusions. Across cancers and patients these type of mutations have different prevalences, with exogenous factors (e.g., exposure to carcinogens such as tobacco or UV light) playing also an important role on the distribution of the somatic signals (Alexandrov et al. 2013; Ramazzotti et al. 2019; Rosenthal et al. 2016). For instance, haematological cancers such as leukaemia have far few Single Nucleotide Variants (SNVs) than lung or colon adenocarcinomas; similarly, certain ovarian cancer subtypes have hundreds of copy number events, and few SNVs (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020).

2.1 The clonal evolution model

To understand tumour growth we use the clonal evolution model (Nowell 1976). Tumours start from one healthy cell, triggered by a somatic driver mutation (Figure 1). From this cell x a new subclone seeds and expand through nested generations; the genetic variation of these cells is called *intra-tumour heterogeneity* (ITH) and can be detected by sequencing (Sottoriva et al. 2015). From an evolutionary perspective, this variation is *neutral within this expansion* (i.e., it is non-functional), which means that all the progeny of x has

the same *fitness*. Here x is Most Recent Common Ancestor (MRCA) of the clone, i.e., the most ancient ancestor from which the sequenced cells originate. This can be identified, at least conceptually, if we percolate backwards the tumour phylogenetic tree starting from its leaves, which are the sequenced cancer cells. When we sample a tumour, we always find an MRCA, which is not the MRCA of the whole tumour unless the biopsy contains all and only the whole sets of cancer cells in the tumour. For this reason, the MRCA is somewhat an arbitrary ancestor that depends on sampling (Caravagna et al. 2020).

The fitness of a clone is an abstract measure of its proliferative potential, usually denoted by s (we discuss mathematical models in Section 3.1). Fitness increases when a cell acquires a new driver mutation; drivers co-occurring in the same cell confer further fitness increase thorough epistatic interactions. There can be several types of epistatic interactions in cancer (Beerenwinkel et al. 2007; Caravagna et al. 2016; Diaz-Uriarte 2018; Diaz-Uriarte and Vasallo 2019; Ramazzotti et al. 2015). For instance, the negative one is important when we study multiple patients, as it can be related to patterns of mutual exclusivity across genes mutated in non-overlapping sets of patients. In this context, when a new subclone is triggered with higher fitness than its MRCA, it can proliferate faster. By enjoying a force of *positive* selection, the new subclone is pushed to colonise the whole tumour mass (in the long run). This evolutionary process is at play with potentially multiple co-existing subclones, and is also affected by random drift.

Cancer clonal dynamics is also heavily affected by *negative selection*, especially when mechanisms of immune response are active or boosted by therapy (Martincorena et al. 2017; Pich et al. 2019; Zapata et al. 2018). Cancer cells that harbour certain neo-antigens are exposed to T-cells, with the effect of becoming depleted by the immune system, a process that can be mathematically modelled (Lakatos et al. 2020; Zapata et al. 2020). In very loose terms, many modern cancer immunotherapies either try to boost or restore the immune system, a natural antagonist of the cancer (Schumacher and Schreiber 2015). As a consequence of the strong immune pressure, resistance to this type of treatments often emerges through complex mechanisms, genetic or epigenetic, that allow for immune evasion. A notable example of genetic mechanism is the copy number loss of heterozygosity of the human leukocyte antigen (HLA) system, which hijacks a machinery that tumour cells use to expose neoantigens (Christopher et al. 2018; Toffalori et al. 2019; Vago et al. 2009).

