Counting unique molecular identifiers in sequencing using a multi-type branching process with immigration

Journal of Theoretical Biology 558 (2023) 111365

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Counting unique molecular identifiers in sequencing using a multi-type branching process with immigration

Serik Sagitov [a,*], Anders Ståhlberg [b,c,d]

[a] Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden
[b] Sahlgrenska Center for Cancer Research, Department of Laboratory Medicine, Institute of Biomedicine, University of Gothenburg, Sweden
[c] Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Sweden
[d] Region Västra Götaland, Sahlgrenska University Hospital, Department of Clinical Genetics and Genomics, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Sequencing
PCR amplification rate
Unique Molecular Identifier
Tree-bookkeeping
PCR branching process
Growing immigration

ABSTRACT

Detection of extremely rare variant alleles, such as tumor DNA, within a complex mixture of DNA molecules is experimentally challenging due to sequencing errors. Barcoding of target DNA molecules in library construction for next-generation sequencing provides a way to identify and bioinformatically remove polymerase induced errors. During the barcoding procedure involving t consecutive PCR cycles, the DNA molecules become barcoded by Unique Molecular Identifiers (UMIs). Different library construction protocols utilize different values of t. The effect of a larger t and imperfect PCR amplifications in relation to UMI cluster sizes is poorly described.

This paper proposes a branching process with growing immigration as a model describing the random outcome of t cycles of PCR barcoding. Our model discriminates between five different amplification rates r_1, r_2, r_3, r_4, r for different types of molecules associated with the PCR barcoding procedure. We study this model by focussing on C_t, the number of clusters of molecules sharing the same UMI, as well as $C_t(m)$, the number of UMI clusters of size m. Our main finding is a remarkable asymptotic pattern valid for moderately large t. It turns out that $\mathrm{E}(C_t(m))/\mathrm{E}(C_t) \approx 2^{-m}$ for $m = 1, 2, \dots$, regardless of the underlying parameters (r_1, r_2, r_3, r_4, r). The knowledge of the quantities C_t and $C_t(m)$ as functions of the experimental parameters t and (r_1, r_2, r_3, r_4, r) will help the users to draw more adequate conclusions from the outcomes of different sequencing protocols.

1. Introduction

Massive parallel sequencing is implemented in a wide range of applications within basic research and for clinical practice. Numerous protocols and technologies are developed to accurately detect and quantify differences in molecular sequences and detect variants. In cancer management, sequencing is applied in diagnostics to identify mutations that can be targeted with specific therapies. The standard massive parallel sequencing techniques can detect variants with frequencies down to the range of 1%–5% (Stead et al., 2013; Xu et al., 2014). However, this sensitivity is not sufficient for several emerging applications. For example, detection of circulating tumor-DNA in liquid biopsies requires technologies that have the ability to detect variants with frequencies lower than 0.1% in clinically relevant samples (Heitzer et al., 2019; Andersson et al., 2021; Ignatiadis et al., 2021). The main source of sequencing noise is due to the polymerase induced errors that occur during library construction and sequencing (Filges et al., 2019).

To reduce amplification bias and polymerase induced errors, Unique Molecular Identifiers (UMIs), also known as DNA barcodes, can be used to enable ultra-sensitive sequencing (Kinde et al., 2011; Kivioja et al., 2011). Nowadays, UMIs can be applied to all types of nucleic acids: for the RNA analysis, the UMIs are normally introduced during the reverse transcription step, and for the DNA analysis, UMIs are either introduced by an initial DNA ligation step or by target PCR (Andersson et al., 2021). For PCR based sequencing methods, the UMIs typically consist of 8–12 randomized nucleotides that are experimentally attached to each target DNA molecule through a limited number of PCR cycles followed by a general amplification step (Fig. 1). Compared to the majority of other sequencing methods, PCR-based approaches can introduce variable number of UMIs by using different numbers of barcoding PCR cycles. After sequencing, all reads with the same UMI can be tracked back to the original DNA molecule, allowing to control the polymerase induced errors and quantification biases.

Experimentally it is challenging to introduce UMIs in PCR based approaches, since randomized sequences easily produce non-specific PCR products. To address this challenge, several barcoding PCR cycles

* Corresponding author.
E-mail addresses: serik@chalmers.se (S. Sagitov), anders.stahlberg@gu.se (A. Ståhlberg).

https://doi.org/10.1016/j.jtbi.2022.111365
Received 5 June 2022; Received in revised form 14 September 2022; Accepted 16 November 2022
Available online 19 November 2022

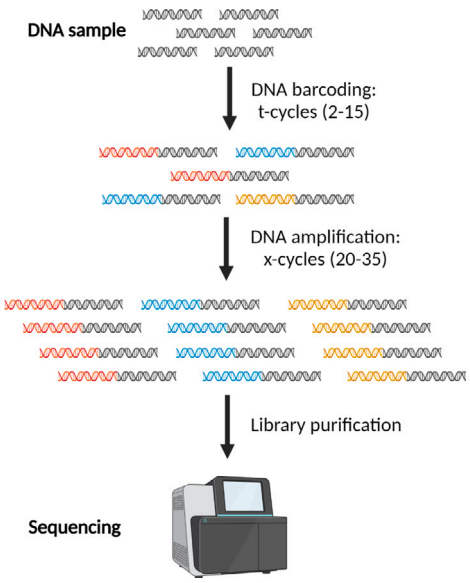

Fig. 1. Overview of an ultra-sensitive sequencing protocol (the figure is created using BioRender.com). The experimental workflow consists of two PCR steps. In the first barcoding PCR step, the UMIs are attached to target DNA molecules during t cycles of amplification. In the second adapter PCR step, sequencing adapters are attached to the barcoded DNA sequences during x cycles of amplification. Typical numbers for the parameters t and x are shown. Finally, libraries are purified and sequenced.

Table 1
The UMI cluster numbers for the perfect PCR amplifications.

Cluster size	1	2	3	4	5	Total
$t=2$	2	0	0	0	0	2
$t=3$	4	2	0	0	0	6
$t=4$	8	4	2	0	0	14
$t=5$	16	8	4	2	0	30
$t=6$	32	16	8	4	2	62

can be applied, which simplifies the experimental protocol (Ståhlberg et al., 2016; Cohen et al., 2018). However, the number of different UMIs and their distribution is not easily estimated with the increasing number of barcoding PCR cycles, limiting the use of UMIs in different applications.

