bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Time-resolved, integrated analysis of clonal genome evolution in parthenogenetic animals and in cancer

Carine Legrand[1,*], Ranja Andriantsoa[1], Peter Lichter[2,3], Frank Lyko[1]

1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, 69120 Heidelberg, Germany

2 Division of Molecular Genetics, German Cancer Research Consortium (DKTK), German Cancer Research Center (DKFZ), Im Neuenheimer ; Feld 280, 69120 Heidelberg, Germany

3 Molecular Precision Oncology, National Center for Tumor Diseases, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany

* Corresponding author: E-mail address: carine.legrand1@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Clonal genome evolution is a key aspect for parthenogenetic species and cancer. While many studies describe precise landscapes of clonal evolution in cancer, few studies determine the underlying evolutionary parameters from molecular data, and fewer integrate theory with data. We derived theoretical results linking mutation rate, time, expansion dynamics, transition to recurrence, and survival. With this, we inferred time-resolved estimates of evolutionary parameters from mutation accumulation, mutational signatures and selection. Using this framework we traced the speciation of the rapidly emerging and invasive marbled crayfish to a time window between 1947 and 1996, which is consistent with biological records. In glioblastoma samples, we determined tumor expansion patterns, and tumor cell survival ratio at resection. Interestingly, our results suggest that the expansion pattern in the primary tumor is predictive of the progress and time to recurrence. In addition, tumor cell survival was always higher after resection and was associated with the expansion pattern and time to recurrence. We further observed selection events in a subset of tumors, with longer and purifying-only selection phases in recurrent tumors. In conclusion, our framework allowed a time-resolved, integrated analysis of key parameters in clonally evolving genomes, and provided novel insights into the evolutionary age of marbled crayfish and the progression of glioblastoma.

Keywords

Clonal genome evolution, Glioblastoma, Marbled crayfish, Integrative analysis, Mathematical modeling, Mutation rate, Selection, Mutational signatures, Tumor regrowth, Recurrence

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

The evolution of genomes is shaped by many factors, among which the random accumulation of mutations over time plays a fundamental role [1, 2]. Far from being homogeneous, the probability of a mutation depends on many factors such as the genomic location [3], mutator alleles, local nucleotide context or mutagenic exposures [4]. Other genomic modifications include recombination in sexual reproduction, copy number variants and genomic rearrangements, gene transfers and hybridization. The capacity of any genomic modification to be inherited is partly stochastic, for instance through genetic drift [5], but can be favored or disfavored by positive or negative selection. Genome evolution could be observed historically via the mere measurement of phenotypes [6], and can now be determined precisely using high-throughput sequencing in parallel with experimental or cohort settings, such as mutation accumulation experiments, or the analysis of genetic trios [7, 8].

Clonal genome evolution is shaped by a more limited set of mechanisms. Mutation rate, selection and variant frequencies are key parameters, which determine the speed of evolution, and which function under the influence of selection pressure. While truly clonal genome evolution is rare in animals, it is recognized as a necessary diversification mechanism within an organism. Prominent examples include hematopoiesis and the immune response, which involves a large number of antibodies. Selection has also been studied in clonally evolving genomes [9, 10, 11, 12]. In the context of tumor genome evolution, some studies have claimed that there can be a fully neutral evolution, at least in an established tumor [13, 9]. However, this is discussed controversially [9, 14, 15, 16, 17]. In particular, the usefulness of allele

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

58 frequency data to infer selection is still unclear. A limited number of studies addresses
59 the inference of the timeline, using coalescent approaches [18], or using a
60 probabilistic framework [19]. Few studies bring together the interplay of the mutation
61 rate, selection and timeline in clonal genome evolution.

62 Cancer constitutes a disease based on clonal genome evolution, defined by somatic
63 mutations, copy number variants, large-scale chromosome anomalies and germline
64 risk variants. Recently, multiple-region sampling and emerging single-cell sequencing
65 have provided an unprecedented view on tumor heterogeneity and cancer cell
66 phylogenies [20, 21, 22, 23, 24, 25], while evolutionary game theory models shed
67 light on the interaction of treatment with cancer evolution and allow treatment
68 adaptation [26, 27]. In glioblastoma, a detailed analysis of tumor trajectories revealed
69 a common tumorigenesis onset via specific chromosome gains or losses, while driver
70 mutations occurred later and led to rapid growth [19]. Furthermore, mathematical
71 modeling yielded time estimates for tumorigenesis ranging from 2 to 7 years before
72 diagnosis [19].

73 Animal models have played a key role in understanding various aspects of tumor
74 formation [28, 29, 30, 31]. Due to its particular mode of asexual reproduction, the
75 marbled crayfish (*Procambarus virginalis*) represents an ideal animal model to study
76 clonal genome evolution [32]. The animals are currently colonizing diverse habitats in
77 a process that is associated with emerging genetic differentiation [33]. Interestingly,
78 marbled crayfish appears to be an evolutionary young species, as their first emergence
79 can be traced back to a specific event in 1995 [34]. If confirmed, this exceptionally
80 young evolutionary age would represent a highly distinctive feature of the model
81 system.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, we aimed to establish a novel framework for analyzing clonal genome evolution in marbled crayfish and in cancer. To this end, we reformulated the dependence of mutation accumulation on variant allele frequency. We used the resulting equation to prove that selection is undecidable using variant allele frequency only, and to determine the links between the various parameters. We enriched this framework by integrating the non-synonymous to synonymous ratio and mutational signatures, in order to estimate selection, time course and tumor expansion parameters. We took care to evaluate uncertainties, using bootstrap. We applied our approach to the clonally evolving marbled crayfish and to recently published samples of primary and recurrent glioblastoma tumors [19]. For both, we provided a detailed view of mutation accumulation, selection, and time. For marbled crayfish, this resulted in a time estimate for its origin. In glioblastoma samples, we further determined tumor expansion parameters and tumor cell survival at the transition between the primary and recurrent tumors.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2 Results

Theoretical results on the mutation rate, allele frequency, growth and survival

To gain insights on the mutation rate in a clonal genome, we studied the theoretical properties of mutation accumulation *dM*, in relation to mutation frequency *f*. We first used the expression of *dM*:

$$dM(t) = \mu(t) \cdot \pi(t) \cdot G \cdot 2 \cdot \omega(t) \cdot \gamma(t) \cdot N(t) \cdot dt, \quad (1)$$

where t is the time, μ is the mutation rate, π is the ploidy, G is the genome size, ω is the growth rate, γ is the survival rate and N is the number of animals, or cells in the context of cancer (for a detailed description of this expression, and for the demonstration of the following equations, see the Supplementary Demonstration). We stratified this expression to each subclone, since animal or cell lineages are likely to have different evolutionary parameters. Then, we introduced the observed mutation frequency $f_i(t\,;t_r)$. This is because *f* was observed not at occurrence, but at the time of retrieval or resection of a genomic DNA sample, t_r, and because a mutation in subclone i was then diluted among all subclones present at t_r. We obtained the following expression (2) for $f_i(t\,;t_r)$:

$$f_i(t\,;t_r) = \frac{K_{i,r}}{N_i(t) \cdot \pi_i(t)}. \quad (2)$$

Equation (2) means that *f* is inversely proportional to *N* and to ploidy. The term $K_{i,r}$ is a constant for subclone *i*, which accounts for the actual time of appearance of the mutation and for the dilution of subclone i in the sample (see assumptions in Supplementary Demonstration).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next, we needed intermediate results about the increment of *1/f*, *d(1/f)* and the increment of *N*, *dN*. Note that we used the inverse allele frequency because this leads to simpler equations, and stays equivalent. Using calculus (see mathematical proof in Supplementary Demonstration), this led to expressions (3) and (4):

$$dN_i = \omega_i(t) \cdot \gamma_i(t) \cdot N_i(t) \cdot dt \tag{3}$$

and

$$d\left(1/f_i(t;t_r)\right) = \omega_i(t) \cdot \gamma_i(t) \cdot N_i(t) \cdot dt \cdot \pi_i. \tag{4}$$