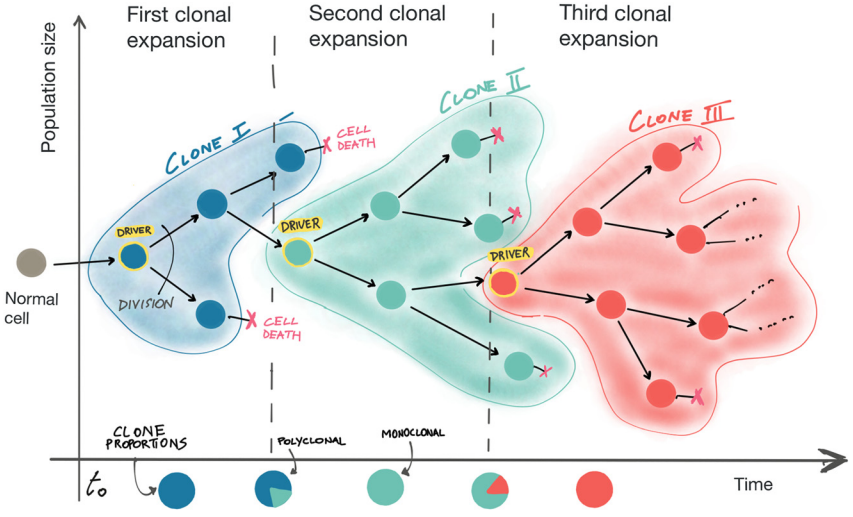

Figure 1: Clonal evolution model with three clones. The tree of cell divisions is represented and the clonal expansions coloured. Each cell division new mutations are acquired, most of them are neutral and do not change the fitness level of the cell where they happen. Others can instead increase proliferation and survival, *de facto* increasing the level of fitness s of the clone. Nested expansions such as these describe a case with $s_0 < s_1 < s_2$. Drift and random death can happen along the tree, as stochastic events. The clones that we can infer from data of this tumour depends on when we observe the process; In general, we make inferences about the ongoing clonal history of the tumour.

2.2 Cancer sequencing data

We can use several types of molecular data to study cancer, but for the problems that we discuss in this review we will refer mostly to bulk DNA sequencing. These are readouts of DNA fragments which we align to a reference genome through standard bioinformatics (DePristo et al. 2011). A common *experimental design* consists in acquiring one or more tumour samples, together with a biopsy of "normal" cells (e.g., from a distant tissue, saliva, blood or anything that is free of cancer cells).

Sequencing reads are processed with bioinformatics tools to identify germline and somatic mutations. This is done by detecting the variation to the reference independently in both the normal and tumour; somatic tumour mutations are then obtained by subtracting the signal of the normal from the tumour (since a normal cell is the tumour's MRCA). The most reliable source of information to infer the evolutionary history of the tumour are simple SNVs, for which we use the substitution frequency of the reference allele v, against the variant allele r. At genome position ℓ, we consider the *Variant Allele Frequency* (VAF)

$$v_\ell = \frac{n_\ell}{d_\ell},$$

where

- the depth of sequencing d_ℓ, defined as the total number of reads that span ℓ;
- the total number of reads n_ℓ with the variant allele r, in position ℓ.

The VAF is a proxy for the prevalence of the mutation in the cell population, which we call *Cancer Cell Fractions* (CCF). If the tumour was diploid, a mutation present in all cells would have CCF equal to 1, and VAF 0.5. In more general cases, when cancers have CNAs (i.e., high levels of aneuploidy), adjustment is a bit more complicated as we need to retrieve the number of copies of the mutation in the genome (see a good review in Dentro et al. 2017).

The reality is more complicated, since with bulk sequencing we often end up sequencing also DNA from normal cells. Therefore $d_\ell - n_\ell$, the number of reference reads, is a composition of reads from *both* normal *and* tumour cells. Therefore in the computation of CCFs we need to adjust frequencies also for sample purity (i.e., the percentage of tumour cells in the bulk sample).

3 Measuring evolution in a single patient

We can use genome sequencing data to measure clonal evolution in a single patient. Ideally, we will be using *whole-genome sequencing* (WGS) data of one, or more, biopsies of the same tumour (which could be either a primary or a metastasis). It is possible to use *whole-exome sequencing* (WES) data, but there are no gold standard rules. The key variable here is the tumour mutational burden, as that affects the VAF distribution we see with a WES or WGS assay.

3.1 Mathematical models of tumour growth

Population genetics is the ground upon which we can formalise the clonal evolution model (Durrett 2002; Ewens 2012; Kimura 1994; Tavaré 1984). There are however some differences between the fields which are worth noting. First, in cancer we have no sexual recombination, which makes modelling easier. Second, evolution in cancer is on a microscopic time-scale, compared to canonical species evolution (e.g., over thousands or millions of years). Third, human cancers are independent realisations of some latent evolutionary processes, as opposite to canonical species that derive from a unique stream of evolution (Caravagna et al. 2018).

In the population genetics literature there are several popular cell growth models; we refer to (Beerenwinkel et al. 2015) for a rich review on these topics. The Moran process, the Wright–Fisher process and the coalescent are the standard models for finite populations of constant size. Branching processes (e.g., the Galton–Watson process) are more general stochastic models for well-mixed populations that have finite, but fluctuating, size. This kind of stochastic branching process has been successful in describing selection-driven bacterial dynamics observed by Luria and Delbrück (Fusco et al. 2016; Kessler and Levine 2013, 2015). Here we give a brief and intuitive description of a possible process, and refer to more advanced literature for details (Beerenwinkel et al. 2015; Nowak 2006; Williams et al. 2016).