Fig. 2 depicts the outcome of three perfectly successful barcoding PCR cycles. The target double-stranded DNA molecule is placed at the top level, $t=0$ (for each double-stranded molecule, the upper segment represents a single-stranded molecule in the direction $5'\to 3'$ and the lower segment represents a single-stranded molecule in the direction $3'\to 5'$). According to Fig. 2, the first PCR cycle produces two double-stranded molecules shown at the level $t=1$: the left pair, consisting of the target sense molecule plus the antisense molecule with a reverse primer, and the right pair, consisting of the target antisense molecule plus the sense molecule with a forward UMI primer. All four single stranded molecules at $t=1$ are incomplete with one generated UMI. The complete molecules, that can be amplified in the second adapter PCR step (Fig. 1), start appearing at the level $t=2$. With perfect PCR amplifications, the sizes of UMI clusters grow geometrically as shown by Table 1. For example at $t=3$, as illustrated on Fig. 2, there are six UMI clusters labelled by A, B, C, D, E, F, with the clusters B and D having size two, and the clusters A, C, E, F being singletons.

In practice, PCR amplifications are imperfect and a single-stranded molecule is successfully amplified with a probability r, i.e., the amplification rate. Despite the attempts to optimize the primers and reaction conditions, the PCR amplification rates are rarely close to 1. With imperfect amplifications, Fig. 2 and Table 1 should be modified to reflect the fact that the outcome of each PCR cycle is random. Fig. 3 illustrates one possible realization of six barcoding PCR cycles, summarized by

Table 2
The UMI cluster numbers of imperfect PCR cycles depicted in Fig. 3.

Cluster size	1	2	3	4	5	Total
$t=2$	0	0	0	0	0	0
$t=3$	1	0	0	0	0	1
$t=4$	4	0	0	0	0	4
$t=5$	4	2	0	0	0	6
$t=6$	6	1	2	0	0	9

Table 2. In particular, at $t=6$, as shown on the bottom rows of Fig. 3 and Table 2, A and D form UMI clusters of size 3, B is a UMI cluster of size 2, and the clusters C, E, F, G, H, I are singletons.

The PCR amplification rate is dependent on both sequence context and sample quality. In particular, the target DNA molecules are often long, containing thousands or even millions base pairs. Moreover, some molecules may have inhibitors attached to the DNA. As a rule, the amplification rates of the original molecules are smaller compared to the later formed DNA molecules. Addressing these features, our model discriminates between five amplification rates (r_1, r_2, r_3, r_4, r) as indicated by Figs. 2 and 3, assuming that

$$0 < r_1, r_2 \le r_3, r_4 \le r \le 1. \tag{1}$$

Here, r_1 and r_2 represent the amplification rates of the original sense and antisense DNA molecules, respectively. The rates r_1 and r_2 may differ, as their sequence contexts are different, since two complementary sequences have different nucleotide sequences. The parameters r_3 and r_4 refer to the amplification rates of the DNA molecules that have been amplified once. Again, r_3 and r_4 may be different since they have different sequence contexts. We assume that both r_3 and r_4 are larger than each of the rates r_1 and r_2, since the DNA molecules become shorter after being amplified once. Finally, we assume that the shortest amplicon is amplified with the highest rate r, being the same for the sense and antisense strands.

This paper introduces and studies a mathematical model for the outcome of the barcoding PCR experiment starting from a single double-stranded DNA molecule. Our stochastic model is defined in terms of the five parameters (r_1, r_2, r_3, r_4, r) using the framework of the multi-type Galton–Watson processes (Haccou et al., 2005). The use of branching processes as a stochastic model for counting the molecules in the repeated PCR amplification cycles is well established in the literature (Krawczak et al., 1989; Kimmel and Alexrod, 2002; Jagers and Klebaner, 2003; Haccou et al., 2005; Lalam, 2006). In Pflug and von Haeseler (2018), the branching process approach is applied to the second, adapter PCR step mentioned in Fig. 1. However, to our knowledge, the use of branching processes is novel for modelling the barcoding PCR step. The multi-type branching process with immigration of this paper is a special example of the Galton–Watson processes with neutral mutations examined in Bertoin (2009).

2. Results

We study the possible outcomes of the barcoding PCR step, starting from a single double-stranded DNA molecule. Our stochastic model for counting the unique UMIs is built upon an efficient bookkeeping system for the DNA barcoding procedure, presented in Section 2.1. We discriminate between six different types of single-stranded molecules emerging during the barcoding PCR procedure and introduce a multi-type Galton–Watson process with immigration describing the random process of reproduction of the single-stranded molecules, see Section 2.2.

We study this model by focussing on C_t, the number of clusters of molecules sharing the same UMI, as well as $C_t(m)$, the number of UMI clusters of size m. The underlying branching properties of the model yield recursive formulas for the expected values $\mathrm{E}(C_t)$ and $\mathrm{E}(C_t(m))$, see Appendix. Our main finding, based on the analysis of the proposed Galton–Watson process, is the following asymptotic result

$$\mathrm{E}(C_t(m))/\mathrm{E}(C_t) \to 2^{-m}, \quad m \ge 1, \quad t \to \infty. \tag{2}$$

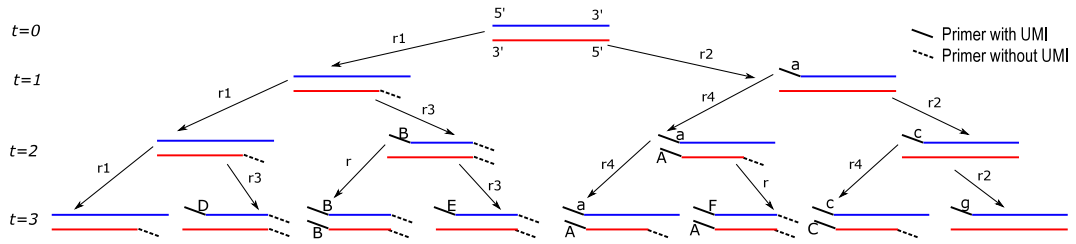

Fig. 2. Schematic representation for three cycles of barcoding PCR assuming that all fourteen PCR amplifications were successful. The two target DNA strands are shown in blue and red (sense and antisense, respectively). The forward primers with UMI and the reverse primers are shown as the solid and dashed black segments. A single-stranded molecule needs primer sequences at both ends to be complete for amplification in the downstream adapter PCR, such molecules are marked with the capital letters. The non-capitalized letters indicate incomplete molecules with different UMIs. Altered letters indicate different UMIs, while the same letter indicates an identical UMI. The amplification rates r_1, r_2, r_3, r_4, r, in general, may differ from each other.