Notably, we have made the assumption that ploidy is constant in order to obtain equation (4). As a result, using equations (1), (3) and (4), it was possible to obtain the dependence (5) of mutation accumulation on frequency *f*, in each subclone *i*:

$$dM_i(t) = \mu_i(t) \cdot G \cdot K_{i,r} \cdot d\left(1/f_i(t;t_r)\right). \tag{5}$$

Finally, mutation accumulation overall was simply obtained as the sum of (5) in all subclones *i*. Because the observed frequencies $f_i(t;t_r)$ are comparable between subclones, we used *f* in the following, while $K_{i,r}$ continued to account for the time and proportion differences between subclones among the sample. This yielded the following equation (Supplementary Demonstration):

$$dM(t) = \left(\sum_i \mu_i(t) \cdot K_{i,r}\right) \cdot G \cdot d(1/f). \tag{6}$$

Equation (6) states that mutation accumulation *dM* is proportional to *d(1/f)*. Furthermore, *dM/d(1/f)* can be constant, meaning that M(1/f) is linear, if the mutation rates μ_i are constant. Conversely, when M(1/f) is linear, then the mutation rates are certainly, but not automatically, constant (Supplementary Demonstration). Equation

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

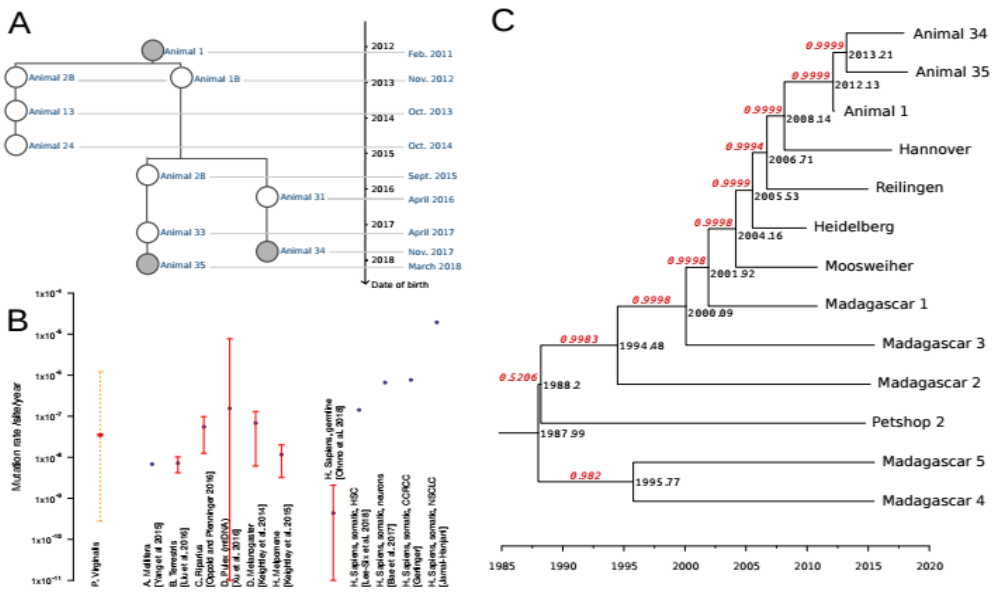
(6) also excludes any role of selection on $M(1/f)$, in agreement with [11, 12], and within the frame of assumptions (Supplementary Demonstration).

These theoretical results provided the foundation for calculating mutation rate variations from the curve M(1/f), and time-resolved estimates, in *P. virginalis* and in glioblastoma samples in the following. In addition, these results allowed to derive the expressions between time, mutation rate, survival rate and growth rate in glioblastoma samples.

Mutation rate estimates and a timed coalescent tree for *P. virginalis*

In order to infer evolutionary parameters of the *P. virginalis* genome, we first assessed the mutation rate. We used whole-genome sequencing of a line of direct descendants from our laboratory colony of *P. virginalis*, that were sampled over a period of seven years (Fig. 1A). The mutation rate was calculated as the average number of de novo mutations in animals 34 and 35 as compared to animal 1, per nucleotide and per year. From these samples, we obtained a mutation rate equal to $\mu = 3.51 \cdot 10^{-8} / nt / y$ (95% confidence interval (CI): $\left[1.67 \cdot 10^{-8}; 5.35 \cdot 10^{-8}\right] / nt / y$, range: $\left[1.45 \cdot 10^{-10}; 5.47 \cdot 10^{-6}\right] / nt / y$). The mutation rate of *P. virginalis* is comparable to known mutation rates from other arthropods and falls between the human germline mutation rate and the somatic mutation rate of human cancers (Fig. 1B).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



158

159

160 **Figure 1.** Mutation rate of *P. virginalis* and coalescent. (A) Genealogy of laboratory
161 animals, with sequenced animals marked in grey. (B) Comparison of the calculated
162 mutation rate for *P. virginalis* with mutation rates for other arthropods and for *H.*
163 *sapiens*. Where available, standard deviations (red) and range (orange) are also
164 indicated. (C) Coalescent tree based on a constant mutation rate and sequences of
165 sampled animals. The posterior probability of each branch is indicated in red.

166

167 We also made a first evaluation of the evolutionary age of *P. virginalis*, using the
168 Bayesian evolutionary analysis software BEAST [35] for the sequences of 13 animals
169 (Fig. 1C), using 10 million states (see Methods for details). The convergence is
170 attained and effective sample sizes were adequately large (2984 or higher). The
171 results showed that animals 1, 34 and 35 correctly clustered together, as well as
172 animals from German wild populations (Hannover, Reilingen, Moosweiher) and from
173 the likely foundational laboratory lineage of the German wild populations

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Heidelberg). Furthermore, samples from Madagascar formed a separate branch. Interestingly, Petshop 2 [32] was nested in the branch of animals from Madagascar. This is consistent with the notion that the Malagasy population was founded by an animal that was originally obtained from a German pet shop. Posterior probabilities (Fig. 1C, red annotations) indicate highly probable branching for all but the top coalescent event, which has 0.5206 probability. From this tree, the most recent common ancestor of the 13 animals occured in 1988 (95% CI: $\lceil 1986.1; 1989.8 \rceil$). This is broadly consistent with the first documented appearance of *P. virginalis* in 1995 [34].

