ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Kato M, Vasco DA, Sugino R, Narushima D, Krasnitz A. 2017 Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *R. Soc. open sci.* **4**: 171060.
http://dx.doi.org/10.1098/rsos.171060

Received: 4 August 2017
Accepted: 31 August 2017

Subject Category:
Genetics

Subject Areas:
bioinformatics/computational biology/genomics

Keywords:
bioinformatics, cancer genomics, copy-number alteration, molecular evolution, single-cell sequencing, coalescent theory

Author for correspondence:
Mamoru Kato
e-mail: mamkato@ncc.go.jp

[†]Present address: Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, 2, Avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg, EU.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3881761.

THE ROYAL SOCIETY
PUBLISHING

Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer

Mamoru Kato[1], Daniel A. Vasco[2,†], Ryuichi Sugino[3], Daichi Narushima[1] and Alexander Krasnitz[4]

[1]Department of Bioinformatics, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuuoo-ku, Tokyo 104-0045, Japan
[2]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158, USA
[3]School of Advanced Sciences, The Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan
[4]Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, One Bungtown Road, Cold Spring Harbor, NY 11724, USA

MK, 0000-0002-8485-8316

Single-cell sequencing is a promising technology that can address cancer cell evolution by identifying genetic alterations in individual cells. In a recent study, genome-wide DNA copy numbers of single cells were accurately quantified by single-cell sequencing in breast cancers. Phylogenetic-tree analysis revealed genetically distinct populations, each consisting of homogeneous cells. Bioinformatics methods based on population genetics should be further developed to quantitatively analyse the single-cell sequencing data. We developed a bioinformatics framework that was combined with molecular-evolution theories to analyse copy-number losses. This analysis revealed that most deletions in the breast cancers at the single-cell level were generated by simple stochastic processes. A non-standard type of coalescent theory, the multiple-merger coalescent model, aided by approximate Bayesian computation fit well with the data, allowing us to estimate the population-genetic parameters in addition to false-positive and false-negative rates. The estimated parameters suggest that the cancer cells underwent sweepstake evolution, where only one or very few parental cells produced a descendent cell population. We conclude that breast cancer cells successively substitute in a tumour mass, and the high reproduction of only a portion of cancer cells may confer high adaptability to this cancer.

1. Introduction

The idea that tumour progression can be viewed as a Darwinian process goes back to the 1970s, when it led to the concept of 'clonal expansion' [1]. During clonal expansion, tumour cells acquire rare advantageous mutations and then undergo rapid population expansion due to selection. This evolutionary process in tumours is strongly supported by recent genomic studies employing next-generation sequencing for a tumour mass, i.e. a mixture of tumour cells [2–7].

A more direct approach for evolutionary analysis is to study the genomes of individual cells. Single-nucleus sequencing (SNS) is a promising technology that generates high-resolution data, illustrating the genetic heterogeneity of cancer cells and providing an excellent tool for investigating the molecular evolution of cancer cells [8–13]. In SNS, single cells are isolated from tumour tissue by flow cytometry or micromanipulation, and then short DNA fragments (typically 50–200 bp) derived from a single cell are sequenced using a next-generation sequencer. The short sequenced reads are mapped to the human reference genome and then copy-number alterations (CNAs) or point mutations present in single tumour cells are identified by bioinformatics analysis. In particular, CNAs are detected based on the rationale that a larger number of reads mapping to a genomic region reflects a higher copy number in the region [8,14].

A recent study combined flow cytometry with next-generation sequencing and identified CNAs in the genomes of individual cells sampled from tumour tissues obtained from two patients with breast cancer [8]. Phylogenetic analysis of these CNA profiles revealed the existence of genetically distinct subpopulations, each of which was composed of homogeneous cells. These results suggested that breast cancer cells do not gradually evolve, but evolve rapidly between otherwise quiescent evolutionary periods. The results of other studies identified point mutations in the exomes of single cancer cells, and principal-component, phylogenetic-tree and allele-frequency analyses of these mutations revealed the mutational landscape of renal cell carcinoma and the monoclonal origin of essential thrombocythaemia [9,10].

Extensive efforts to develop analytical methods for cancer SNS data are mainly focused on the reconstruction of evolutionary trees such as phylogenetic trees (dendrograms in which the nodes represent cancer cells) [15,16] and mutation trees (dendrograms in which the nodes represent mutation sites) [17,18]. However, these methods assume data on point mutations, not CNAs. Tree reconstruction using SNS CNA data was introduced in a previous study [8], which employed the neighbour-joining method [19] based on the Euclidean distance between the integer copy numbers of cells. Although this distance shows some relatedness between cells, it is a population-genetic metric that may reflect correct genealogical relationships but is not well confirmed. Recently, a pipeline program to analyse SNS CNA data was developed [20]; however, this program focuses on quality control and CNA calling, and uses tree-reconstruction methods for which the distances have not been validated as appropriate metrics for phylogenetic-tree inference using copy numbers (e.g. the Euclidean distance between integer copy numbers). It is necessary to use a valid metric for reconstructing phylogenetic trees that reflects correct genealogies and further to develop an evolutionary model for understanding the dynamics of cancer cells that underlie the reconstructed trees.

For this purpose, we developed a population-genetic framework combined with bioinformatics techniques that analyses SNS CNA data, where cancer cells were treated as individuals of a non-sexually reproducing species. Based on this framework, we decoded integer copy numbers in the previous SNS CNA data [8] into genetic alleles, revealing an unexpectedly simple allelic nature for the breast cancers. We further found that individual cancer genomes fit well with an extended type of coalescent model, namely a multiple-merger coalescent (MMC) model [21] (reviewed in [22]) rather than the standard Kingman coalescent model, which is derived from the classic Wright–Fisher model [23]. MMC modelling allows multiple lineages to be merged simultaneously, based on a probability distribution for the number of merged lineages, whereas the Kingman coalescent model only allows the merger of two lineages. Our current findings explain why the phylogenetic tree showed distinct clades composed of homogeneous cells in a previous study [8], and suggest the underlying microscale dynamics of cancer cells in this cancer type.