The growth of a population of cells can be described as a Markov birth-death process that starts from a single cell, which at every step undergoes a probabilistic choice (independently from all other cells): the cell either proliferates by generating two distinct daughter cells (asymmetrical cell division), or it dies (Williams et al. 2016; Tung and Durrett 2020). There are also symmetric division cases, often linked to the cancer stem cell model, where only one daughter cell is produced and the ancestor cell retained (Shackleton et al. 2009). Despite this change the conceptualisation of the model is not very different, and the inter-event probability follows an exponential distribution (i.e., it is memoryless, as it depends on the current state and other parameters).

3.1.1 A single clonal expansion

It is easy and instructive to visualise the special case of a single clone with fitness advantage $s > 0$ (*monoclonal expansion*), as in Sottoriva et al. (2015), Williams et al. (2016, 2018). The process of cell division can be described by two stochastic events with mass action rate functions f_i

$$\begin{array}{cccc} \texttt{(divide)} & x \to 2\,x & f_1(\mathbf{x}) = \lambda(1+s)x \\ \texttt{(die)} & x \to \emptyset & f_2(\mathbf{x}) = \beta x. \end{array}$$

Both events are linear in x, and the total *exit rate* of the process in state $\mathbf{x}$ is

$$a_0(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) = \big[\lambda(1+s) + \beta\big]x.$$

Since we count cells, the process state is always a discrete vector – this leads to a Continuous Time Markov Chain model, with absorbing state $\mathbf{0}$. In the general case with k clones states are elements from $\mathbb{N}^k$; with a single population the state reduces to a scalar $\mathbf{x} \in \mathbb{N}$. The division event (`divide`) depends on the baseline cell division rate $\lambda > 0$, with s the *selection coefficient* of the clone. A value of $s = 0.2$ represents a 120% increased fitness, relative to baseline; here this means a 20% faster growth, but in alternative models this could relate to a 20% increase survival rate for the offspring. In this model s can predict the outcome of competitions among clones with different fitness values (Antal and Krapivsky 2010; Khan et al. 2018).

The overall model is characterised by the conditional density function $p(\mathbf{x}, t|\mathbf{x}_0, t_0)$ – the probability of being in state $\mathbf{x}$ at time t, given the initial start in state $\mathbf{x}_0$ at time t_0 – which obeys the ordinary differential equation

$$\begin{aligned} \partial_t p(\mathbf{x}, t|\mathbf{x}_0, t_0) =& p(\mathbf{x}-1, t|\mathbf{x}_0, t_0)a_1(\mathbf{x}-1) + p(\mathbf{x}+1, t|\mathbf{x}_0, t_0)a_2(\mathbf{x}+1) \\ & -p(\mathbf{x}, t|\mathbf{x}_0, t_0)\,(a_1(\mathbf{x}) + a_2(\mathbf{x})). \end{aligned} \tag{1}$$

This is the equation for the general case $\mathbf{x} > 1$; the special case with $\mathbf{x} = 1$ where there are no jumps from the $\mathbf{0}$-absorbing state is discussed in Williams et al. (2018).

Analytical solutions to this master equation equation can be obtained as shown in Williams et al. (2018) within the context of cancer, and even more in general cases in an earlier book by Bailey (1990). Samples from the density function (i.e., realisations of the process) that solves this equation can be obtained through the Gillespie approach (Gillespie 1977) – i.e., if the chain is in state $\mathbf{x}$ at time t we can determine *when* the next reaction event will happen, and *what* that will be:

- the time-to-event lag τ follows an exponential density $\lambda \sim \text{Exp}(a_0(\mathbf{x}))$;
- the conditional density that that event is $j = \{1, 2\}$ follows $j \sim f_j(\mathbf{x})/a_0(\mathbf{x})$.

3.1.2 Including genotypes and mutations

To complete a cancer model we need to add *cell genotypes*, on top of clonal expansions, allowing for cells to simulate intra-tumour heterogeneity. We read out genotypes from sequencing, which we can be abstractly model as binary vectors $g \subseteq \{0,1\}^r$ indicating presence or absence of r mutations. Every cell division, a daughter cell inherits the genotype g from his ancestor, plus w new random mutations (w new entries 1 in g). The process is recursive and triggers every cell division (notice that the genotype per se does not affect the jump rates, which depend only on fitness). The total number of newly acquired mutations w follow a Poisson density with constant rate $\mu > 0$, $w \sim \text{Poisson}(\mu) \equiv p(w|\mu)$.

3.1.3 Generalisation with multiple clones

The overall process contains cell division, death, and genotypes, to distinguish cells; it needs to be extended with clones. For every clone, the selection coefficient s is constant through time; in a monoclonal expansion all w mutations are neutral (Williams et al. 2016). To model clones that experience either positive or negative selection, we need to allow for a somatic mutation to change the value of s, the fitness coefficient. This can be done by introducing a probability $\eta > 0$ of a new mutation to be a driver; this event is independent of the probability of a new mutation, so the joint density of sampling one driver among w mutations is

$$p(1 \text{ driver among } w \text{ new mutations}) = \eta(1-\eta)^{w-1} p(w|\mu)$$

where $p(w|\mu)$ is the Poisson density described above.