Dynamics of mutation accumulation in *P. virginalis*

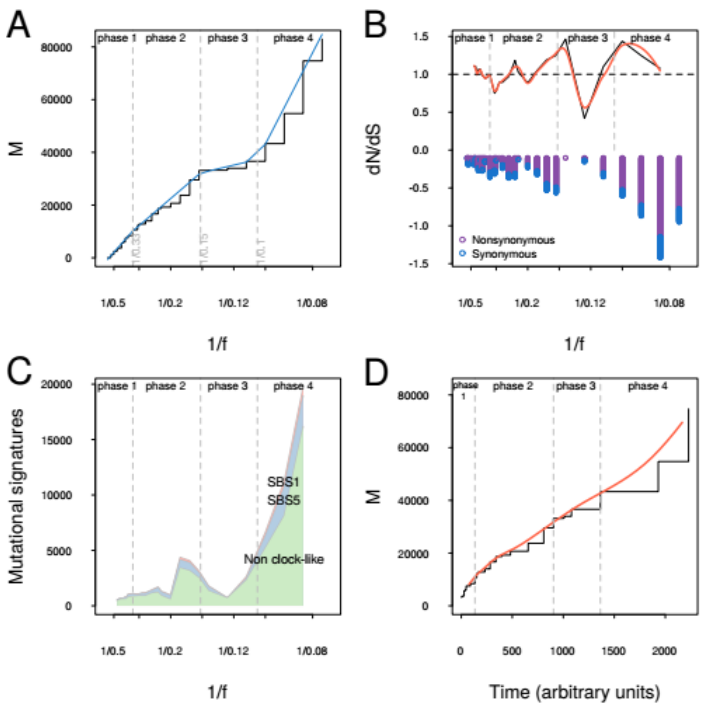
Knowing from theoretical results that mutation accumulation *M* as a function of (*1/f*) informs on the mutation rate, we looked at the dynamics of $M(1/f)$ in *P. virginalis*. The curve suggested that the mutation rate changed over time, with 4 phases defined by segmented regression (Fig. 2A; p=0.06). For example, the mutation rate was reduced in phase 3, and increased in phase 4 (Fig. 2A).

Since selection (denoted *s*) is not observable using *M(1/f)*, according to equation (6), we used the ratio of non-synonymous to synonymous mutations as a proxy for *s* (Fig. 2B). This profile showed values close to 1 for phases 1 and 2, suggesting the absence of selection (Fig. 2B). During phases 3 and 4, we detected *s* values >1, and <1, respectively (Fig. 2B), suggesting phases of positive and negative selection respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

196 We then used previously established clock-like mutational single-base signatures
197 (SBS1 and SBS5) [36, 37, 18] as a proxy for the time course of mutation
198 accumulation (Fig. 2C). We further assumed that the arrow of time from past to
199 present corresponds with the arrow of increasing *1/f*. We calculated the integral of the
200 clock-like components of mutation accumulation (see Methods for the details), which
201 yielded a time course in arbitrary units (Fig. 2D). The slope of this curve is
202 proportional to the mutation rate as a function of time. According to Fig. 2D, this
203 mutation rate exhibited little variation. As a result, our framework allowed the
204 analysis of mutations and selection dynamics at allele frequency resolution and at
205 time resolution.

206



207

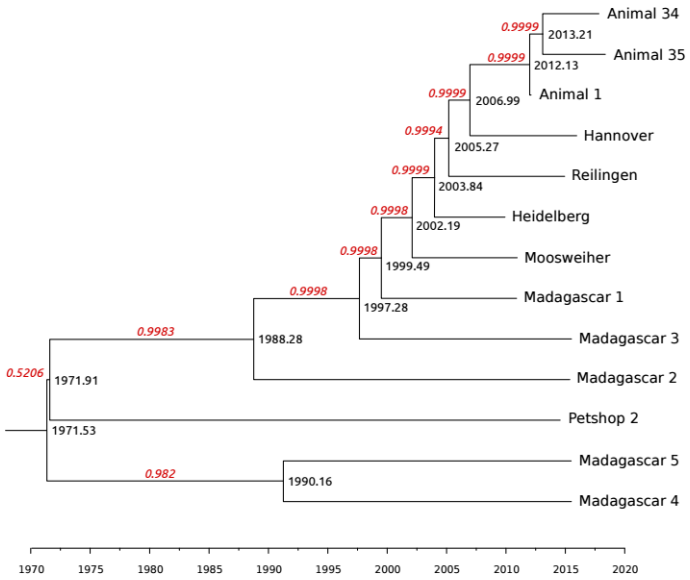
208 **Figure 2.** Mutation accumulation, selection and time course of *P. virginalis* genome
209 evolution. (A) Mutation accumulation as a function of the inverse allele frequency $1/f$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(black) and phases from automated segmentation (breakpoints in grey, segments in blue). (B) Non-synonymous to synonymous ratio (dNdS). The smoothed ratio is shown in red. (C) Comparison of clock-like and non-clock-like mutational signatures. (D) Mutation accumulation as a function of time. Smoothened mutation accumulation is shown in red.

Integrated analysis of P. *virginalis* genome evolution

Since we had obtained two complementary sources of information on the time course of mutation accumulation in *P. virginalis* (coalescent and mutation rate profile), we integrated both approaches. Assuming that the mutation rate of the most recent past corresponded to the mutation rate calculated for animals 1, 34 and 35, we recalculated the coalescent times and thus obtained a consolidated coalescent tree (Fig. 3). The resulting time estimate for the most recent common ancestor was 1971 $[1946.9; 1996.2]$ (95% CI) which is again consistent with the first report of the appearance of *P. virginalis* in 1995 [34].



bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3. Coalescent tree of *P. virginalis* evolution after integration of the mutation rate profile. The tree, branch posterior probabilities, and coalescence times of animals 1, 34 and 35 are unchanged, while the coalescence times of other animals were matched to the relative mutation rate profile derived from Fig.2D.

Clonal evolution landscape in glioblastoma tumors

Since glioblastoma is a high-grade glioma with systematic recurrence and poor patient survival, a better understanding of evolutionary parameters in this context would be of considerable importance. We therefore sought to apply our framework to a published set of whole-genome sequencing data of primary and recurrent glioblastoma tumors [19]. This study estimated in particular the age of primary tumors [19], allowing further integration in the following. Based on the curve $M(1/f)$, we generated mutation rate profiles (Fig. 4A, see Suppl. Fig. 1A for individual samples), which we further segmented into phases (Fig. 4A, $p<2.2\cdot10^{-16}$). The results indicated distinct variations in the mutation rate (Standard Deviation SD=63%, Inter-Quartile-Range IQR=133% relative to the mean mutation rate; Fig. 4A). Considering all primary tumor samples, SD varied from 54% to 141% (IQR: 71%-178%; Suppl. Table 2), and recurrent samples showed comparable variations (SD: 29%-139%, IQR: 29%-172%).

We next analyzed mutation selection using the dN/dS ratio. Taking confidence bounds into account, the results were compatible with neutral selection for most tumors (Fig. 4B, Suppl. Fig. 1B per sample). However, 11 primary tumor samples (2, 3, 6, 18, 22, 24, 30, 31, 35, 40 and 42) showed evidence of negative (purifying) selection during brief intervals, for instance sample 35 (Suppl. Fig. 1B for sample 2,