2. Results

2.1. The nature of deletion alleles

Here, we analysed data generated in a previous study [8]. In the previous study, SNS was performed to identify integer copy numbers along binned chromosome regions (see the electronic supplementary material for more details regarding the data) for 100 single cells in a tissue designated as T10 and 100 single cells in tissues designated as T16P and T16M, which we collectively designate as T16 hereafter. Integer copy numbers ranged from 0 to 42, where '2' represents the original diploid state, '0' and '1' represent deletions, and numbers greater than 2 indicate amplifications. The tissues were sampled from two patients diagnosed with ductal breast carcinoma. Tissue T10 was the primary carcinoma, and tissue T16 consisted of primary breast and metastatic liver carcinomas. Sampled cancer cells are considered as random samples because cells taken from the macro-dissected tissues were randomly selected and sorted by flow cytometry, and classified into 'subpopulations' with different ploidies based only on their DNA content, without any preference for particular cell types [8]. In this study, we focused only on the copy-number data and did not analyse somatic point mutations because there were few common sites among the cells, due to the low genomic coverage per cell (approx. 6% of the genome per cell).

First, we observed that the copy-number profiles were unexpectedly simple for cancer. The copy-number changes mostly involved 1-copy losses or gains (66% for T10 and 41% for T16), or, at most 2-copy losses or gains (87% for T10 and 76% for T16) (electronic supplementary material, figure S1*a*). In addition, the patterns along the binned chromosome regions were mostly simple, such as a pattern of 2 copies changing to 1 copy and back to 2 copies along the chromosome, from the start to end positions (electronic supplementary material, figure S1*a*). These simple copy-number patterns motivated us to perform a deeper analysis using population genetics. CNA segments were usually (91% for T10 and 96% for T16) composed of either amplifications or deletions, allowing us to analyse amplifications and deletions separately. Because the number of deletion states (0 or 1 copy) is lower than that of amplifications (3–42 copies), we focused on deletions to avoid theoretical complications in subsequent analyses. The deletion patterns were also simple (figure 1*a*).

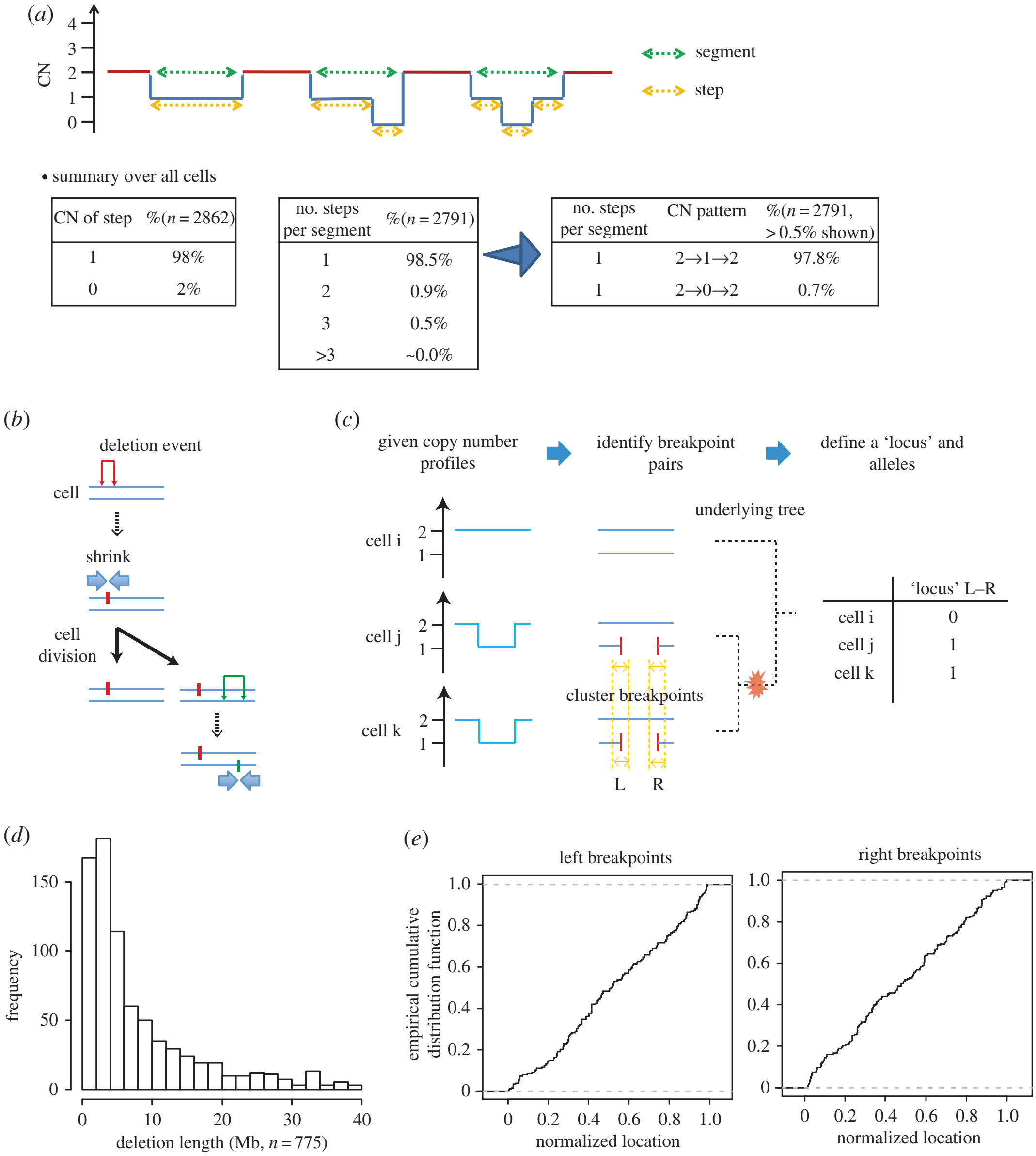
Next, to perform population-genetics analysis, we considered the simple model of deletions illustrated in figure 1*b*, which assumes that any deletion event leaves a unique pair of right and left breakpoints on a chromosome. Under this assumption, all variable loci are bi-allelic, with original and derived allelic states. Because all chromosomes in a single cell are co-inherited by progeny cells, we treated them as if they evolved as a single chromosomal unit. Because paired homologous chromosomes are also inherited together, we assigned two different coordinate systems to a pair of homologous chromosomes, as if one homologous chromosome was physically linked to another by their ends.