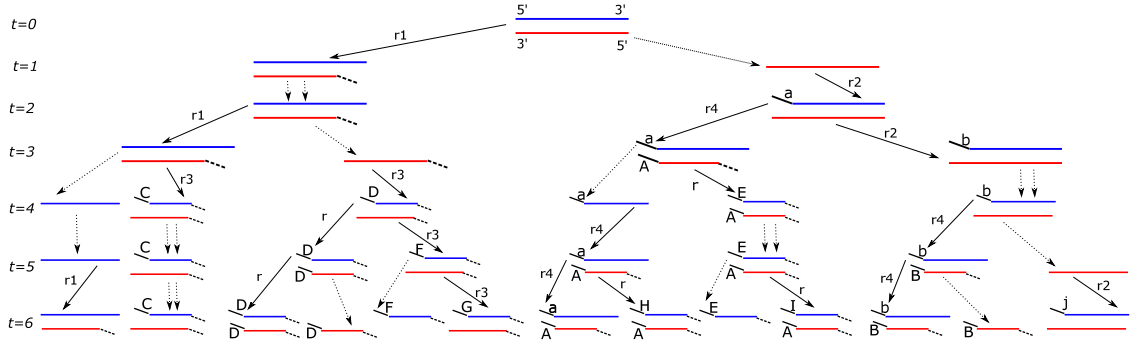

Fig. 3. A possible output of the first six cycles for the barcoding PCR step with imperfect amplifications. The failed PCR amplifications are marked by the dashed arrows and dashed double-arrows.

The remarkable feature of this relationship is that the limits are the same irrespectively of the parameter values (r_1, r_2, r_3, r_4, r). This approximation may work well already at $t = 10$, as illustrated below.

2.1. Tree-bookkeeping system for barcoding PCR

In this section we introduce a convenient bookkeeping system for the outcome of multiple barcoding PCR cycles. We start by considering the simple case of perfect PCR amplifications. Compared to the schematic representation of Fig. 2, the tree-graph approach of Fig. 4 allows us to neatly depict more than three PCR cycles using the same space. At any tree level t, the vertical branches (lineages) of the tree on Fig. 4 are pairwise connected designating the double-stranded molecules of Fig. 2.

Out bookkeeping system distinguishes between six different types of the tree lineages. At any level t there is exactly one target sense lineage labelled by 0, and one target antisense lineage labelled by 1. The other four types of lineages T_1, T_2, T_3, T_4 are generated according to the following rules:

$$0 \to 0 + T_1, \quad 1 \to 1 + T_2, \quad T_1 \to T_1 + T_3, \quad T_2 \to T_2 + T_4, \\ T_3 \to T_3 + T_4, \quad T_4 \to T_4 + T_3, \tag{3}$$

where

- the lineages $0, 1, T_1, T_2$ represent incomplete molecules,
- the lineages T_3, T_4 represent complete molecules,
- newly generated lineages T_2 and T_3 represent molecules with a novel UMI,
- the lineages T_4 represent molecules that inherit the UMI of the parental molecule.

Let $Z_t^{(i)}$ stand for the number of the T_i-lineages at the level t, for $i = 1, 2, 3, 4$ and $t \geq 0$. Given

$$(r_1, r_2, r_3, r_4, r) = (1, 1, 1, 1, 1), \tag{4}$$

the evolution of the vector $(Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}, Z_t^{(4)})$ is deterministic and in accordance with rule (3) satisfies the following recursions

$$Z_t^{(1)} = Z_{t-1}^{(1)} + 1,$$
$$Z_t^{(2)} = Z_{t-1}^{(2)} + 1,$$
$$Z_t^{(3)} = Z_{t-1}^{(3)} + Z_{t-1}^{(1)} + Z_{t-1}^{(4)},$$
$$Z_t^{(4)} = Z_{t-1}^{(4)} + Z_{t-1}^{(2)} + Z_{t-1}^{(3)},$$

valid for $t \geq 1$ under the initial condition

$$Z_0^{(1)} = Z_0^{(2)} = Z_0^{(3)} = Z_0^{(4)} = 0. \tag{5}$$

From these recursions, it is easy to see that $Z_t^{(1)} = Z_t^{(2)} = t - 1$, implying $Z_t^{(3)} = Z_t^{(4)}$. Moreover,

$$Z_t^{(3)} = 2Z_{t-1}^{(3)} + t - 2,$$

so that

$$Z_t^{(2)} + Z_t^{(3)} = 2(Z_{t-1}^{(2)} + Z_{t-1}^{(3)}) + 1,$$

yielding

$$Z_t^{(2)} + Z_t^{(3)} = 2^t - 1, \quad t \geq 0. \tag{6}$$

At any given level t, we split the set of complete lineages into the clusters of lineages representing the UMI clusters of molecules sharing the same UMI. According to our bookkeeping system, there are two different types of lineage clusters: T_2-*clusters* stemming from the T_2-lineages and T_3-*clusters* stemming from the T_3-lineages. A T_2-cluster consists of the T_4-lineages, which are the daughter lineages of the stem T_2-lineage. A T_3-cluster consists of the stem T_3-lineage and its daughter T_4-lineages. Observe a principal difference between these two types of

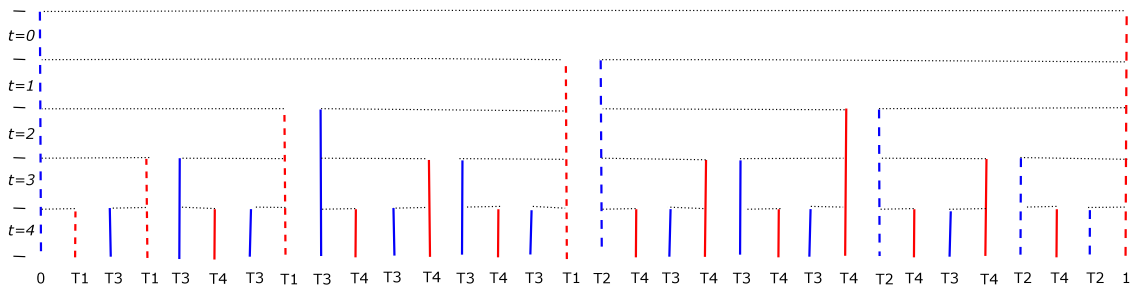

Fig. 4. A tree-graph summary of the four cycles of perfect barcoding PCR. Each vertical lineage represents a molecule that, once appeared, persists over the succeeding PCR cycles. The dashed and solid branches discriminate between the incomplete and complete molecules. Labels $0, 1, T_1, T_2, T_3, T_4$ indicate different types of the molecules with and without UMIs, where the different levels t of the tree represent the consecutive PCR cycle numbers.