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

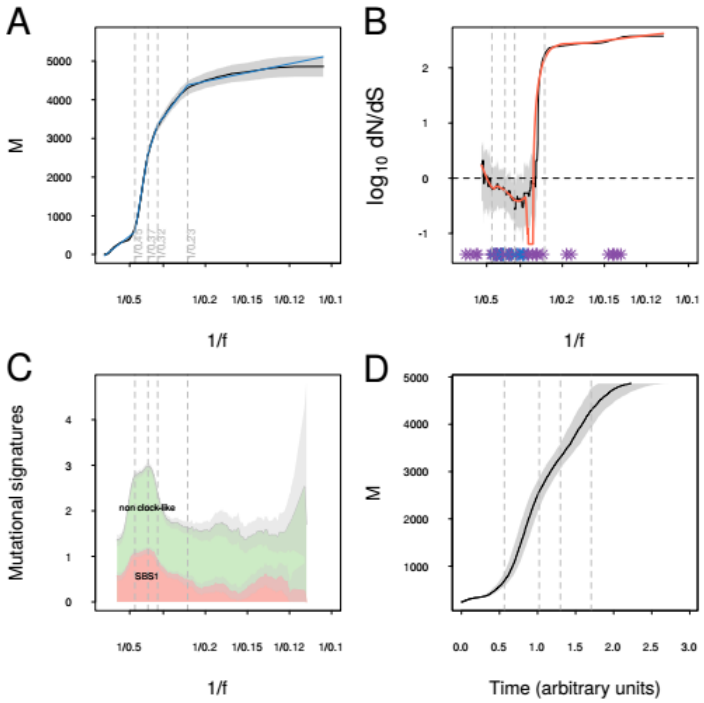
1/f in [1/0.5; 1/0.3]). We also observed evidence for positive selection in two primary tumor samples (Samples 2 and 7, Suppl. Fig. 1B). Interestingly, 7 out of 9 recurrent tumor samples (samples 3, 4, 13, 19, 27, 30 and 35) underwent prolonged phases of negative selection (for example, sample 4, 1/f in [1/0.5; 1/0.1], Suppl. Fig. 1B), while 2 samples (31 and 38) still exhibited short phases of negative selection. No recurrent tumor sample showed any significant phase of positive selection.

As a first step to determine the timeline of tumor evolution, we analyzed canonical tumor-associated single-base substitution (SBS) signatures [4]. More specifically, we compared the prevalence of the stable, clock-like SBS1 signature to the other, non clock-like SBS signatures (Fig. 4C, see Suppl. Fig. 1 for individual samples). This includes SBS5, which is non clock-like in glioblastoma according to [36] (Table1). This was confirmed in most samples analyzed, since there was little correlation between SBS1 and SBS5 (ρ_P=-0.060 with IQR=0.143 in primary tumors; ρ_P=0.067, IQR=0.136 in recurrences). Surprisingly, the few samples under selection displayed larger correlation coefficients, suggesting that SBS5 might also be clock-like in this subset of samples (Suppl. Fig. 2, p=0.01916, ρ_P=0.32 in primary tumors; p=0.02188, ρ_P=0.52 in recurrences). The non-clock like SBS signature prevalence was 3.334 (IQR = 0.477) fold higher than the clock-like SBS1 signature in the primary tumors (4.049 fold higher, IQR= 1.074, in recurrences; Suppl. Table 2).

Using the information on the clock-like signature SBS1, and using equation (9) (Methods), we reconstructed M as a function of time (Fig. 4D, Suppl. Fig. 1D per sample), in arbitrary units. The slope of this curve is proportional to the mutation rate per time unit. Similar to the mutation rate per division in Fig. 4A, the mutation rate per time unit exhibited large variations, with SD in [22.8%; 117.9%] and IQR in

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

273 [26.5%; 142.4%] in primary tumors, and SD in [28.1%; 466.4%] and IQR in [20.6%;
274 117.0%] in recurrent tumors (Suppl. Table 2). In conclusion, the evolutionary
275 landscape of glioblastoma revealed notable variations in the mutation rate per division
276 and per unit of time in all samples. Some evidence for selection was also detectable,
277 with a putative association with the clock-like status of the SBS5 signature.



278

279 **Figure 4.** Mutation accumulation, selection and time dynamics of a representative
280 glioblastoma tumor (patient 1, primary tumor). (A) Mutation accumulation as a
281 function of the inverse allele frequency $1/f$ (black) and phases from automated
282 segmentation (breakpoints are indicated as dashed vertical lines, segments are
283 indicated in blue). (B) Non-synonymous to synonymous ratio dNdS. Purple and blue
284 stars show non-synonymous and synonymous mutations, respectively. The
285 smoothened ratio is shown in red. (C) Clock-like and non-clock-like mutational
286 signatures. (D) Mutation accumulation as a function of time.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

287

288 **Expansion parameters of the primary and recurrent tumors**

289 Using mutation accumulation as a function of 1/f or time from Fig. 4A, D and
290 equation (17) (Methods), we could reconstruct the product $\omega\gamma N$ of the tumor
291 parameters growth rate ω, tumor cell survival rate γ and number of cells N (Fig. 5A).
292 This product corresponds to the expansion parameters of the tumor, hence yielding
293 insights into the way how the tumor develops during the primary and recurrent
294 phases, respectively (left and right panels of Fig. 5A). The curves $\omega\gamma N$ for sample 1
295 and other samples (Fig. 5A, Suppl. Fig. 3A for individual samples) displayed an
296 overall increase in the primary tumor, except for samples 5 and 16. This increase was
297 also observed in the recurrence phase for 29 samples out of 42, while 13 samples
298 displayed an overall decrease. A decrease might be attributed to a declining growth
299 rate ω or tumor cell survival rate γ, while a decrease of N can be excluded in
300 principle (unobservable). Furthermore, $\omega\gamma N$ curves sometimes had a simple form with
301 one local minimum, and sometimes a more complex pattern, with one or several
302 additional local maxima and minima (Fig. 5A, Suppl. Fig. 3A for individual samples).

303 We then looked at a possible association between the patterns of the $\omega\gamma N$ curve in the
304 primary tumors, and the time to the recurrence (Fig. 5C). We first sorted curves into
305 the following categories 1: convex, 2: stable, then convex, 3: double convex, 4:
306 increasing (Suppl. Fig. 4A-D). The pattern of these curves in the primary tumors was
307 associated to the differential time to recurrence, though this was non-significant after
308 p-values adjustment (p=0.04035, p_{adj}=0.28245, Suppl. Fig. 4E, F). Further, while
309 some patterns remained identical between the primary tumor and the recurrence,
310 curves of type 2 preferably led to types 1 or 3 (Suppl. Fig. 4G, n=42). Conversely,

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

type 3 never led to type 2. We then sought to confirm this manual analysis with a systematic approach, relying on segmentation and automatic detection of minima and maxima. Interestingly, we observed an association of the variance of time to recurrence with the presence of one local maximum (p=0.0181, p_{adj}=0.0362, n=20). A similar trend, non-significant at 5% type I error level, appeared for the count of maxima during recurrence (Fig. 5D, p=0.0367, p_{adj}=0.0734, n=20). The number of local maxima of $\omega\gamma N$ curves was correlated between the primary tumor and recurrence, in the whole set of 42 samples (ρ_P= 0.34). Hence, the patterns of $\omega\gamma N$ curves in the primary tumor are indicative of the expansion pattern in the recurrence, as well as the time to recurrence. The time to recurrence was 17.4 months (SD=12.9) in the subgroup with no maximum, suggesting a more favorable prognosis as compared to the subgroup with at least 1 maximum (8.7 months, SD=3.1), although this did not reach statistical significance (p=0.2485, n=20). The possible more favorable prognosis for the subgroup with no maximum was better explained by a larger variance (16.77 times higher if no maximum, p=9.343×10^{-5}, p_{adj}=1.868×10^{-4}, 95% CI= [4.04 ; 239.06]) than in the subgroup with 1 maximum or more.