Based on this model, we extracted deletion events as alleles. From the integer copy-number profiles for each cell, we obtained the left and right breakpoint pairs of the deletion events, using a simple greedy algorithm (figure 1*c*; electronic supplementary material for the details). We then performed clustering analysis to align the breakpoints across cells because breakpoints may fluctuate due to noise (figure 1*c*; electronic supplementary material). We defined a 'genetic locus' or 'polymorphic site' by a unique breakpoint pair and assigned a derived allele to cells that harboured the deletion. The ancestral allele was assigned to cells without the deletion. Results from principal-component analysis indicated that we successfully decoded the deletion alleles (electronic supplementary material, figure S1*b*).

We observed simple distributions for the deletion allele sizes and breakpoint positions. The number of deletion alleles gradually decreased according to size, which appeared to follow a reciprocal ($1/x$) distribution (figure 1*d*; electronic supplementary material, figure S1*c*). The breakpoints were distributed roughly uniformly across the genome (figure 1*e*). These data indicated that the deletion events occurred largely based on a relatively simple stochastic process, which was amenable to molecular-evolution analysis.

2.2. Multiple-merger coalescent

We drew phylogenetic trees for the subpopulations defined previously [8]. Of all the subpopulations, the tree of the hypodiploid subpopulation (HP) had features expected from MMC theory: skewed branching and multiple mergers within a close distance (fig. 1 in [24]), as shown in figure 2*a* (by the neighbour-joining method) and in electronic supplementary material, figure S2 (by the unweighted pair group method with arithmetic mean [UPGMA]). The HP subpopulation most closely reflected the nature of



CN of step	%(n = 2862)
1	98%
0	2%

no. steps per segment	%(n = 2791)
1	98.5%
2	0.9%
3	0.5%
>3	~0.0%

no. steps per segment	CN pattern	%(n = 2791, > 0.5% shown)
1	2→1→2	97.8%
1	2→0→2	0.7%

	'locus' L–R
cell i	0
cell j	1
cell k	1

Figure 1. The nature of deletion alleles. Results from T10 only are shown because T16 showed essentially the same tendencies. (*a*) Profile patterns of the integer copy number (CN). The horizontal axis represents the chromosomal position. The general CN pattern '2→*n*→ ... →*m*→2' indicates that the copy numbers of a segment changed from 2 copies to *n* copies, ... and to *m* copies finally back to 2 copies along a chromosome. (*b*) Evolutionary model of deletions. Every deletion event (the inverted U-shaped marks shown in red and green) leaves a unique pair of left and right breakpoints as fingerprints on a homologous chromosome (blue lines). (*c*) Procedure to convert copy-number profiles into alleles. 'L' and 'R' represent positions to the left and right of breakpoints, and an 'L–R' pair defines a locus. The symbols '0' and '1' represent the ancestral and derived alleles, respectively. (*d*) Distribution of deletion allele lengths. We excluded deletion alleles larger than the size of a chromosome level (40 Mb). (*e*) Distributions of breakpoint locations. The locations were normalized with respect to chromosome lengths.

deletions of all the subpopulations because HP was dominated by many copy-number losses [8]. Hence, we sought to analyse HP cells using an MMC model.

The objective of using an MMC model was to estimate various parameters that included a parameter related to the probability of multiple mergers. One of the simplest MMC models is the β-coalescent MMC model [22]. We used a modified β-coalescent model that included an exponential growth term (Material and methods section). Inclusion of the exponential growth term was justified by clinical observations with several types of cancer, including breast cancer [25]. We also modelled the occurrence of false-positive and false-negative errors in the data, because SNS data may have considerable errors.

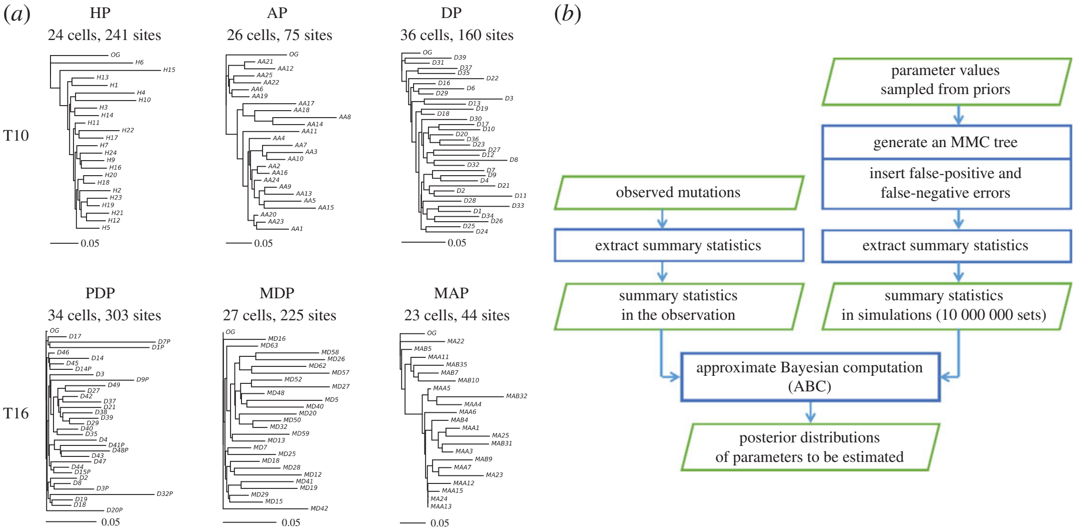

(a)
HP
24 cells, 241 sites
AP
26 cells, 75 sites
DP
36 cells, 160 sites
T10
0.05
PDP
34 cells, 303 sites
MDP
27 cells, 225 sites
MAP
23 cells, 44 sites
T16
0.05
(b)
parameter values sampled from priors
generate an MMC tree
insert false-positive and false-negative errors
observed mutations
extract summary statistics
extract summary statistics
summary statistics in the observation
summary statistics in simulations (10 000 000 sets)
approximate Bayesian computation (ABC)
posterior distributions of parameters to be estimated