For any new driver event we sample a new $\hat{s}$ for the associated cell. In practice, one can set $s_0 = 0$ for the initial selection coefficient of the tumour-initiating cell, so that every new subclone with $s > 0$ is s percent more fit than baseline. The actual distribution to draw values for s is subject to modelling (Williams et al. 2020; Zapata et al. 2020). The new clone can enjoy either positive ($\hat{s} > s$, which triggers a new subclonal expansion) or negative ($\hat{s} < s$, which triggers depletion of the new population) selection. The new selection coefficient $\hat{s}$ also determines the speed of these dynamics – i.e., how long it takes for the subclone to colonise the overall population, or go extinct. For practical purposes one can model the prevalence of a clone as a density, and assume that very small clones (e.g., <1% of the overall mass) are too small to detect by sequencing. Mathematically, in the long run, the predicted distribution for this process is totally concentrated towards states where a single clone survives.

Some simulators are already available to generate samples from this stochastic process as well as from its spatial extensions (Heide, Webpage: https://github.com/T-Heide/TEMULATOR 2020 (accessed December 6, 2020; Chkhaidze et al. 2019). The utility of these models is that they also include a data-generation process that mimicks the effect of other confounders observed in real data (e.g., tumour purity and sequencing coverage, which can make this analysis more complicated but are not discussed here).

3.2 Clonal deconvolution from bulk sequencing

Bulk WGS can be used to detect subclonal expansions that are currently ongoing in a tumour sample (Turajlic et al. 2019); this means determining colours for the tree in Figure 1. From a single patient we can only identify ongoing expansions of the populations competing for fixation. We cannot infer all past clonal dynamics and, for instance, resolve the order of clonal mutations whose expansions are already fixated in the current sample.

The deconvolution problem is approached by using VAF or CCF values from read counts data, ideally from WGS (Dentro et al. 2017). This problem draws on some popular clustering models in the context of Machine Learning, usually parametric and non-parametric formulations of Dirichlet mixtures (Bishop 2006). The intuition is that mutations with similar allele frequencies are likely to co-occur in the same cell. The mixture density has the general form

$$p(\mathbf{x}|\pi, \boldsymbol{\theta}) = \sum_{i=i}^{k} \pi_i f(\mathbf{x}|\boldsymbol{\theta}),$$

where $\sum_{i=i}^{k} \pi_i = 1$. For the likelihood of each component, we can model read counts by Binomial or Beta-Binomial distributions, or we can model VAF or CCF values by Beta or Gaussian distributions (Caravagna et al. 2020; Deshwar et al. 2015; Miller et al. 2014; Nik-Zainal et al. 2012; Roth et al. 2014). For example, if we use VAF and Binomials we have

$$f(\mathbf{x}|\theta) = \text{Bin}(n_\ell | d_\ell, v_\ell).$$

The model for n mutations $\mathbf{x}$ has usually a $n \times k$ latent variable matrix $\mathbf{z}$ that assign mutations to clusters. From the output clustering assignments we compose cancer clones, assembling a *clone tree* via the *pigeonhole principle*.[1] We can identify genotypes percolating on a clone tree that represents ancestral relations.

While this approach is neat, the joint presence of neutral within-clone dynamics and positive selection requires caution[2] in the interpretation of these clusters (Caravagna et al. 2020). In practice, we are neglecting non-functional ITH from the overall picture. Some earlier work on the master equation for a stochastic Luria–Delbrück model of bacterial growth can be adapted to cancer, providing insights for the shape of the ITH signal – see (Kessler and Levine 2013) for a nice recap on those results, as well as some new findings. The large population solution for the probability of having m mutants at time t follows a fat-tail Landau distribution

$$p(m) = \frac{1}{\mu N} f_{\text{Landau}}\left(\frac{m}{\mu N} - \log \mu N + \gamma - 1\right),$$

where N is population size, μ the mutation rate and γ a constant. The asymptotic behavior of f_{Landau} can be approximated as the inverse squared of m, which is a power-law[3] model for neutral growth (Caravagna et al. 2020; Williams et al. 2016, 2018) (Figure 2). By this reason one can include a power-law density (f_{PowerLaw}) on top of a Dirichlet mixture model with $k-1$ Betas (f_{Beta}), i.e.,

$$p(\mathbf{x}|\pi, \boldsymbol{\theta}) = \pi_* f_{\text{PowerLaw}}(\mathbf{x}|\boldsymbol{\theta}) + \sum_{i=1}^{k} \pi_i f_{\text{Beta}}(\mathbf{x} \mid \boldsymbol{\theta}),$$

where $\sum_{i=i}^{k} \pi_i = 1 - \pi_*$, *de facto* integrating a model – the power law – to perform tumour subclonal deconvolution (Caravagna et al. 2020). The advantage of this model is that it can retrieve, from the fit distributions, tumour features such as the mutation rate and the age of the identified subclones (see Williams et al. (2018) for the derivation of these quantities).