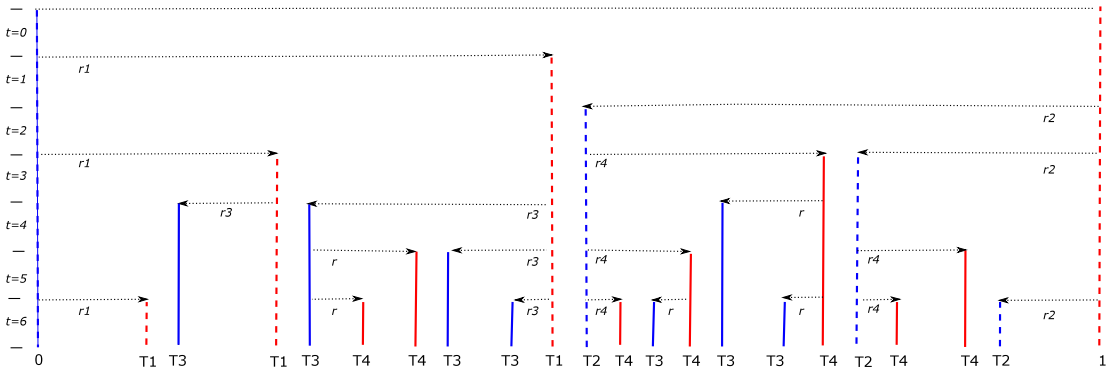

Fig. 5. The tree view of Fig. 3. Depending on the corresponding type of the underlying DNA molecule, the five amplification rates r_1, r_2, r_3, r_4, r, are assigned to the corresponding horizontal arrows.

clusters: while a T_3-cluster includes the stem T_3-lineage, the stem T_2-lineage is not a part of the corresponding T_2-cluster as it represents an incomplete molecule.

If $C_t(m)$ is the number of T_2 and T_3-clusters of size m observed at the level t of the lineage tree, then

$$C_t = C_t(1) + \cdots + C_t(t-1) \tag{7}$$

gives the total number of lineage clusters at the level t. Observe that in the special case (4), we have

$$C_t = Z_t^{(2)} + Z_t^{(3)} - 1,$$

so that by (6),

$$C_t = 2^t - 2, \quad t \geq 1. \tag{8}$$

Furthermore, under (4), we get

$$C_{t+1}(1) = C_t + 2, \quad C_{t+1}(m+1) = C_t(m), \quad m \geq 2, \quad t \geq 1. \tag{9}$$

The first equality in (9) says that each cluster at the level t produces at the next level $t+1$ one novel T_3-cluster, in addition to a new T_2-singleton plus a new T_3-cluster generated by a T_1-lineage. (By a T_2-singleton we mean a T_2-lineage that has not yet produced a daughter lineage.) The second equality in (9) says that each cluster of size m at the level t turns into a cluster of size $m+1$ at the next level $t+1$. By (8) and (9), we obtain

$$C_t(m) = 2^{t-m}, \quad m \geq 1, \quad t \geq m+1, \tag{10}$$

cf Table 1. Relations (8) and (10) imply an interesting rule of thumb saying that given (4), at any given t, the increase of the cluster size by 1 reduces the number of clusters by half. More precisely,

$$C_t(m)/C_t \to 2^{-m}, \quad m \geq 1, \quad t \to \infty. \tag{11}$$

If condition (4) does not hold, so that some PCR amplifications may fail, the cluster numbers $C_t(m)$ and C_t become random. In the Appendix, we obtain recursion relations for the corresponding mean values, from which we will derive (2), a far-reaching extension of the deterministic relation (11).

2.2. Multi-type branching process with immigration

We are going to write $\xi \sim \mathrm{Ber}(r)$ to say that the random variable ξ has a Bernoulli distribution with

$$\mathrm{P}(\xi = 1) = r, \quad \mathrm{P}(\xi = 0) = 1 - r.$$

Assuming that the amplification rates (r_1, r_2, r_3, r_4, r) satisfy (1), we restate (3) in the form

$$0 \to 0 + \xi_0 T_1, \quad T_1 \to T_1 + \xi^{(1)} T_3, \quad T_3 \to T_3 + \xi^{(3)} T_4, \tag{12}$$

$$1 \to 1 + \xi_1 T_2, \quad T_2 \to T_2 + \xi^{(2)} T_4, \quad T_4 \to T_4 + \xi^{(4)} T_3, \tag{13}$$

involving six random inputs

$$\xi_0 \sim \mathrm{Ber}(r_1), \quad \xi_1 \sim \mathrm{Ber}(r_2), \quad \xi^{(1)} \sim \mathrm{Ber}(r_3),$$
$$\xi^{(2)} \sim \mathrm{Ber}(r_4), \quad \xi^{(3)}, \xi^{(4)} \sim \mathrm{Ber}(r).$$

We clarify these relations by referring to $T_3 \to T_3 + \xi^{(3)} T_4$, which says that a T_3-lineage existing at any given level t, at the next level $t+1$ infallibly reproduces itself and produces a new lineage of type T_4 with probability r.

The sequence of random vectors

$$\{(Z_t^{(1)}, Z_t^{(2)}, Z_t^{(3)}, Z_t^{(4)}) : \quad t = 0, 1, \ldots\} \tag{14}$$

forms a Markov chain with the initial state (5). Treating relations (12) and (13) as the reproduction rules involving four types of individuals along the discrete time t, we may view this Markov chain as a multi-type branching process with immigration (Mode, 1970). This is an example of a decomposable multi-type branching process, since the types T_3 and T_4 may generate each other but not the types T_1 and T_2. There are two sources of immigration for this four-type branching

process: the 0-lineage generates the T_1-individuals at the rate r_1, and the 1-lineage generates the T_2-individuals at the rate r_2. The types T_1 and T_2 without directly communicating with each other, give rise to the types T_3 and T_4 respectively.