Figure 2. Phylogenetic trees and MMC. (*a*) Phylogenetic trees reconstructed by the neighbour-joining method. HP, AP, DP, PDP, MDP and MAP represent respective subpopulations of hypodiploid, aneuploid, diploid, primary diploid, metastatic diploid and metastatic aneuploid cells, which were defined previously [8]. OG represents the outgroup: no deletions at all sites. (*b*) Flow chart of our ABC. (*c*) The posterior distributions for the parameters of the MMC model. (*d*) The posterior distributions for the parameters of the Kingman population-growth model. (*e*) The posterior distributions for the parameters of the Kingman population-constant model. (*f*) Site-frequency spectrum under the MAP estimates. Sample size $n = 23$. (*g*) Distribution of the number of merged lineages under the MAP estimates. For (*f*,*g*), the results of 100 000 replications in the simulations were used.

Our model thus had seven parameters: the growth rate (α), the parameter (β) of the distribution that represents the rate of multiple mergers, population mutation rate (θ), false-positive and false-negative rates, and the numbers of false-positive and false-negative sites. The parameter α takes non-negative values, and a value of zero represents a constant population size. Values of β are defined within the range of 0–2. When β has a value close to 2, the rate distribution for the number (m in equation (4.1) in Material and methods section) of lineages to be merged has a large value for two lineages and smaller values for greater than two lineages. Indeed, the limit of 2 for β represents that only mergers of two lineages occur, as in Kingman coalescent models. When β has a value close to 0, it has larger values for greater than two lineages, meaning that multiple mergers (more than two lineages) tend to occur. Mutational events occur following a Poisson distribution with the mean of $\theta \times l_b$ on a branch of a coalescent tree, where l_b is the branch length. False-positive sites are sites where alleles are originally copy-number neutral for all cells but are misjudged as deletions for some cells. False-negative sites are sites where alleles are originally deletion alleles for all cells but are misjudged as neutral for some cells.

We estimated these parameters in the framework of approximate Bayesian computation (ABC) (figure 2*b*) [26,27]. We used the features and summary statistics listed in table 1. The detailed reasons for selecting the features are described in electronic supplementary material, table S1. For comparison, we also used the models of Kingman coalescent with a constant population size and Kingman coalescent with a population growth. We obtained the posterior distributions (figure 2*c*–*e*) and maximum *a posteriori* probability (MAP) estimates (table 2). In the MMC model, the MAP result of β was 1.6, and the ratio of the posterior probabilities to the β value of nearly 2 (1.999) was 12.2. Hence, it appeared that HP cancer cells were better modelled by multiple mergers than by Kingman two-branch mergers.

Formally, we calculated the posterior probabilities of the three models in model selection when we used the multinomial logistic regression with explanatory variables for the summary statistics. The posterior probabilities were 1.00, 0.00 and 0.00 for the MMC, Kingman population-constant and Kingman population-growth models, respectively, which suggests that MMC was the best model. We confirmed that it was possible to distinguish the three models when we used the multinomial logistic regression by performing a leave-one-out cross-validation analysis of the misclassification rates of the models (misclassification rate of only 1.7% on average; electronic supplementary material, table S2).

In addition, we performed two analyses that complement the analysis of models' posterior probabilities [28]. We first performed the goodness-of-fit test using, as the test statistics, the distance between the accepted summary statistics and observed summary statistics for each model. The p-values were 0.37, 0.18 and 0.05 for the MMC, Kingman population-constant and Kingman population-growth