Tumor cell survival at the transition from primary to recurrent tumor

Since the time difference between the resection of the primary tumor and the resection of the recurrence is known for a subset of samples [19], this allowed us to calibrate the time course from arbitrary units into real units (Eq. 10 in Methods, Fig. 5B, and Suppl. Fig. 3B per sample). Furthermore, the transition from the primary to the

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

recurrent tumor can be expressed formally (Eq. 11, Methods) and simplified using continuity assumptions (Methods), resulting in the following equation:

$$\left(\frac{dM}{dt}\right)_P = \frac{(1/f)_P}{(1/f)_R} \cdot \frac{\gamma_P}{\gamma_R} \cdot \left(\frac{dM}{dt}\right)_R. \quad (7)$$

Equation (7) describes how mutation accumulation of primary tumors at resection (index P) and at the initiation of recurrence (index R), are linked by only their respective allele frequencies and tumor survival rate. An important assumption is that the mutation rate (μ) and the growth rate (ω) are constant between the primary tumor and the recurrence, based on the argument that these are intrinsic characteristics of the tumor, which are unlikely to change notably over a short time span. Conversely, in expression (7), the ratio for tumor survival rate in primary tumor to the recurrence, γ_R/γ_P, cannot be taken as constant, since tumor cell survival changes drastically at resection and following treatment [19]. The ratio was therefore set to 300:1, based on the consideration that oxygen, room to expand and other resources supply might be much higher after resection of the primary tumor. Using this value, and since other parameters were known, we reconstructed the time course of mutation accumulation for the primary tumor (Fig. 5B, Eqs. 14, 15 in Methods), determining a progression time of 6 years for the primary tumor of sample 1, in agreement with published data [19]. This suggested that the chosen value for the tumor cell survival ratio was a reasonable assumption for this sample.

Using equation (7) in the reverse way, we determined the tumor survival ratio from time estimates. Previous analyses [19] indicated that the primary tumors could have emerged from 2 to 7 years before diagnosis. Using the values of 2 years and 7 years as

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

355 the lowest and highest limits for the time course of the primary tumor, we could
356 determine a range for the value of the tumor survival ratio γ_R/γ_P for each individual
357 sample (Eq. 16 in Methods, Fig. 5E, Suppl. Table 3). As a result, the lowest value of
358 the ratio γ_R/γ_P, corresponding to a tumor emergence about 2 years before diagnosis,
359 was always higher than 1, and had a median value of 27.8 (95% CI=[17.4; 54.0],
360 n=20 samples). As for the upper bound for the range of γ_R/γ_P ratio, corresponding to
361 tumor emergence 7 years before diagnosis, the median γ_R/γ_P ratio was 97.5 (95%
362 CI=[60.9; 189.0]). These results indicated that tumor cell survival was higher at the
363 start of the recurrence than at the end of the primary tumor growth. Notably, the
364 variability between samples was considerable, with some samples being close to the
365 unity ratio (samples 39, 40, 42), indicating similar tumor cell survival before and after
366 resection. Not surprisingly, γ_R/γ_P ratios were associated to the time to recurrence
367 (Fig. 5F, adjusted-R^2=0.44, p=8.386×10^{-4}, p_{adj}=1.258×10^{-3} for regression line 1,
368 adjusted-R^2=0.61, p=2.883×10^{-5}, p_{adj}=8.649×10^{-4} for regression curve 2), with higher
369 γ_R/γ_P ratios corresponding to shorter time to recurrence. Further, the variance of
370 ratios γ_R/γ_P was also associated to the number of local maxima in the primary tumors
371 (p=0.01545, p_{adj}=0.04635). This further suggested that the expansion pattern of the
372 primary tumor can predict the progression to recurrence.

373

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

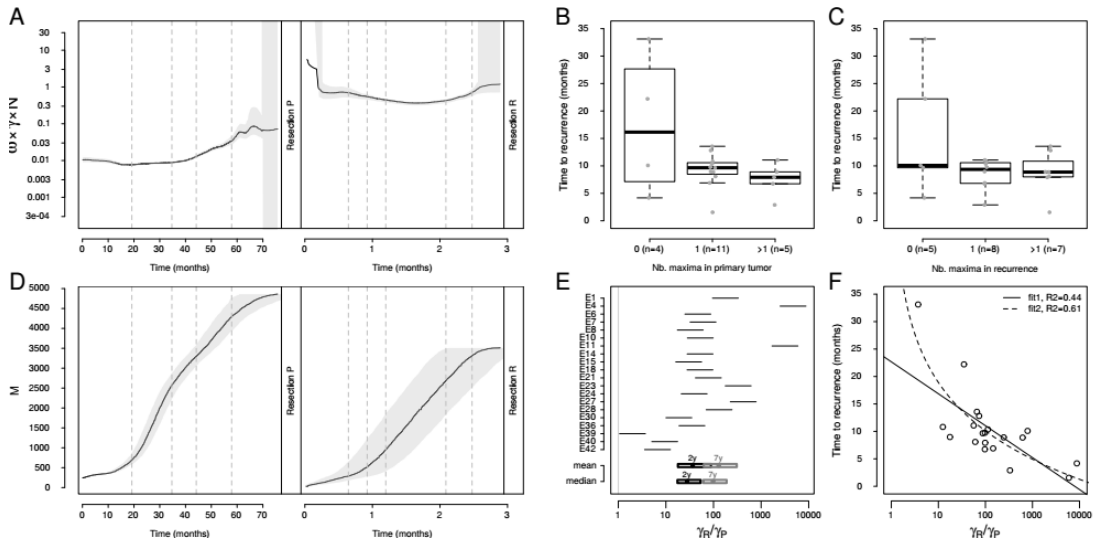


Figure 5. Transition of primary tumor to recurrent tumor for patient 1. (A) Dynamics of growth rate ω times tumor cell survival rate γ times number of cells N, for (P) the primary tumor and (R) the recurrence. (B) Time to recurrence dependence on ωγN characteristics in the primary tumor, and (C) in the recurrence. (D) Time-resolved mutation accumulation for primary tumor and recurrence. (E) Neoplastic cell survival in recurrence relative to the primary tumor, denoted γ_R/γ_P, for the lower and higher limits where tumor emergence dates back to 2 years, or to 7 years. (F) Dependence of time to recurrence on the γ_R/γ_P ratio. Fit1 corresponds to a linear regression of time versus $\log_{10}(\gamma_R/\gamma_P)$, with intercept=19.511 (standard error SE=2.544) and slope=-5.819 (SE=1.455), fit2 corresponds to a linear regression of time versus $\log_{10}(\log_{10}(\gamma_R/\gamma_P))$, with intercept=18.922 (SE=1.806) and slope=-29.321 (SE=5.285).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Discussion

Our study suggests a key role of evolution dynamics in the primary tumor and of tumor cell survival at resection, in glioblastoma. This and other aspects examined in this study were brought to light thanks to the intimate combination and integration of theoretical results with molecular data, in order to determine detailed and time-resolved characteristics of clonal genome evolution. In that, our study is the first to demonstrate the viability and strength of such a comprehensive approach. Replication and extension studies should help complement our results with additional insights and potential clinical applications.

Our study also validated the clonal genome model *P. virginalis*, with a base mutation rate for this animal of $3.51 \cdot 10^{-8}/nt/y$, close to the mutation rate observed in human somatic evolution in healthy tissues and in cancer, though some discrepancies may arise because of differentiation, adaptation timescales, and environmental switches [38]. This model was instrumental in developing our approach, and further demonstrated the utility of integrating different sources of information, which yielded a refined estimated time to the most recent common ancestor in 1971 (95% confidence limits: $\lfloor 1946.9; 1996.2 \rceil$), in agreement with first reports of this animal in 1995 [34], and consistent with a very young evolutionary age of the species.