In general, it is possible to extend these models to consider multiple spatially separated biopsies of the same tumour, so we can measure ITH in space by detecting clusters of alleles that move differently across biopsies. In this case to model the data we use multivariate Binomial distributions, where each dimension corresponds to one of the biopsies – therefore alleles that move have distinct Binomial parameters. If there are w dimensions the likelihood is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{w} \text{Bin}\left(n_{\ell,i} \middle| d_{\ell,i} v_{\ell,i}\right).$$

where the data is now indexed by each biopsy, and the dimensions are assumed to be independent. Note that

---

1 This is a constructive combinatorial principle to assemble a tree: it states that the total CCFs for the descendants of a node, cannot be larger than the CCF of the node (Deshwar et al. 2015; Nik-Zainal et al. 2012).

2 This is discussed in detail in Caravagna et al. (2020). The key point is the polyphyletic nature of neutral tails, which invalidates the implicit assumption that mutations at the same frequency co-occur in the same cell. For instance, in Figure 2 yellow and blue mutations that are present in 50% of the sequenced cells originate into two separate branches of the tree.

3 A type-I Pareto density that can capture $p(m)$ follows the generic density

$$f(x|\alpha, x_\star) = \alpha x_\star^\alpha \Big/ \left(x^{\alpha+1}\right),$$

where $x_\star$ is the scale value, i.e., the value such that $f(x|\alpha, x_*) = 0$ for $x < x_\star$, and α is the shape of the power law.

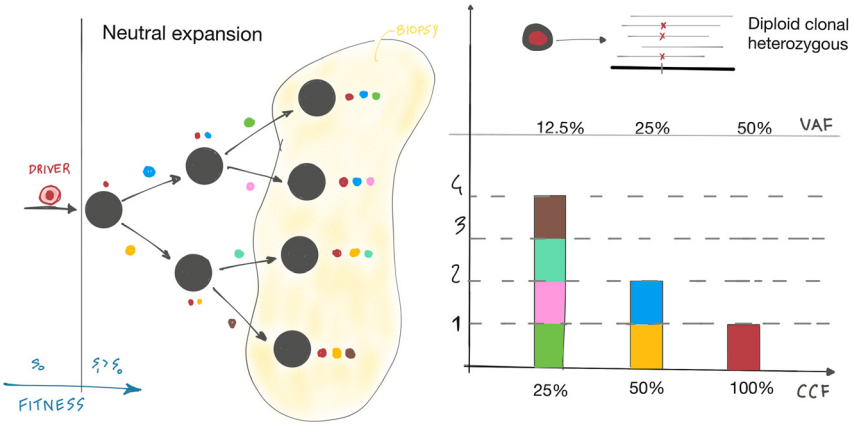

Figure 2: Intra-clone evolutionary dynamics are neutral and can be mathematically modelled (with exact solutions for the steady-state distributions for the process). If we neglect drift, the tree of cell divisions is balanced –in the plot we annotate on tree branches the mutations that accumulate. If we find the proportion of tumour cells that harbour a set of mutations, we can infer the cell lineage history (what mutation happened first). Note in this case the scaling power law for the Cancer Cell Fraction (CCF) and corresponding Variant Allele Frequency (VAF) values, assuming a diploid cancer genome.

even if the dimensions are independent, in the latent variables of the model each input point is assigned to a cluster after considering the product likelihood on all dimensions. The introduction of multivariate extensions in the overall picture however leaves a number of open problems: in particular, it turns out that many multivariate distributions are equally observable in cases with genuine spatially-measurable positive selection, or in cases that are actually neutral, but affected by strong spatial sampling biases. A detailed discussion of the role of spatial sampling in the deconvolution has been recently presented in Caravagna et al. (2020).