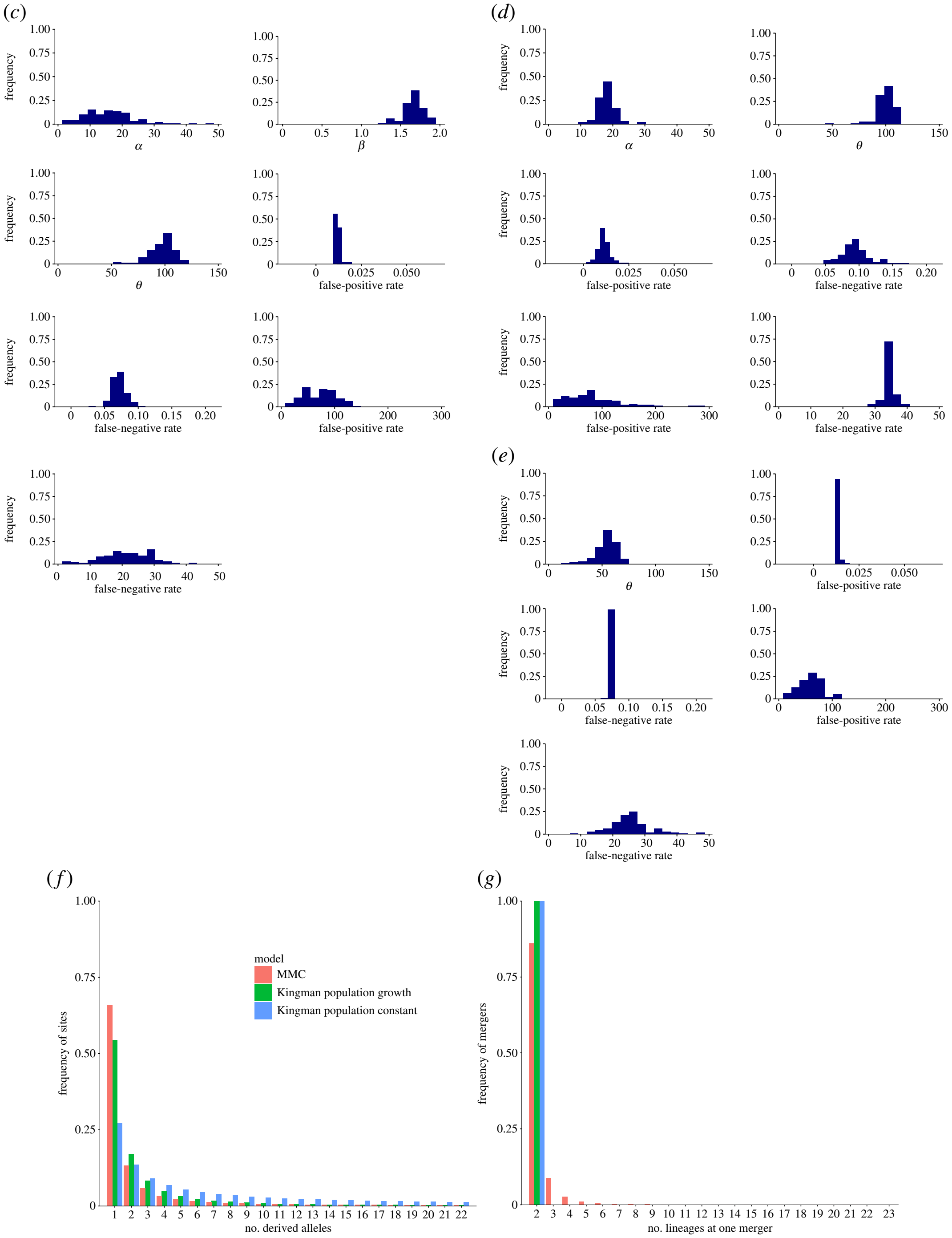


Figure 2. (*Continued*.)

models, respectively. This suggests that the Kingman population-growth model was significantly deviated from observed data, and that the MMC model was the best fit among the three.

Second, we performed posterior predictive checks for each model, where we checked the concordance between summary statistics calculated from observed data and summary statistics calculated from coalescent simulations performed secondarily, based on 1000 sets of parameter values sampled from the initially obtained posterior distributions for the parameters (electronic supplementary material, figure S3). In addition, we checked the concordance between the observed summary statistics and summary statistics secondarily simulated under the MAP estimates (electronic supplementary material, figure S3). These results showed that the summary statistics were reasonably reproduced for all of the models, although the Kingman population growth and constant models were least concordant in the predictive checks based on the posteriors and MAP estimates, respectively. In summary, the MMC model was always the best model across all the three analyses of model selection.

Table 1. Features and summary statistics. The reasons for selecting these features are listed in electronic supplementary material, table S1.

feature	summary statistics
number of mutation sites	the number itself
allele frequencies at all sites	10, 30, 50, 70 and 90% quantiles
distances between all cell pairs in a tree	10, 30, 50, 70 and 90% quantiles
all branch lengths in a tree	10, 30, 50, 70 and 90% quantiles
associations (r^2) between all site pairs	10, 30, 50, 70 and 90% quantiles

Table 2. MAP estimates. α and β represent the population-growth rate and the parameter of the distribution that describes the rate of multiple mergers, respectively. See the text for more information on α and β. Here, θ is the population mutation rate.

parameter	the MMC model	Kingman population-constant model	Kingman population-growth model
α	17.1	n.a.	17.7
β	1.64	n.a.	n.a.
θ	102.9	52.1	97.2
false-positive rate	0.012	0.013	0.010
false-negative rate	0.068	0.072	0.098
false-positive sites	89	72	70
false-negative sites	21	26	35

We also calculated an allele-frequency spectrum (figure 2*f*). This spectrum was drawn based on the estimated parameters without false-positive or false-negative errors, because a spectrum (electronic supplementary material, figure S4) constructed directly from the observed data might have been contaminated by false positives and negatives. For comparison purposes, we calculated spectra for Kingman population-constant and population-growth models. The β coalescent with growth showed an intense frequency at the smallest number of derived alleles and sharp drops in the frequencies at large numbers, particularly at the second smallest number (figure 2*f*). The Kingman growth model showed a less intense frequency at the smallest number, but did not show as sharp a drop in the frequency at the second smallest number. The Kingman constant model did not have an intense frequency at the smallest number and had long-tail frequencies at large numbers. We also computed a distribution for the number of merged lineages observed during the simulation. Mergers of more than two lineages were found (figure 2*g*). Both Kingman models were two-lineage mergers by definition.

3. Discussion

In this study, we developed a computational framework that integrates bioinformatics copy-number algorithms with population-genetics theory. Using this approach, we quantitatively analysed the previous SNS CNA data in breast cancers [8]. Our analyses of copy-number profiles and deletion alleles demonstrated that their patterns were unexpectedly simple for cancer. Other investigators proposed the fractal globule model to explain the $1/x$ distribution of CNA sizes in typical bulk-cell sequencing [29,30], and our analysis demonstrated that this observation held true at the single-cell level. The $1/x$ distribution, together with the uniform distribution of breakpoint positions in chromosomes, may serve as a future simulation framework for modelling stochastic processes of CNAs in cancer cells. Lengths around branch mergers in the HP tree (figure 2*a*) mostly appeared short enough to be approximated with multiple mergers; indeed, MMC fit better to the CNA data than it did to the Kingman coalescent models.

A phylogenetic analysis by Navin *et al*. [8] demonstrated the presence of distinct subpopulations composed of homogeneous cancer cells; no clear intermediate subpopulations were found in the breast cancer cells they examined. The absence of intermediate subpopulations can be explained by 'sweepstake' reproductive processes underlying the MMC model. Unlike the Wright–Fisher model or the Kingman coalescent model, MMC is characterized by great variance in the number of descendants:

the MMC models are coalescent processes in species with 'sweepstake' reproduction such as fish and parasites, in which only one or very few individuals produce descendants [21,22,31]. The population is composed of very few genotypes. Therefore, cancer cells within the same subpopulation were genetically *homogeneous* in the previous study [8]. On the other hand, the time of allele fixation in the sweepstake reproduction modelled in MMC is short; hence, many divergences (substitutions) tend to accumulate between two incipient populations [22]. This is the reason why *distinct* subpopulations were observed in the previous study [8].