The reproduction rules (12) and (13) yield the following recursive relations

$$Z_t^{(1)} = Z_{t-1}^{(1)} + \xi_{t,0}, \quad Z_t^{(2)} = Z_{t-1}^{(2)} + \xi_{t,1}, \tag{15}$$

$$Z_t^{(3)} = Z_{t-1}^{(3)} + \sum_{j=1}^{Z_{t-1}^{(1)}} \xi_{t,j}^{(1)} + \sum_{j=1}^{Z_{t-1}^{(4)}} \xi_{t,j}^{(4)}, \qquad Z_t^{(4)} = Z_{t-1}^{(4)} + \sum_{j=1}^{Z_{t-1}^{(2)}} \xi_{t,j}^{(2)} + \sum_{j=1}^{Z_{t-1}^{(3)}} \xi_{t,j}^{(3)}, \tag{16}$$

involving independent Bernoulli random values

$$\xi_{t,0} \sim \mathrm{Ber}(r_1), \quad \xi_{t,1} \sim \mathrm{Ber}(r_2), \quad \xi_{t,j}^{(1)} \sim \mathrm{Ber}(r_3),$$
$$\xi_{t,j}^{(2)} \sim \mathrm{Ber}(r_4), \quad \xi_{t,j}^{(3)}, \xi_{t,j}^{(4)} \sim \mathrm{Ber}(r),$$

each indicating whether the underlying PCR amplification is successful or not.

In the current setting, the supercritical four-type branching process (14) could be described in terms of a single type branching process with a growing immigration. The specific reproduction rules (12) and (13) allow the types T_3 and T_4 to be treated as a single type, say T, such that the type T individuals produce $(1+r)$ offspring on average. In terms of the process (14), the number Z_t of T-individuals at time t can be expressed as the sum

$$Z_t = Z_t^{(3)} + Z_t^{(4)},$$

and the Markov chain $(Z_t)_{t\geq 0}$ can be treated as a branching process with growing immigration. By (16), the number of type T immigrants at time t is given by the sum $I_t = I_t^{(1)} + I_t^{(2)}$ of two independent random variables

$$I_t^{(1)} = \sum_{j=1}^{Z_{t-1}^{(1)}} \xi_{t,j}^{(1)}, \qquad I_t^{(2)} = \sum_{j=1}^{Z_{t-1}^{(2)}} \xi_{t,j}^{(2)}$$

having binomial distributions $I_t^{(1)} \sim \mathrm{Bin}(t, r_1 r_3)$ and $I_t^{(2)} \sim \mathrm{Bin}(t, r_2 r_4)$. Observe that the generating function for the number of immigrants $h_t(s) = \mathrm{E}(s^{I_t}) = \mathrm{E}(s^{I_t^{(1)}})\mathrm{E}(s^{I_t^{(2)}})$ is computed explicitly

$$h_t(s) = (1 - r_1 r_3 + r_1 r_3 s)^t (1 - r_2 r_4 + r_2 r_4 s)^t, \quad 0 \leq s \leq 1, \quad t \geq 0. \tag{17}$$

According to Theorem 2b from Section 4 of Rahimov (2021), the long term population size growth of the supercritical branching process with growing immigration is regulated by its reproduction rate $(1+r)$:

$$(1+r)^{-t} Z_t \to W \text{ in } L_2, \quad t \to \infty. \tag{18}$$

Here, the limit W is a strictly positive random variable, whose Laplace transform

$$\mathrm{E}(e^{-\lambda W}) = \prod_{k=1}^{\infty} h_k(\lambda(1+r)^{-k})$$

is determined by the five amplification rates (r_1, r_2, r_3, r_4, r) in terms of the generating functions (17) and the limiting Laplace transform $\phi(\lambda)$ for the branching process without immigration satisfying the functional quadratic equation

$$\phi((1+r)\lambda) = (1-r)\phi(\lambda) + r\phi^2(\lambda).$$

The main concern of this paper is not the decomposable multitype branching process (14) per se, but certain functionals thereof, especially the number of clusters of size m,

$$C_t(m) = X_t(m) + Y_t(m), \quad 1 \leq m \leq t-1,$$

where $X_t(m)$ is the number of T_2-clusters and $Y_t(m)$ is the number of T_3-clusters of size m at the level t. For $m \geq 1$, let

$\mathbb{Y}_t(m)$ be the set of T_3-lineages which have exactly $(m-1)$ daughter lineages of the type T_4 at the level t,

$\mathbb{X}_t(m)$ be the set of T_2-lineages which have exactly m daughter lineages of the type T_4 at the level t,

$\mathbb{X}_t(0)$ be the set of T_2-singletons at the level t.

Then by (16), we have

$$Y_t(1) = \sum_{j=1}^{Z_{t-1}^{(1)}} \xi_{t,j}^{(1)} + \sum_{j=1}^{Z_{t-1}^{(4)}} \xi_{t,j}^{(4)} + \sum_{j\in\mathbb{Y}_{t-1}(1)} (1 - \xi_{t,j}^{(3)}), \tag{19}$$

$$X_t(m) = \sum_{j\in\mathbb{X}_{t-1}(m-1)} \xi_{t,j}^{(2)} + \sum_{j\in\mathbb{X}_{t-1}(m)} (1 - \xi_{t,j}^{(2)}), \tag{20}$$

$$Y_t(m) = \sum_{j\in\mathbb{Y}_{t-1}(m-1)} \xi_{t,j}^{(3)} + \sum_{j\in\mathbb{Y}_{t-1}(m)} (1 - \xi_{t,j}^{(3)}), \tag{21}$$

where $1 \leq m \leq t-1$.

Furthermore, let $X_t(0)$ be the number of T_2-singletons at the level t. (To illustrate, the example of Fig. 5 gives $X_0(0) = X_1(0) = X_5(0) = 0$, $X_2(0) = X_3(0) = X_4(0) = X_6(0) = 1$.) Then, due to (15) and (16),

$$X_t(0) = \xi_{t,1} + \sum_{j\in\mathbb{X}_{t-1}(0)} (1 - \xi_{t,j}^{(2)}). \tag{22}$$

Since

$$Z_t^{(2)} = X_t(0) + X_t(1) + \cdots + X_t(t-1),$$

the total number of clusters at the level t equals

$$C_t = C_t(1) + \cdots + C_t(t-1) = X_t(1) + Y_t(1) + \cdots + X_t(t-1) + Y_t(t-1),$$

which entails

$$C_t = Z_t^{(2)} - X_t(0) + Z_t^{(3)}. \tag{23}$$

In the expression (23) for the total number of clusters C_t, the dominating term is $Z_t^{(3)}$, which according to (18) is of order $(1+r)^t$. Observe that relation (18) implies that both the mean number of clusters $\mathrm{E}(C_t)$ and the standard deviation $\sqrt{\mathrm{Var}(C_t)}$ are growing proportionally to $(1+r)^t$ as $t \to \infty$.

Our results concerning the expected values, presented in the Appendix, are illustrated by Fig. 6. On the right panel of Fig. 6, the four lines, representing different combinations of the parameter values, almost coincide demonstrating that asymptotic relation (2) works well already for $t = 10$.