3.2.1 Software for deconvolution

In the last years many software tools have been developed for deconvolution from bulk sequencing – see Rosenthal et al. (2017) for a review. One of the most famous is PyClone, which can integrate clonal copy numbers and tumour purity on top of somatic VAFs obtained from deep sequencing assays, in a Bayesian model learnt by Markov Chain Monte Carlo (Roth et al. 2014). Similar to PyClone is DPclust, which leverages WGS to circumvent the need for deep sequencing and allows mutations to reside on subclonal copy numbers – i.e., CNAs that appear in a subset of cells in the whole tumour. An interesting alternative, which also applies a Bayesian clustering method but restricts only to mutation data is SciClone, which uses a variational approximation to the posterior parameters (Miller et al. 2014). There are also many other methods that approach the problem from other technical angles, e.g., via integer linear programming and the like; we refer to Rosenthal et al. (2017) for a broad review of these tools.

Importantly, all the above mentioned tools approach the deconvolution problem for the perspective of Binomial, Beta-Binomial or Gaussian mixtures, disregarding the power law model presented in this review. As of today, the only tool that integrates both perspectives is MOBSTER (Caravagna et al. 2020).

4 Conclusions

The field of Cancer Evolution has recently emerged, in which we seek to model tumour growth and response to therapy through the lens of evolution. From several NGS technologies we seek to extract the evolutionary trajectories that describe a cancer. In this review, we have summarised some of the basic principles

underpinning the application of mathematical modelling to clonal evolution, the conceptual framework upon which we can study tumour growth. Also, we have provided an overview of the inferential approach that can be used to learn, from cancer DNA sequencing data, cancer clonal dynamics.

In this respect, many new technologies promise to deliver new measurements which we can use to study tumour evolution. Among these we want to mention both single-cell technologies (Gawad et al. 2016; Navin 2015), with their multiomics extensions (Chappell et al. 2018; Macaulay et al. 2017), and spatial sequencing approaches (Burgess 2019; Ståhl et al. 2016). Multi-omics assays can probe multiple molecules from the same cell – e.g., the DNA and the RNA – and pose challenges for data integration and mathematical modelling (Colomé-Tatché and Theis 2018; Nam et al. 2021; Stuart and Satija 2019). In multi-omics data we have multiple measurements for the same cell, and the integration seems conceptually more intuitive; in many cases, however, we have multiple data types generated from different cells, and an explicit integration has to be carried out (Argelaguet et al. 2018, 2020; Campbell et al. 2019; Milite and Caravagna, In preparation 2021). Spatial extensions for single-cell RNA sequencing are also extremely important to identify patterns of spatial transcriptomics. These new technologies allow, to a different degree, to probe the cancer RNA profiles from tissue slices, obtaining a measurement associated with a geographical position (Vickovic et al. 2019). This information makes it easier to decouple the spatial signal, since that becomes explicitly associated to a location. While still in their infancy, these technologies can be used to understand spatial patterns of cell segregation and localised drug response (Berglund et al. 2018; Moncada et al. 2018).

With those data becoming increasingly available, there is the need of extending the approaches mentioned in this review in order to accommodate specificities of the new data (see (Lähnemann et al. 2020) for a very recent review on these topics).

Acknowledgments: I wish to thank Guido Sanguinetti for inviting me to write this review, and Marc Williams for useful discussions on branching process modelling.
Author contributions: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. Nature 500: 415–421.

Antal, T. and Krapivsky, P. (2010). Exact solution of a two-type branching process: clone size distribution in cell division kinetics. J. Stat. Mech. Theor. Exp. 2010: P07028.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-omics factor analysis a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. 14: e8124.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J.C., and Stegle, O. (2020). Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 21: 1–17.

Bailey, N.T. (1990). *The elements of stochastic processes with applications to the natural sciences*, 25. John Wiley & Sons, New York.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. (2007). *Conjunctive Bayesian networks*. International Statistical Institute (ISI) and the Bernoulli Society for Mathematical Statistics and Probability, JSTOR, pp. 893–909.

Beerenwinkel, N., Schwarz, R.F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: mathematical models and computational inference. Syst. Biol. 64: e1–e25.

Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., et al. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nat. Commun. 9: 1–13.

Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. CA Cancer J. Clin. 69: 127–157.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer, New York.

Burgess, D.J. (2019). Spatial transcriptomics coming of age. Nat. Rev. Genet. 20: 317.

Campbell, K.R., Steif, A., Laks, E., Zahn, H., Lai, D., McPherson, A., Farahani, H., Kabeer, F., O'Flanagan, C., Biele, J., et al. (2019). Clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. Genome Biol. 20: 1–12.

Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., Moreno, V., Antoniotti, M., and Mishra, B. (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. PNAS 113: E4025–E4034.

Caravagna, G., Giarratano, Y., Ramazzotti, D., Tomlinson, I., Graham, T.A., Sanguinetti, G., and Sottoriva, A. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. Nat. Methods 15: 707.