The expansion profile in the primary tumors and also the tumor cell survival ratio around resection, were both associated with the time to recurrence, and with the expansion profile during recurrence. This suggests a predictive value for evolutionary

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

parameters in the primary tumor, with respect to timing and expansion characteristics during recurrence. Logically, earlier diagnosis should in principle prevent the primary tumor from reaching a complex expansion profile and hence lead to a more favorable prognosis. Few studies have proposed biomarkers or a mechanistic explanation based on clonal evolution [39, 40, 41, 42, 43, 44], so that a potential prognostic biomarker based on the expansion profile would be valuable.

Interestingly, tumor cell survival was systematically higher in the recurrence. This supports the notion that tumor regrowth is more aggressive after surgical resection of primary glioblastoma tumors [45, 46], possibly because resection-induced astrocyte injury can support faster growth [46]. Additionally, space and oxygen can promote tumor regrowth [47, 48, 49], while a stronger immune response after resection would reduce it. In this regard, our large range of tumor cell survival rates might reflect the different balances between these (and other) parameters.

The impossibility to observe selection using allele frequency alone had been suggested before [11, 12], and is shown here, in the frame of minimal, reasonable assumptions [50, 51, 52]. This prompted us to use the dNdS ratio instead. While most glioblastoma samples showed no or little signs of selection, about 25% of samples did exhibit short (in primary tumors) or longer (in recurrences) phases of selection. This provides an important complement to recent studies, that have either claimed a major role of selection [3] or its complete absence [13, 9]. These seemingly contradictory findings may be explained by the pace of tumor growth. In slowly growing populations, the genome is shaped by random drift and selection [50, 51]. Conversely, faster growth rates render random drift negligible [50, 51] and selection less

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

observable [53]. This indicates that the observed presence of selection in a subset of samples might correspond to phases of slower tumor growth.

We further noted variations of the mutation rate, both in *P. virginalis* and in glioblastoma. This supports the argument that the mutation rate should not be considered constant [1, 54, 55]. Known mechanisms can explain these variations, including a temporarily more pronounced effect of error-prone mechanisms, hypoxia-induced mutagenesis, or transcription-associated mutagenesis [1, 56]. Interestingly, the mutational signature SBS5 exhibited a correlation with the clock-like, m5C-deamination related signature SBS1, in a subset of glioblastoma samples under selection. While this could be a spurious finding, it could also help to understand the etiology of SBS5.

It will be important to validate the prognostic value of evolution characteristics in independent tumor datasets. In addition, since our characterization relies on the entire tumor, with subclones inherently considered in the mathematical framework, it would be interesting to complement the analysis with local biopsies or on the single-cell level. Assuming that complex evolution patterns in the primary tumor are explained by its adaptation, this could shed light on the identity of responsible subclones or interaction of subclones, and possibly provide novel mechanistic explanations. Also, integration of multi-omics datasets could help disentangle the individual roles of tumor expansion parameters and the connections between genotypes and phenotypes.

In conclusion, our integrated analysis of theoretical results, mutation accumulation, dNdS ratio and mutational signatures revealed a detailed picture of the expansion

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

452 dynamics, time course and survival of tumor cells in glioblastoma samples, and
453 validated the marbled crayfish as a useful animal model for studying clonal genome
454 evolution. In particular, our results suggested that tumor dynamics as well as the time
455 to recurrence were predicted by the parameters of the primary tumor, with a longer
456 time to recurrence, and hence a longer patient survival in the subgroup with the least
457 complex dynamics. Remarkably, survival of neoplastic cells was shown to be
458 systematically higher after resection than before resection, and a lower survival of
459 neoplastic cells in the recurrence was associated with a longer time to recurrence.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4 Methods

***Procambarus virginalis* samples.** Freshwater crayfish samples from [32] were used. Additionally, samples from animal 1, Madagascar 1 sample and Moosweiher sample were resequenced. Animal 1 corresponds to the lab strain, acquired from a pet shop (Suppl. Table 1). New genomic DNA samples were taken from animal 34 and animal 35, which, as animal 1, also correspond to lab strains animals, and which are direct offsprings of animal 1. These new samples were prepared and submitted for whole genome sequencing following the protocol already described. The genealogy and birth date of animals were retrieved from laboratory records and field records (Suppl. Table 1). Sequence data was trimmed using Trimmomatic v0.32 (settings: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:40, adapter sequence: TruSeq3-PE). Next, trimmed data was mapped to Pvir genome assembly v04, using bowtie2 (v2.2.6, setting: --sensitive). Aligned reads were sorted, cleared from duplicates, sorted and indexed using samtools. Subsequently, variant calling was performed using freebayes v0.9.21-g7dd41db (parameters: --report-all-haplotype-alleles -P 0.7 -p 3 --min-mapping-quality 30 --min-base-quality 20 --min-coverage 6 --report-genotype-likelihood-max).

Glioblastoma Multiforme samples. The glioblastoma primary and recurrent tumor samples correspond to the WGS cohort already described in Koerber *et al.* (2019). In particular, summary information can be found in supplementary table 1 of [19]. After approval of the research project, access to the SNP data of primary and recurrent tumor samples, as well as time to recurrence when available, was granted.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Study of mutation accumulation. An infinitesimal increment of mutations was defined from the mutation rate, ploidy, cell survival, growth and number of cells:

$$dM(t) = \mu(t) \cdot \pi(t) \cdot G \cdot 2 \cdot \omega(t) \cdot \gamma(t) \cdot N(t) \cdot dt. \quad (8)$$

We have stratified this expression for each subclone (see Supplementary Demonstration for details about this and for the proof of each of the following steps). We have then determined the relationship between the observable allele frequency of a mutation, which is the one obtained after sequencing and SNP calling, and the features of the subclone where this mutation appeared. Next, we have determined a formula for the increment of the number of cells *dN* and for the increment of inverse allele frequency *d(1/f)*. For this latter increment, we have made the assumption that ploidy was constant. Using intermediate equations, we could then deduce the mutation accumulation dM_i as a function of inverse allele frequency $1/f_i$, in each subclone i. Finally, the equation for mutation accumulation over all subclones *dM* was obtained by summing the individual contributions dM_i of each subclone.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation annotation and dNdS ratio. Mutations were annotated as synonymous or non-synonymous (including splice or stopgain mutations) using SNPdat v1.0.5. The dNdS ratio was calculated as the quotient of non-synonymous mutations by synonymous mutations in a sample, divided by the average quotient in the full genome. The average quotient of non-synonymous to synonymous in humans was taken equal to 3.34951759 (hg19).