3. Discussion

The number of approaches and applications that use UMIs in sequencing is rapidly increasing. In cancer diagnostics, the use of UMIs is crucial since it allows to correct for both polymerase induced errors and amplification biases (Johansson et al., 2020). Many sample types and matrices are challenging to analyse due to the limited amounts of DNA and enzymatic inhibitors. In this paper we propose a convenient bookkeeping system for annotating the emerging UMI clusters during t consecutive barcoding PCR cycles.

The proposed tree based bookkeeping system leads to a branching process model for the counts $C_t(m)$ of the UMI clusters of sizes $m = 0, 1, \ldots, m-1$. Our model distinguishes between five PCR amplification rates (r_1, r_2, r_3, r_4, r). A key feature of interest for such a model is the set of proportions $C_t(m)/C_t$, $m = 1, \ldots, t-1$, where $C_t = C_t(1) + \cdots + C_t(t-1)$ is the total number of the UMI clusters.

The main theoretical finding of this paper, convergence (2), claims that the ratio between the expected counts $\mathrm{E}(C_t(m))/\mathrm{E}(C_t)$ is approximately 2^{-m} regardless of the underlying parameters (r_1, r_2, r_3, r_4, r). At the moment, we are lacking an illuminating mathematical explanation of this surprisingly simple asymptotical pattern. To double check this key result, we made computer calculations based on the recursive

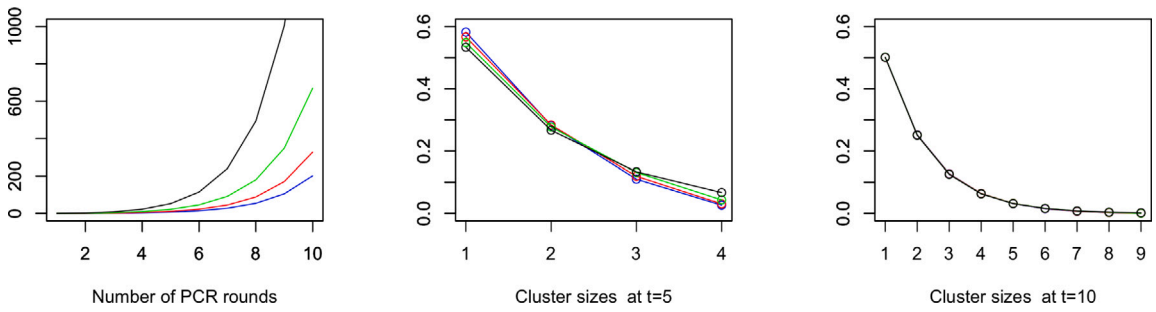

Fig. 6. Left panel shows the expected values of the total number of clusters $\mathrm{E}(C_t)$. Middle panel ($t = 5$) and right panel ($t = 10$) show the plots of $\mathrm{E}(C_t(m))/\mathrm{E}(C_t)$ over the cluster sizes $m = 1, \dots, t-1$. Different colours represent different sets of the parameters (r_1, r_2, r_3, r_4, r): black $(1, 1, 1, 1, 1)$, green $(0.6, 0.6, 0.7, 0.8, 0.9)$, red $(0.4, 0.4, 0.6, 0.7, 0.9)$, and blue $(0.3, 0.3, 0.4, 0.5, 0.9)$.

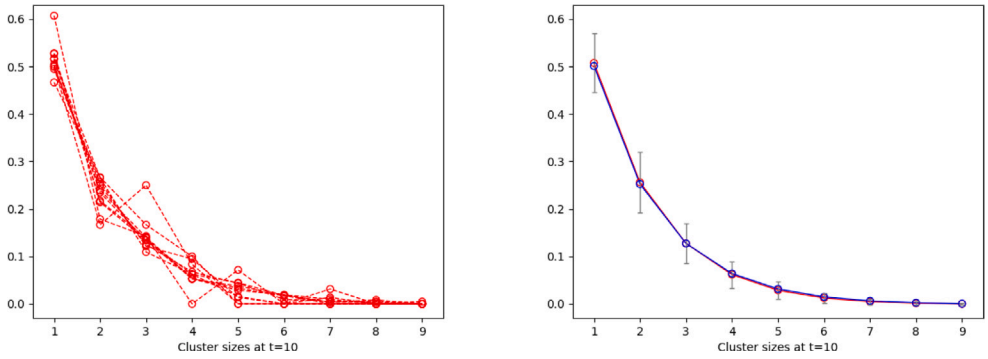

Fig. 7. Simulation results for the proportions $C_t(m)/C_t$, $m = 1, \dots, t-1$ with $t = 10$ and the amplification rates $(r_1, r_2, r_3, r_4, r) = (0.3, 0.3, 0.4, 0.5, 0.9)$. The left panel presents ten individual simulation results. The right panel summarizes 1 000 000 simulation results for the proportions $C_t(m)/C_t$: the red line connects the simulation averages, the vertical intervals attached to the red line show the means ± one standard deviations obtained from the simulations. Also on the right panel, the blue line connects the theoretical values for $\mathrm{E}(C_t(m))/\mathrm{E}(C_t)$, this is the same blue line as on the right panel of Fig. 6.

equations given in Appendix. The results shown in Fig. 6, demonstrate that approximation (2) works well even for moderately large values of t, see the right panel of Fig. 6.

We *hypothesize* a biologically more relevant asymptotic result

$$\mathrm{E}(C_t(m)/C_t) \to 2^{-m}, \quad m \geq 1, \quad t \to \infty. \tag{24}$$

To address this hypothesis, a simulation study based on our model was performed for a particular choice of the amplification rates $(r_1, r_2, r_3, r_4, r) = (0.3, 0.3, 0.4, 0.5, 0.9)$. The simulation results summarized by Fig. 7, support the approximation formula (24) for the number of cycles $t = 10$.

Relation (2) is proven to hold in the limit of infinitely many PCR cycles during UMI labelling. In practice, the number of cycles is typically small. For the practical applicability, it is desirable in future research, to obtain the rate of convergence results, which would imply a recommendation of a lower bound for the number of cycles for (2), or even (24), to apply with a reasonable accuracy.