One important prediction by MMC is that alleles under positive selection theoretically may have a probability of 1 to become fixed [22]. The possibility that even a slightly advantageous allele can be fixed under a little genetic drift may be related to numerous 'passenger' mutations observed in recent cancer-genomics studies [2,32].

There are several biological and medical implications if the cancer data fit the MMC model.

(1) To understand how cancer is generated in a human body, cancer genomics employing typical next-generation sequencing for bulk cells estimates the order of dysfunctional genes from the variant allele frequencies in a tumour tissue sample, based on the idea that older variants have higher variant allele frequencies [33,34]. For example, if variants in *KRAS* and *TP53* show variant allele frequencies of 50% and 30%, it is estimated that the *KRAS* variant occurred before the *TP53* variant. In this example, *KRAS* is interpreted as a possible initiating factor for this cancer. This reconstruction holds true in the Wright–Fisher and Kingman models; however, it is not true in the MMC model because higher variant frequencies may just reflect variants occurring in a rapidly expanding subpopulation [35]. If the data fit the MMC model well, this order reconstruction method may be incorrect.

(2) As with the management of marine species [31], the reproductive skew in the MMC model has implications for the management of cancer treatment. The reproductive skew is represented by a heavy-tailed $Cx^{-\beta}$ distribution for the probability of having x or more offspring, where C is a positive constant [31,36]. If data from a cancer patient fit with the MMC model, it suggests that killing cancer cells randomly with anti-cancer drugs would be ineffective because the surviving cells with very high reproduction located in the heavy tail of the distribution will surely re-emerge. It is much more effective to distinguish such cells using biomarkers and kill them directly. The cancer stem cell hypothesis suggests that only a small portion of cancer cells with stem cell properties generate mitotic descendent cells, which constitute almost all of the cancer cell population [37]. This hypothesis may be associated with the high reproductive skew represented in the MMC model, and some markers (e.g. $\text{CD44}^{+}/\text{CD24}^{-}$ for breast cancer) to distinguish cancer stem cells have already been developed.

(3) If the cancer data fit the MMC, it is disadvantageous to take a wait-and-see approach because even slightly advantaged variants may spread through the population and cancer cells rapidly evolve; thus, an estimated β can serve as an index to represent the malignancy of the cancer.

To our knowledge, this is the first study to apply MMC modelling to cancer SNS data. Branching processes have been often used to model cancer evolution [13,38]. Branching processes are a time-forward type of model, while MMC is a time-backward type of model. The standard Kingman model as a backward model can be derived from the Wright–Fisher model as a forward model [23]. In this light, it is interesting that recent theoretical studies indicated a relationship between MMC as a backward model and branching processes as a forward model [39,40].

Although multiple deletion events may share the same breakpoint pairs, we ascertained that virtually all deletions in the dataset arose from single events (see the electronic supplementary material). One caveat in our analysis is that we only examined simple deletions defined from copy-number profiles. We did not identify deletions within amplifications, let alone amplifications or point mutations. Future studies are warranted to include these mutations. Moreover, if data with a sufficient number of cells were obtained from every dissected sector of a tissue, 'geographic' differences in tissues could be addressed in the future. The mutational model of CNAs depicted in figure 1*b* and also suggested in a previous study [41] is in principle applicable to germ-line copy-number variations (CNVs) and therefore may also be helpful for improving population-genetic studies of CNVs [42,43].

4. Material and methods

4.1. Phylogenetic tree

Because we extracted genetic alleles from copy-number profiles, we applied standard phylogenetic-construction methods that are used for point mutations. For the distance, we used the p-distance [44] because multiple occurrences were unlikely to occur, as described in the electronic supplementary material. We used the neighbour-joining method [19] for agglomeration, in addition to the UPGMA method. We constructed trees for subpopulations with greater than 20 cells.

4.2. Multiple-merger coalescent

In our coalescent simulation, we used a β-coalescent model modified to include population growth. We based our simulation procedures on reference [24]. In the β-coalescent model [22], the rate of merger of m lineages from k active lineages is represented by

$$\left.\begin{aligned}\lambda_{k,m} &= \int_0^1 x^{m-2}(1-x)^{k-m}\left\{\frac{1}{\Gamma(2-\beta)\Gamma(\beta)}x^{1-\beta}(1-x)^{\beta-1}\right\}\,\mathrm{d}x\\ &= \frac{B(m-\beta,k-m+\beta)}{B(2-\beta,\beta)},\end{aligned}\right\} \tag{4.1}$$

where Γ and B represent the gamma and beta functions, respectively, and β is the parameter. The number of lineages to be merged is first sampled from the probabilities:

$$p_{k,m} = \binom{k}{m}\lambda_{k,m}. \tag{4.2}$$

Next, particular lineages to be merged are randomly sampled. This process is repeated until no lineages remain.

Coalescent time t_k under the assumption of a constant population size is simulated in [24] as follows:

$$t_k \sim \text{Exponential}(p_k) \tag{4.3}$$

and

$$p_k = \sum_{m=2}^{k} p_{k,m}, \tag{4.4}$$

where t_k is sampled from an exponential distribution with a rate of p_k. Intuitively, coalescent events occur, following a Poisson process with the average number of occurrences of p_k in coalescent time units.