Mutational signatures. Mutational signatures for human subjects were downloaded from the COSMIC database (https://cancer.sanger.ac.uk/signatures/ ; version 3.1 as of 11.08.2020). Mutation data was binned using a bin half-width of 0.5 on the inverse allele frequency. Exposure of binned data was determined using R 3.5.2 with package YAPSA (version 1.8.0). Uncertainty on mutational signatures were determined by bootstrap resampling of mutations and generation of the binned data and YAPSA exposures on the resampled data. We have used 1000 bootstrap replicates as a compromise between an ideally larger (1M) number of replicates, and reasonable computing time. Large mutation sets (>100,000 mutations) were subsampled to 50,000-60,000 mutations for the bootstrap analysis. Mean, median, percentiles and 95% confidence bounds were determined using the resulting bootstrap distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Time course. We have utilized clock-like mutational signatures SBS1 and SBS5 as a surrogate indicator of time (for glioblastoma, SBS1 only, in agreement with [36]). We obtained an increment of time by integrating the number of mutations which are clock-like over an increment of inverse allele frequency *1/f*. Considering that the number of mutations is also proportional to the number of cells in the tumor, we further wanted to standardize the count of clock-like mutations, by dividing this count by the number of cells *N*. Since 1/*f* is proportional to *N* [9], we multiply by *f* instead of dividing by *N*. This yielded the formula for determining time *t* from integration of clock-like mutational signatures θ over the range of the inverse allele frequency *1/f*:

$$t_{a.u.} = \int_{(1/f)min}^{(1/f)max} \theta \cdot f \cdot d(1/f). \quad (9)$$

This evaluation of time is in arbitrary units (a.u.). For some glioblastoma samples, the time-to-relapse is known. We have used this time-to-relapse, denoted here τ, to calibrate the time course in the recurrence to real units, in months:

$$t = \tau \cdot t_{a.u.} / max(t_{a.u.}). \quad (10)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to propagate the time calibration to the time course of the primary tumor, it was necessary to determine a practicable link between these two phases. To this aim, we have looked at the ratio of mutation accumulation between the end of primary tumor (subscripted 'P', taken as the last 5% time points) and start of recurrence (subscript 'R', first 5% time points). The passage from primary tumor to recurrence effectively corresponds to the instant of primary tumor resection. Using equation (X) above, this ratio could be written as follows:

$$\frac{(dM/dt)_P}{(dM/dt)_R} = \frac{(\mu \cdot \pi \cdot G \cdot 2 \cdot \omega \cdot \gamma \cdot N)_P}{(\mu \cdot \pi \cdot G \cdot 2 \cdot \omega \cdot \gamma \cdot N)_R}. \quad (11)$$

The constants normalized out of this ratio. Further, we have assumed that ploidy π, mutation rate μ and division rate ω stay constant over this short period, because they are inherent features of the tumor cells. However, the count of tumor cells *N* wasn't constant. We expressed it as the ratio of inverse allele frequency, since it is proportional to *N:*

$$\frac{N_P}{N_R} = \frac{(1/f)_P}{(1/f)_R}. \quad (12)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Equation (12) is perturbed in practice by mutations which are not de novo in the recurrence, but inherited from the primary tumor. Ideally, only de novo mutations should be included to perform this calculation. Finally, the survival rate of tumor cells, γ, also couldn't be considered constant, and we had no indicator or surrogate for this value. For this reason we have set an arbitrary value for the survival at end of primary tumor, relatively to the start of recurrence, $\gamma_P/\gamma_R = 1/300$.

Using the above, dM/dt at end of primary tumor could be determined:

$$\left(\frac{dM}{dt}\right)_P = \frac{(1/f)_P}{(1/f)_R} \cdot \frac{\gamma_P}{\gamma_R} \cdot \left(\frac{dM}{dt}\right)_R . \quad (13)$$

Since the number of mutations at end of primary tumor was known, and since the rest of parameters was known, the time in real units at the end of primary tumor could be calculated as follows:

$$dt_P = \frac{(dM)_P}{\frac{(1/f)_P}{(1/f)_R} \cdot \frac{\gamma_P}{\gamma_R} \cdot \left(\frac{dM}{dt}\right)_R} . \quad (14)$$

Finally, the time course of the primary tumor in arbitrary units was scaled to real units, using the known point at the end of primary tumor:

$$dt = dt_{a.u.} \cdot \frac{dt_P}{dt_{a.u.,P}} . \quad (15)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tumor cell survival ratio. For time calibration to real units, we have made an assumption on tumor cell survival ratio γ_R/γ_P to determine real time in the primary tumor. To quantify the survival ratio, we have proceeded the other way around, using an assumption on the time course in the primary tumor, in order to determine the ratio γ_R/γ_P. We have taken the assumption that the time between the most recent common ancestor (TMRCA) lies either 2 years or 7 years before primary tumor resection. These durations correspond to the shortest and longest time spans from Koerber et al. (2019). The calculation of the tumor cell survival ratio was done using the following equation:

$$\frac{\gamma_R}{\gamma_P} = \frac{\left(\frac{dM}{dt}\right)_R}{\frac{(1/f)_R}{(1/f)_P} \cdot \left(\frac{dM}{dt}\right)_P}. \tag{16}$$

Tumor expansion profile. From equation (4) in Suppl. Demonstration, time and 1/f are proportional, with modulators growth rate ω, tumor cell survival rate γ, and number of tumor cells *N*:

$$d(1/f) \propto \omega \cdot \gamma \cdot N \cdot dt. \tag{17}$$

As a consequence, dividing increment d(1/f) by increment d(t) yielded the product ωγN, which we denoted expansion profile (or pattern).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

572 **Expansion profile analysis.** We first noticed and identified 4 curve categories by
573 visual inspection (Suppl. Fig. 4), and next manually classified expansion curves to
574 one of these 4 categories. For the automated procedure, curve segments for M(1/f)
575 were determined using R package segmented (v1.1-0), using objective R^2 set to
576 0.9995, and using the lowest number of segments which attained this objective,
577 limited to a maximum of 20 segments. For curves $\omega\gamma N(t)$, the number of segments
578 was tailored to allow annotation of most visible local minima or maxima. The count
579 of extrema was then derived from the changed slope from one segment to the next.
580 This count had to be further curated in a subset of samples.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation rate of *Procambarus virginalis*. Mutation accumulation between animals 1 and its offsprings animals 34 and animals 35 was used to calculate the mutation rate. SNP variants were examined in terms of quality and coverage (Suppl. Fig. 5). Quality at least 35 and coverage at least 50 and no more than 200 was retained for the main estimate of the mutation rate. Coverages 200 and higher exhibited altered SNP distribution (Suppl. Fig. 5), and have thus been excluded because possibly corresponding to a distinct part of *P. viriginalis* genome (possibly highly repetitive and variable domains). For the minimum estimate of the mutation rate, we have used more stringent quality filters, with quality >39 and coverage in [50-200]. For the maximal estimate of the mutation rate, we have used relaxed quality filters, with quality >35 and coverage in [25-200]. Subsequently, the mutation rate per nucleotide per year was calculated as the count of biallelic mutated nucleotides in animal 34 (respectively, animal 35) as compared to animal 1, divided by the count of nucleotides in the triploid genome of *P. virginalis*, and divided by the period of time, in years, between animal 1 and animal 34's births (respectively, birth date of animal 35). We have made a thorough uncertainty analysis. We first determined standard deviation on the count of mutations observed assuming that this count follows a Poisson distribution of new mutations (genotyping uncertainty). Second, we used a third of the total uncertainty on time of birth as the standard deviation for the date of birth. Third, we calculated the standard deviation between animal 34 and animal 35's mutation rates. Finally, we combined these three standard deviation components using a quadratic sum (since considering that at least two of these variance components follow a normal distribution).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Coalescent time. Time to most recent common ancestor for *P. virginalis* samples was determined using Bayesian evolutionary analysis by sampling trees (BEAST v1.10.4 ; Note: at the time of download and installation, BEAST2 didn't offer any additional model, but rather a multiplatform component which wasn't needed here). Mutation data with quality >35 and coverage depth >15 was used in this analysis (a coverage cutoff of 25 was not justified here because samples other than animals 1, 34 and 35 possessed a notably lower average sequencing depth). Samples birth dates were used as tip dates. Further BEAST parameters used were: simple substitution model with estimated base frequencies, strict clock, skyride coalescent prior, Markov chain Monte Carlo length of 10M. The resulting dates were rescaled to match the exact time durations known for animals 1, 34 and 35. We further modulated the coalescent time with the mutation rate profile in replacement of the strict clock, using the following equation:

$$t_0 = t_f - \frac{1}{\mu_r} \cdot \left(t_f - t_{b,0}\right). \quad (18)$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this equation, t_f is the end timepoint, t_0 is the initial time point, here the time to most recent common ancestor, $t_{b,0}$ is the initial time point before rescaling and μ_r is the temporal mean of the mutation rate profile divided by the mutation rate at the end timepoint. The mutation rate profile was taken as the slope of dM/d(1/f), in agreement with equation (6). For the first time point, the temporal mean could be calculated immediately, using the full profile, but for intermediate coalescent timepoints t_i, the time interval where the temporal mean should be calculated was not known beforehand. Because of that, we first used the full mutation rate profile to estimate $t_{i,1}$. Then, we proceeded by iterations, using $t_{i,1}$ to recalculate $\mu_r(t_{i,1})$ which in turn was used to determine $t_{i,2}$. We repeated these steps ten times, which ensured $|t_{i,10} - t_{i,9}| < 0.001$ for all i.