Our model of the barcoding PCR step uses five different amplification rates as the key model parameters (r_1, r_2, r_3, r_4, r). In sequencing, amplification rates are rarely assessed and there is no general method to determine (r_1, r_2, r_3, r_4). In the framework of quantitative PCR, the overall amplification rate r can be assessed by standard curves (Svec et al., 2015). The amplification rate r varies between assays due to different sequence context as well as between samples due to sample inhibition (Bar et al., 2003; Ruijter et al., 2021). The amplification rate also decreases during the last PCR cycles, when reagents become sparse. Scientist working with PCR and sequencing are experimentally used to the fact that some samples and sequences suffer from poor amplification rates or that the original DNA molecules never become amplified. In future studies, it will be important to verify our model with experimental data to estimate the importance of different model parameters. Such a verified model will be valuable in development of improved sequencing protocols and our ability to detect and quantify individual DNA molecules with single nucleotide resolution.

We assume that both r_3 and r_4 are larger than each of the rates r_1 and r_2, since the DNA molecules become shorter after being amplified once. The shortest amplicon is assumed to have the highest amplification rate r. In our current model, the single rate r disregards the difference between the sense and antisense strands. In future studies, it will be interesting to investigate the effect of introducing different amplification rates for the sense and antisense strands, as well as another amplification rate parameter for the downstream adapter PCR step.

CRediT authorship contribution statement

Serik Sagitov: Methodology, Formal analysis, Writing – review & editing. **Anders Ståhlberg:** Conceptualization, Validation, Writing – review & editing.

Declaration of competing interest

A.S. declares stock ownership and is a board member in Tulebovaasta, Iscaff Pharma and SiMSen Diagnostics AB.

Data availability

No data was used for the research described in the article.

Acknowledgements

We are grateful to professor Peter Jagers for bringing us together and for fruitful discussions. Fig. 7 was produced by Hongui Zhan and Yizhe Gu, two master students at the Chalmers University of

Technology, who performed a computer simulation study based on our model. This research was partially funded by Region Västra Götaland, Sweden; Swedish Cancer Society (22-2080); Swedish Research Council (2020-01008); Swedish Childhood Cancer Foundation (MTI2019-0008 and 2020-0007); the Swedish state under the agreement between the Swedish government and the county councils, the ALF-agreement (ALFGBG-965065); Sweden's Innovation Agency and the Sjöberg Foundation.

Appendix. Expected numbers of clusters

Turning to the expected numbers of clusters, denote

$$c_t = \mathrm{E}(C_t), \quad c_t(m) = \mathrm{E}(C_t(m)), \quad 1 \le m \le t-1.$$

Here we show first that

$$c_t = \alpha(1+r)^t + \alpha_1 t - \alpha_2 + \alpha_3(1-r_4)^t + \alpha_4(1-r)^t, \tag{A.1}$$

where

$$\alpha := (r_1 r_3 + r_2 r_4)/(2r^2),$$
$$\alpha_1 := r_2(1 - r_4 r^{-1}), \quad \alpha_2 := r_1 r_3 r^{-2} + r_2 r_4^{-1},$$
$$\alpha_3 := r_2 r_4^{-1}, \quad \alpha_4 := (r_1 r_3 - r_2 r_4)/(2r^2),$$

and then derive the main result (2) of this paper. Notice that $\alpha_4 = 0$ if $r_1 r_3 = r_2 r_4$, and in the deterministic case (4), relation (A.1) turns into (8).

Put

$$z_t^{(i)} = \mathrm{E}(Z_t^{(i)}), \quad x_t(m) = \mathrm{E}(X_t(m)), \quad y_t(m) = \mathrm{E}(Y_t(m)).$$

The proofs of (A.1) and (2) rely on the recursive relations

$$x_t(0) = r_2 + x_{t-1}(0)(1-r_4), \tag{A.2}$$
$$y_t(1) = r_3 z_{t-1}^{(1)} + r z_{t-1}^{(4)} + y_{t-1}(1)(1-r), \tag{A.3}$$
$$x_t(m) = x_{t-1}(m-1)r_4 + x_{t-1}(m)(1-r_4), \quad 0 \le m \le t-1, \tag{A.4}$$
$$y_t(m) = y_{t-1}(m-1)r + y_{t-1}(m)(1-r), \quad 0 \le m \le t-1, \tag{A.5}$$

following from (22), (20), (19), and (21). In particular, relation (A.3) is obtained by taking the expectations in (19).

Proof of (A.1). From (15) we obtain

$$z_t^{(1)} = r_1 + z_{t-1}^{(1)}, \quad z_t^{(2)} = r_2 + z_{t-1}^{(1)}, \quad z_0^{(1)} = z_0^{(2)} = 0,$$

so that $z_t^{(1)} = r_1 t$, $z_t^{(2)} = r_2 t$, and therefore by (16),

$$z_t^{(3)} = z_{t-1}^{(3)} + r_1 r_3 (t-1) + r z_{t-1}^{(4)},$$
$$z_t^{(4)} = z_{t-1}^{(4)} + r_2 r_4 (t-1) + r z_{t-1}^{(3)}.$$

This yields

$$z_t^{(3)} + z_t^{(4)} = (r_1 r_3 + r_2 r_4)(t-1) + (1+r)(z_{t-1}^{(3)} + z_{t-1}^{(4)})$$
$$= (r_1 r_3 + r_2 r_4) \sum_{j=1}^{t-1} (t-j)(1+r)^{j-1} = 2\alpha((1+r)^t - rt - 1),$$

where we used the relation

$$\sum_{j=1}^{t-1} (t-j)(1+r)^{j-1} = (1+r)^t \sum_{j=2}^{t} (j-1)(1+r)^{-j} = r^{-2}((1+r)^t - rt - 1).$$

On the other hand, we have

$$z_t^{(3)} - z_t^{(4)} = (r_1 r_3 - r_2 r_4)(t-1) + (1-r)(z_{t-1}^{(3)} - z_{t-1}^{(4)})$$
$$= (r_1 r_3 - r_2 r_4) \sum_{j=1}^{t-1} (t-j)(1-r)^{j-1} = 2\alpha_4((1-r)^t + rt - 1),$$

so that

$$z_t^{(3)} = \alpha(1+r)^t - r_2 r_4 r^{-1} t - r_1 r_3 r^{-2} + \alpha_4(1-r)^t,$$

$$z_t^{(4)} = \alpha(1+r)^t - r_1 r_3 r^{-1} t - r_2 r_4 r^{-2} - \alpha_4(1-r)^t.$$