Caravagna, G., Heide, T., Williams, M.J., Zapata, L., Nichol, D., Chkhaidze, K., Cross, W., Cresswell, G.D., Werner, B., Acar, A., et al. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. Nat. Genet. 52: 898–907.

Chappell, L., Russell, A.J., and Voet, T. (2018). Single-cell (multi) omics technologies. Annu. Rev. Genom. Hum. Genet. 19: 15–41.

Chkhaidze, K., Heide, T., Werner, B., Williams, M.J., Huang, W., Caravagna, G., Graham, T.A., and Sottoriva, A. (2019). Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLoS Comput. Biol. 15: e1007243.

Christopher, M.J., Petti, A.A., Rettig, M.P., Miller, C.A., Chendamarai, E., Duncavage, E.J., Klco, J.M., Helton, N.M., O'Laughlin, M., Fronick, C.C., et al. (2018). Immune escape of relapsed AML cells after allogeneic transplantation. NEJM 379: 2330–2341.

Colomé-Tatché, M. and Theis, F.J. (2018). Statistical single cell multi-omics integration. Curr. Opin. Syst. Biol. 7: 54–59.

Dentro, S.C., Wedge, D., and Van Loo, P. (2017). Principles of reconstructing the subclonal architecture of cancers. Cold Spring Harb. Perspect. Med. 7: a026625.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491.

Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., and Morris, Q. (2015). Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 16: 1–20.

Diaz-Uriarte, R. (2018). Cancer progression models and fitness landscapes: a many-to-many relationship. Bioinformatics 34: 836–844.

Diaz-Uriarte, R. and Vasallo, C. (2019). Every which way? On predicting tumor evolution using cancer progression models. PLoS Comput. Biol. 15: e1007246.

Durrett, R. (2002). Basic models. In: *Probability models for DNA sequence evolution*. Springer, New York, pp. 1–66.

Ewens, W.J. (2012). *Mathematical population genetics 1: theoretical introduction*, 27. Springer Science & Business Media, New York.

Fusco, D., Gralka, M., Kayser, J., Anderson, A., and Hallatschek, O. (2016). Excess of mutational jackpot events in expanding populations revealed by spatial Luria–Delbrück experiments. Nat. Commun. 7: 12760.

Gatenby, R.A., Silva, A.S., Gillies, R.J., and Frieden, B.R. (2009). Adaptive therapy. Canc. Res. 69: 4894–4903.

Gawad, C., Koh, W., and Quake, S.R. (2016). Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. 17: 175.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81: 2340–2361.

Greaves, M. and Maley, C.C. (2012). Clonal evolution in cancer. Nature 481: 306–313.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158.

Heide, T. (2020). The TEMULATOR package to generate synthetic tumour sequencing data, Webpage. https://github.com/T-Heide/TEMULATOR (Accessed 6 December 2020).

Househam, J., Cross, W.C., and Caravagna, G. (2021). An automated quality checking tool for clonal copy number changes and single nucleotide variant calls from whole genome sequencing data. In preparation.

Kessler, D.A. and Levine, H. (2013). Large population solution of the stochastic Luria–Delbrück evolution model. PNAS 110: 11682–11687.

Kessler, D.A. and Levine, H. (2015). Scaling solution in the large population limit of the general asymmetric stochastic Luria–Delbrück evolution process. J. Stat. Phys. 158: 783–805.

Khan, K.H., Cunningham, D., Werner, B., Vlachogiannis, G., Spiteri, I., Heide, T., Mateos, J.F., Vatsiou, A., Lampis, A., Damavandi, M.D., et al. (2018). Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial. Canc. Discov. 8: 1270–1285.

Kimura, M. (1994). *Population genetics, molecular evolution, and the neutral theory: selected papers*. University of Chicago Press, Chicago, IL, United States.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biol. 21: 1–35.

Lakatos, E., Williams, M.J., Schenck, R.O., Cross, W.C., Househam, J., Zapata, L., Werner, B., Gatenbee, C., Robertson-Tessi, M., Barnes, C.P., et al. (2020). Evolutionary dynamics of neoantigens in growing tumors. Nat. Genet. 52: 1057–1066.

Macaulay, I.C., Ponting, C.P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. Trends Genet. 33: 155–168.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal patterns of selection in cancer and somatic tissues. Cell 171: 1029–1041.

McGranahan, N. and Swanton, C. (2017). Clonal heterogeneity and tumor evolution: past, present, and the future. Cell 168: 613–628.

Milite, S. and Caravagna, G. (2021). Genotyping copy number alterations from single-cell RNA sequencing of cancer cells. In preparation.

Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S., Griffith, O.L., Vij, R., Tomasson, M.H., Graubert, T.A., Walter, M.J., et al. (2014). Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS Comput. Biol. 10: e1003665.