We scaled the waiting times to include the effect of population growth, following the standard approach in coalescent theories [45]. This approach focuses only on changes in the coalescent rate due to changes in population size, i.e. smaller population size is associated with a higher coalescent rate, and scales the coalescent time appropriately. We assumed an exponential growth model:

$$N(t) = N_0\mathrm{e}^{-\alpha t}, \tag{4.5}$$

where N_0 is the population size at the present and t is the time before the present; α is the growth rate measured in coalescent time units ($4N_0$ generations). Exponential growth can be included in a coalescent model by scaling time as follows:

$$t_k^* \sim \text{Exponential}(p_k), \tag{4.6}$$

where t_k^* is sampled from an exponential distribution with a rate of p_k. The time is then scaled with the growth rate α:

$$t_k = \frac{1}{\alpha}\log(1+\alpha t_k^*\mathrm{e}^{-\alpha v_{k+1}}), \tag{4.7}$$

and

$$v_{k+1} = \sum_{i=k+1}^{n} t_i, \tag{4.8}$$

where n is the sample size. We thus obtained the coalescent tree and time.

For each branch of an MMC tree, we sampled the number of mutational events from a Poisson distribution with the mean of $\theta \times l_\mathrm{b}$, where θ is the population mutation rate (in coalescent time units) and l_b is the branch length. We then placed mutational events onto the branch. As shown in figure 1*c*, we treated the breakpoint pair of a deletion as a point mutation that follows the infinite-site model [46].

4.3. Approximate Bayesian computation

We first sampled values from prior distributions, assuming uniform distributions:

α: {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100},
β: {0.001, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 1.999},
θ: {10, 20, 30, 40, 50, 60, 70, 80, 90, 100},
False-positive rate: {0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1},
False-negative rate: {0.02, 0.04, 0.06, 0.08, 0.10, 0.12, 0.14, 0.16, 0.18, 0.2},
Number of false-positive sites: {10, 20, 30, 40, 50, 60, 70, 80, 90, 100},
Number of false-negative sites: {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}.

Here, the false-positive sites are sites that had copy-number-neutral alleles for all cells originally, but were misjudged to have a deletion allele for at least one cell. Conversely, false-negative sites are sites that had deletion alleles for all cells originally, but were misjudged to have a neutral allele for at least one cell.

Under a set of parameter values, we generated an MMC tree with mutations. From the tree, we obtained DNA sequences where derived and ancestral alleles were represented as ‘1’ and ‘0’, respectively. We then simulated false positives and negatives by performing Bernoulli’s trials with the probability of the given false-negative and false-positive rates, and then flipped the alleles (‘1’ to ‘0’ or ‘0’ to ‘1’) based on the outcomes of the trials, respectively. The same method was applied to Bernoulli’s trials for single sites where the alleles were all ‘1’s (or ‘0’s) across the cells. It follows that we obtained sites where at least one ‘1’ (or ‘0’) was flipped over. Then, we added such sites to the DNA sequence data up to the given number of false-negative (or false-positive) sites. In this way, we simulated DNA sequences with false positives and false negatives.

We then extracted five features and their summary statistics, as given in table 1. The reasons for selecting these features are described in electronic supplementary material, table S1. The reason for using the summary statistics of quantiles is that we wished to use information as close to the distribution itself as possible. We repeated these processes 10 000 000 times to obtain 10 000 000 sets of summary statistics.

Using the ‘abc’ package [47] of R, we compared the summary statistics obtained with the simulated data with those with the observed data, based on ABC with the ridge regression adjustment (method=‘ridge’ in the ‘abc’ function) [26]. We determined the acceptance rate to be 0.001%, based on prediction errors calculated from 100 cross-validations for each parameter at different acceptance rates by the ‘cv4abc’ function (electronic supplementary material, table S3). We used features of the tree reconstructed from the neighbour-joining method for the ABC features related to a tree.

For the population-growth Kingman model, we fixed α to 0. For the population-constant Kingman model, we further fixed β to 1.999 (approx. 2). For these models, we performed the same ABC procedures that were performed for the β coalescent with growth.

In model selection analysis, we used the ‘postpr’ function to calculate the posterior probabilities of the three models in the multinomial logistic regression, and used the ‘cv4postpr’ function to perform a leave-one-out cross-validation analysis for the misclassification rates of the three models. We used the ‘gfit’ function to perform a leave-one-out cross-validation for the goodness-of-fit test using a statistic of the distance between the accepted summary statistics and the observed summary statistics.

Data accessibility. The datasets supporting this article have been uploaded as part of the electronic supplementary material. The source codes for MMC simulations are freely available from Dryad (http://dx.doi.org/10.5061/dryad.71jp0) [48].

Authors’ contributions. M.K. designed the study, performed analyses and wrote the manuscript. D.A.V. and D.N. performed analyses and wrote the manuscript. R.S. performed analyses. A.K. modified the study design and directed the study. All the authors gave their final approval for publication.

Competing interests. We have no competing interests.

Funding. This work was supported by the National Institutes of Health (ARRA 1RC2 CA148532-1 to A.K.), MEXT (25134721, 25830141, and 15K06916 to M.K.), AMED (15ck0106013h0002 to M.K.), the Foundation for the Promotion of Cancer Research (to M.K.) and the Kato Memorial Bioscience Foundation (to M.K.).

Acknowledgements. We acknowledge Bernard Fendler, Hideki Innan, Jude Kendall, Michael Wigler, Mickey Atwal, Nicholas Navin, Shuhei Mano, Tatsuhiko Tsunoda and Timour Baslan for providing useful information and suggestions. We thank Eisaku Furukawa and Momoko Nagai for providing technical assistance.