By (23), we have

$$c_t = z_t^{(2)} + z_t^{(3)} - x_t(0),$$

and the stated formula (A.1) follows from the obtained expression for $z_t^{(3)}$ and the next consequence of (A.2):

$$x_t(0) = r_2 r_4^{-1}(1 - (1-r_4)^t). \tag{A.6}$$

Proof of (2). Relation (A.3) implies

$$y_t(1) = \alpha r(1+r)^{t-1} - r_2 r_4 r^{-1} - \alpha_4 r(1-r)^{t-1} + y_{t-1}(1)(1-r),$$

which entails

$$y_t(1) = \sum_{j=1}^{t-1} \left(\alpha r(1+r)^{t-j} - r_2 r_4 r^{-1} - \alpha_4 r(1-r)^{t-j} \right)(1-r)^{j-1}$$
$$= \tfrac{1}{2}\alpha(1+r)((1+r)^{t-1} - (1-r)^{t-1}) - r_2 r_4 r^{-2}(1 - (1-r)^{t-1})$$
$$- \alpha_4 r(1-r)^{t-1}(t-1)$$
$$= \tfrac{1}{2}\alpha(1+r)^t - r_2 r_4 r^{-2} - \alpha_4 r(1-r)^{t-1}(t-1) + \alpha_5(1-r)^{t-1},$$

where $\alpha_5 = r_2 r_4 r^{-2} - \frac{1}{2}\alpha(1+r)$. Thus,

$$y_t(1) \sim \tfrac{1}{2}\alpha(1+r)^t, \quad t \to \infty.$$

Using this as the initiation step for the induction over m based on recursion (A.5), we find that

$$y_t(m) \sim 2^{-m}\alpha(1+r)^t, \quad m \ge 1, \quad t \to \infty.$$

This gives (2) in view of (A.1) and

$$c_t(m) = x_t(m) + y_t(m), \quad 1 \le m \le t-1,$$

where $x_t(m) = o((1+r)^t)$ in accordance with (A.4) and (A.6).

References

Andersson, D., Kristiansson, H., Kubista, M., Ståhlberg, A., 2021. Ultrasensitive circulating tumor DNA analysis enables precision medicine: experimental workflow considerations. Expert Rev. Mol. Diagn. 21, 299–310.

Bar, T., Ståhlberg, A., Muszta, A., Kubista, M., 2003. Kinetic Outlier Detection (KOD) in real-time PCR. Nucleic Acids Res. 31, e105.

Bertoin, J., 2009. The structure of the allelic partition of the total population for Galton–Watson processes with neutral mutations. Ann. Probab. 37, 1502–1523.

Cohen, J.D., Li, L., Wang, Y., Thoburn, C., Afsari, B., Danilova, L., Douville, C., Javed, A.A., Wong, F., Mattox, A., Hruban, R.H., Wolfgang, C.L., Goggins, M.G., Dal Molin, M., Wang, T.L., Roden, R., Klein, A.P., Ptak, J., Dobbyn, L., Schaefer, J., Silliman, N., Popoli, M., Vogelstein, J.T., Browne, J.D., Schoen, R.E., Brand, R.E., Tie, J., Gibbs, P., Wong, H.L., Mansfield, A.S., Jen, J., Hanash, S.M., Falconi, M., Allen, P.J., Zhou, S., Bettegowda, C., Diaz, L.A. Jr, Tomasetti, C., Kinzler, K.W., Vogelstein, B., Lennon, A.M., Papadopoulos, N., 2018. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 359, 926–930.

Filges, S., Yamada, E., Ståhlberg, A., Godfrey, T.E., 2019. Impact of polymerase fidelity on background error rates in next-generation sequencing with unique molecular identifiers/barcodes. Sci. Rep. 9, 3503.

Haccou, P., Jagers, P., Vatutin, V., 2005. Branching processes: variation, growth, and extinction of populations. In: Cambridge Studies in Adaptive Dynamics. Cambridge University Press.

Heitzer, E., Haque, I.S., Roberts, C.E., Speicher, M.R., 2019. Current and future perspectives of liquid biopsies in genomics-driven oncology. Nature Rev. Genet. 20, 71–88.

Ignatiadis, M., Sledge, G.W., Jeffrey, S.S., 2021. Liquid biopsy enters the clinic—Implementation issues and future challenges. Nat. Rev. Clin. Oncol. 18, 297–312.

Jagers, P., Klebaner, F.C., 2003. Random variation and concentration effects in PCR. J. Theoret. Biol. 224, 299–304.

Johansson, G., Kaltak, M., mniceanu, C.Rî., Singh, A.K., J., Lycke., Malmeström, C., Hühn, M., Vaarala, O., Cardell, S., Ståhlberg, A., 2020. Ultrasensitive DNA immune repertoire sequencing using unique molecular identifiers. Clin. Chem. 66, 1228–1237.

Kimmel, M., Alexrod, D.E., 2002. Branching processes in biology. Interdiscip. Appl. Math. 19.

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B., 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc. Natl. Acad. Sci. USA 108, 9530–9535.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2011. Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods 9, 72–74.

Krawczak, M., Reiss, J., Schmidtke, J., Rösler, U., 1989. Polymerase chain reaction: replication errors and reliability of gene diagnosis. Nucleic Acids Res. 17, 2197–2201.

Lalam, N., 2006. Estimation of the reaction efficiency in polymerase chain reaction. J. Theoret. Biol. 242, 947–953.

Mode, C.J., 1970. Multitype Branching Processes: Theory and Applications. American Elsevier.

Pflug, F.G., von Haeseler, A., 2018. TrUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers. Bioinformatics 34, 3137–3144.

Rahimov, I., 2021. Homogeneous branching processes with non-homogeneous immigration. Stoch. Qual. Control 36, 165–183.

Ruijter, J.M., Barnewall, R.J., Marsh, I.B., Szentirmay, A.N., Quinn, J.C., van Houdt, R., Gunst, Q.D., van den Hoff, M.J.B., 2021. Efficiency correction is required for accurate quantitative PCR analysis and reporting. Clin. Chem. 67, 829–842.

Stead, L.F., Sutton, K.M., Taylor, G.R., Quirke, P., Rabbitts, P., 2013. Accurately identifying low-allelic fraction variants in single samples with next-generation sequencing: Applications in tumor subclone resolution. Hum. Mutat. 34, 1432–1438.

Ståhlberg, A., Krzyzanowski, P.M., Jackson, J.B., Egyud, M., Stein, L., Godfrey, T.E., 2016. Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res. 44 (105).

Svec, D., Tichopad, A., Novosadova, V., Pfaffl, M.W., Kubista, M., 2015. How good is a PCR efficiency estimate: Recommendations for precise and robust qPCR efficiency assessments. Biomol. Detect. Quantif. 3, 9–16.

Xu, H., DiCarlo, J., Satya, R.V., Peng, Q., Wang, Y., 2014. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics 15 (244).