Moncada, R., Wagner, F., Chiodin, M., Devlin, J.C., Baron, M., Hajdu, C.H., Simeone, D.M., and Yanai, I. (2018). Building a tumor atlas: integrating single-cell RNA-Seq data with spatial transcriptomics in pancreatic ductal adenocarcinoma. bioRxiv 254375, https://doi.org/10.1101/254375.

Nam, A.S., Chaligne, R., and Landau, D.A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat. Rev. Genet. 22: 3–18.

Navin, N.E. (2015). The first five years of single-cell cancer genomics and beyond. Genome Res. 25: 1499–1507.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012). The life history of 21 breast cancers. Cell 149: 994–1007.

Nowak, M.A. (2006). *Evolutionary dynamics: exploring the equations of life*. Harvard University Press, Cambridge, MA.

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

Pich, O., Muiños, F., Lolkema, M.P., Steeghs, N., Gonzalez-Perez, A., and Lopez-Bigas, N. (2019). The mutational footprints of cancer therapies. Nat. Genet. 51: 1732–1740.

Ramazzotti, D., Caravagna, G., Loohuis, L.O., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., and Mishra, B. (2015). Capri: efficient inference of cancer progression models from cross-sectional data. Bioinformatics 31: 3016–3026.

Ramazzotti, D., Lal, A., Liu, K., Tibshirani, R., and Sidow, A. (2019). De novo mutational signature discovery in tumor genomes using sparse signatures. bioRxiv 384834.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 17: 1–11.

Rosenthal, R., McGranahan, N., Herrero, J., and Swanton, C. (2017). Deciphering genetic intratumor heterogeneity and its impact on cancer evolution. Annu. Rev. Cell Biol. 1: 223–240.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S.P. (2014). Pyclone: statistical inference of clonal population structure in cancer. Nat. Methods 11: 396–398.

Schumacher, T.N. and Schreiber, R.D. (2015). Neoantigens in cancer immunotherapy. Science 348: 69–74.

Shackleton, M., Quintana, E., Fearon, E.R., and Morrison, S.J. (2009). Heterogeneity in cancer: cancer stem cells versus clonal evolution. Cell 138: 822–829.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., et al. (2015). A big bang model of human colorectal tumor growth. Nat. Genet. 47: 209–216.

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353: 78–82.

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. 20: 257–272.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26: 119–164.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. (2020). Pan-cancer analysis of whole genomes. Nature 578: 82.

Toffalori, C., Zito, L., Gambacorta, V., Riba, M., Oliveira, G., Bucci, G., Barcella, M., Spinelli, O., Greco, R., Crucitti, L., et al. (2019). Immune signature drives leukemia escape and relapse after hematopoietic cell transplantation. Nat. Med. 25: 603–611.

Topol, E.J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25: 44–56.

Tung, H.-R., and Durrett, R. (2020). Signatures of neutral evolution in exponentially growing tumors: a theoretical perspective. bioRxiv 1–12.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J.I., Nicol, D., O'Brien, T., Larkin, J., Horswell, S., et al. (2018a). Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal. Cell 173: 581–594.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Horswell, S., Chambers, T., O'Brien, T., Lopez, J.I., Watkins, T.B., Nicol, D., et al. (2018b). Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. Cell 173: 595–610.

Turajlic, S., Sottoriva, A., Graham, T., and Swanton, C. (2019). Resolving genetic heterogeneity in cancer. Nat. Rev. Genet. 20: 404–416.

Vago, L., Perna, S.K., Zanussi, M., Mazzi, B., Barlassina, C., Stanghellini, M.T.L., Perrelli, N.F., Cosentino, C., Torri, F., Angius, A., et al. (2009). Loss of mismatched HLA in leukemia after stem-cell transplantation. NEJM 361: 478–488.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods 16: 987–990.

Williams, M., Werner, B., Heide, T., Curtis, C., Barnes, C., Sottoriva, A., and Graham, T. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. Nat. Genet. 50: 895.

Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. Nat. Genet. 48: 238–244.

Williams, M.J., Zapata, L., Werner, B., Barnes, C.P., Sottoriva, A., and Graham, T.A. (2020). Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dn/ds ratios. Elife 9: e48714.

Zapata, L., Pich, O., Serrano, L., Kondrashov, F.A., Ossowski, S., and Schaefer, M.H. (2018). Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. Genome Biol. 19: 1–17.

Zapata, L., Caravagna, G., Williams, M., Lakatos, E., Abdul-Jabbar, K., Werner, B., Graham, T.A., and Sottoriva, A. (2020). dN/dS dynamics quantify tumour immunogenicity and predict response to immunotherapy. bioRxiv 1–41.