References

1. Nowell PC. 1976 The clonal evolution of tumor cell populations. *Science* **194**, 23–28. (doi:10.1126/science.959840)
2. Wood LD *et al.* 2007 The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113. (doi:10.1126/science.1145720)
3. Beroukhim R *et al.* 2010 The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905. (doi:10.1038/nature08822)
4. Maley CC *et al.* 2006 Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473. (doi:10.1038/ng1768)
5. Ding L *et al.* 2012 Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510. (doi:10.1038/nature10738)
6. Tao Y *et al.* 2011 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**, 12 042–12 047. (doi:10.1073/pnas.1108715108)
7. Bignell GR *et al.* 2010 Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898. (doi:10.1038/nature08768)
8. Navin N *et al.* 2011 Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94. (doi:10.1038/nature09807)
9. Hou Y *et al.* 2012 Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885. (doi:10.1016/j.cell.2012.02.028)
10. Xu X *et al.* 2012 Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895. (doi:10.1016/j.cell.2012.02.025)
11. Zong C, Lu S, Chapman AR, Xie XS. 2012 Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626. (doi:10.1126/science.1229164)
12. Li Y *et al.* 2012 Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience* **1**, 12. (doi:10.1186/2047-217X-1-12)
13. Wang Y *et al.* 2014 Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160. (doi:10.1038/nature13600)
14. Baslan T *et al.* 2012 Genome-wide copy number analysis of single cells. *Nat. Protocols* **7**, 1024–1041. (doi:10.1038/nprot.2012.039)
15. Ross EM, Markowetz F. 2016 OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **17**, 69. (doi:10.1186/s13059-016-0929-9)
16. Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N. 2015 BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* **16**, 36. (doi:10.1186/s13059-015-0592-6)
17. Kim KI, Simon R. 2014 Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinf.* **15**, 27. (doi:10.1186/1471-2105-15-27)
18. Jahn K, Kuipers J, Beerenwinkel N. 2016 Tree inference for single-cell data. *Genome Biol.* **17**, 86. (doi:10.1186/s13059-016-0936-x)
19. Saitou N, Nei M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
20. Garvin T, Aboukhalil R, Kendall J, Baslan T, Atwal GS, Hicks J, Wigler M, Schatz MC. 2015 Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060. (doi:10.1038/nmeth.3578)
21. Eldon B, Wakeley J. 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172**, 2621–2633. (doi:10.1534/genetics.105.052175)
22. Tellier A, Lemaire C. 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Mol. Ecol.* **23**, 2637–2652. (doi:10.1111/mec.12755)
23. Nordborg M. 2007 Coalescent theory. In *Handbook of statistical genetics* (eds DJ Balding, MJ Bishop, C Cannings), pp. 602–635. New York, NY: John Wiley & Sons.
24. Neher RA, Hallatschek O. 2013 Genealogies of rapidly adapting populations. *Proc. Natl Acad. Sci. USA* **110**, 437–442. (doi:10.1073/pnas.1213113110)
25. Friberg S, Mattson S. 1997 On the growth rates of human malignant tumors: implications for medical decision making. *J. Surg. Oncol.* **65**, 284–297.
26. Csillery K, Blum MG, Gaggiotti OE, Francois O. 2010 Approximate Bayesian Computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418. (doi:10.1016/j.tree.2010.04.001)
27. Zhao J, Siegmund KD, Shibata D, Marjoram P. 2014 Ancestral inference in tumors: how much can we know? *J. Theor. Biol*. **359**, 136–145. (doi:10.1016/j.jtbi.2014.05.027)
28. Robert CP, Cornuet JM, Marin JM, Pillai NS. 2011 Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl Acad. Sci. USA* **108**, 15 112–15 117. (doi:10.1073/pnas.1102900108)
29. Fudenberg G, Getz G, Meyerson M, Mirny LA. 2011 High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.* **29**, 1109–1113. (doi:10.1038/nbt.2049)
30. Mirny LA. 2011 The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19**, 37–51. (doi:10.1007/s10577-010-9177-0)
31. Niwa HS, Nashida K, Yanagimoto T. 2016 Reproductive skew in Japanese sardine inferred from DNA sequences. *Ices J. Mar. Sci.* **73**, 2181–2189. (doi:10.1093/icesjms/fsw070)
32. Raphael BJ, Dobson JR, Oesper L, Vandin F. 2014 Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med.* **6**, 5. (doi:10.1186/gm524)
33. Landau DA *et al.* 2013 Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726. (doi:10.1016/j.cell.2013.01.019)
34. Shah SP *et al.* 2012 The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* **486**, 395–399. (doi:10.1038/nature10933)
35. Sargsyan O, Wakeley J. 2008 A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popul. Biol.* **74**, 104–114. (doi:10.1016/j.tpb.2008.04.009)
36. Steinrucken M, Birkner M, Blath J. 2013 Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theor. Popul. Biol.* **87**, 15–24. (doi:10.1016/j.tpb.2013.01.007)
37. Fulawka L, Donizy P, Halon A. 2014 Cancer stem cells—the current status of an old concept: literature review and clinical approaches. *Biol. Res.* **47**, 66. (doi:10.1186/0717-6287-47-66)
38. Bozic I *et al.* 2010 Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA* **107**, 18 545–18 550. (doi:10.1073/pnas.1010978107)
39. Abraham R, Delmas J-F. 2015 β-coalescents and stable Galton-Watson trees. *Latin Am. J. Probab. Math. Stat.* **12**, 451–476.
40. Berestycki N. 2009 Recent progress in coalescent theory. The Brazilian Mathematical Society (SBM) (Sociedade Brasileira de Matemática) 193.
41. Schwarz RF, Trinh A, Sipos B, Brenton JD, Goldman N, Markowetz F. 2014 Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* **10**, e1003535. (doi:10.1371/journal.pcbi.1003535)
42. Kato M, Nakamura Y, Tsunoda T. 2008 An algorithm for inferring complex haplotypes in a region of copy-number variation. *Am. J. Hum. Genet.* **83**, 157–169. (doi:10.1016/j.ajhg.2008.06.021)
43. Kato M *et al.* 2010 Population-genetic nature of copy number variations in the human genome. *Hum. Mol. Genet.* **19**, 761–773. (doi:10.1093/hmg/ddp541)
44. Yang Z. 2006 *Computational molecular evolution*. Oxford, UK: Oxford University Press.
45. Hein J, Schierup MH, Wiuf C. 2005 *Gene genealogies, variation and evolution*. Oxford, UK: Oxford University Press.
46. Kimura M. 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**, 893–903.
47. Csillery K, Francois O, Blum MGB. 2012 abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479. (doi:10.1111/j.2041-210X.2011.00179.x)
48. Kato M, Vasco DA, Sugino R, Narushima D, Krasnitz A. 2017 Data from: Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.71jp0)