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

631 **Statistical analyses.** R [57] was used for all statistical analyses, as well as for
632 bootstrap calculations (except a few instances carried out in Python [58] and summary
633 statistics mean, median, quantiles). All statistical tests were unpaired and two-sided;
634 with the level of significance set at 5%. Segmentation p-values were extracted from
635 the output of R package segmented. Correlation coefficients between SBS1 and SBS5
636 were determined using Pearson method, and summarized by their median and IQR
637 over the 42 samples, and a comparison between the group under selection or not was
638 made using a Wilcoxon rank-sum test. A differential time to recurrence between
639 subgroups in the manually sorted $\omega\gamma N$ curves was assessed using a wilcoxon rank-
640 sum test against curve type 3. Differences on the time to recurrence in the systematic
641 analysis was explored by analysis of variance against the number of maxima in the
642 primary tumors and in the recurrence, with Bonferroni adjustment of *p*-values. The
643 difference of time to recurrence between subgroups based on the number of maxima
644 in the primary tumors was further explored using a non-parametric wilcoxon test and
645 using a F-test of comparison of variances. A possible association of the γ_R/γ_P ratio
646 (n=20) with the pattern of $\omega\gamma N$ curves was investigated using a F-test of comparison
647 of variances. A possible association of the γ_R/γ_P ratio with the time to recurrence was
648 assessed with a linear regression, using a simple or double $\log_{10}$ scale on the γ_R/γ_P
649 ratio, with Bonferroni adjustment.

650

651 **Data availability**

652 Newly sequenced marbled crayfish data have been deposited as a National Center for
653 Biotechnology Information BioProject (to be completed ; accession number XXXX),

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

654 while reanalysed data is accessible as a BioProject as well (accession number
655 PRJNA356499). Glioblastoma data corresponds to accession number :
656 EGAS00001003184 at the European Genome-phenome Archive (EGA).

657

658 **Acknowledgements**

659 We would like to thank Verena Körber and Thomas Höfer for helpful discussions and
660 for providing data, and Julian Gutekunst for discussions about the methods. We would
661 also like to thank Katharina Hanna for data and for crayfish culture, and Sina Tönges
662 for sample processing. We further acknowledge the German Cancer Research Center
663 Genomics and Proteomics Core Facility for whole-genome sequencing.

664

665 **Authors contributions**

666 C.L. and F.L. conceived and designed the analysis. R.A. carried out fieldwork and
667 contributed to the conceptual development, P.L. contributed data, C.L. developed the
668 theory and performed the analysis, C.L. and F.L. wrote the manuscript.

669

670 **Competing interests**

671 The authors declare no competing interests.

672

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] Lynch, M., et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, **17,** 704–714 (2016).

[2] Kent, D. G., and Green, A. R. Order matters: The order of somatic mutations influences cancer evolution. *CSH. Perspect. Med.*, **7,** a027060 (2017).

[3] Martincorena, I., et al. Universal patterns of selection in cancer and somatic tissues. *Cell*, **171,** 1029–1041.e21 (2017).

[4] Alexandrov, L. B., et al. Signatures of mutational processes in human cancer. *Nature*, **500,** 415–421 (2013).

[5] Tataru, P., Simonsen, M., Bataillon, T., and Hobolth, A. Statistical inference in the wright-fisher model using allele frequency data. *Syst. Biol.*, **66,** e30–e46 (2017).

[6] Benton, M. L., et al. The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.*, **22,** 269–283 (2021).

[7] Martincorena, I., and Campbell, P. J. Somatic mutation in cancer and normal cells. *Science*, **349,** 1483–1489 (2016).

[8] Katju, V., and Bergthorsson, U. Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. *Genome Biol. Evol.*, **11,** 136–165 (2019).

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

[10] Rosen, Z., and Bhaskar A, Roch S, S. Y. Geometry of the sample frequency spectrum and the perils of demographic inference. *Genetics*, **210,** 665–682 (2018).

[11] Niida, A., Iwasaki, W. M., and Innan, H. Neutral theory in cancer cell population genetics. *Mol. Biol. Evol.*, **35,** 1316–1321 (2019).

[12] Bozic I, Paterson C1, W. B. On measuring selection in cancer from subclonal mutation frequencies. *Plos Comput. Biol.*, **15,** e1007368 (2019).

[13] Sottoriva, A., et al. A big bang model of human colorectal tumor growth. *Nat. Genet.*, **47,** 209–216 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[14] Noorbakhsh, J., and Chuang, J. H. Uncertainties in tumor allele frequencies limit power to infer evolutionary pressures. *Nat. Genet.*, **49,** 1288–1289 (2017).

[15] Tarabichi, M., et al. Neutral tumor evolution ? *Nat. Genet.*, **50,** 1630–1633 (2018).

[16] Balaparya, A., and De, S. Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. *Nat. Genet.*, **50,** 1626–1628 (2018).

[17] McDonald, T. O., Chakrabarti, S., and Michor, F. Currently available bulk sequencing data do not necessarily support a model of neutral tumor evolution. *Nat. Genet.*, **50,** 1620–1623 (2018).

[18] Baez-Ortega, A., et al. Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science*, **365,** eaau9923 (2019).

[19] Körber, V., et al. Evolutionary trajectories of idhwt glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell*, **35,** 692–704.e12 (2019).

[20] Gerlinger, M., et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, **366,** 883–892 (2012).

[21] Sun, R., et al. Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.*, **49,** 1015–1024 (2017).

[22] Werner, B., et al. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat. Commun.*, **11,** 1035 (2020).

[23] Lee-Six, H., et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, **561,** 473–478 (2018).

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

[25] Ellsworth, D. L., et al. Single-cell sequencing and tumorigenesis: improved understanding of tumor evolution and metastasis. *Clin. Transl. Med.*, **6,** 15–15 (2017).

[26] Basanta, D., Simon, M., Hatzikirou, H., and Deutsch, A. Evolutionary game theory elucidates the role of glycolysis in glioma progression and invasion. *Cell proliferat.*, **41,** 980–987 (2008).

[27] Zhang, J., Cunningham, J. J., Brown, J. S., and Gatenby, R. A. Integrating evolutionary dynamics into treatment of metastatic castrate-resistant prostate cancer. *Nat. Comm.*, **8,** 1816 (2017).

[28] Kersten, K., de Visser, K. E., van Miltenburg, M. H., and Jonkers, J. Genetically engineered

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mouse models in oncology research and cancer medicine. *EMBO Mol. Med.*, **9,** 137–153 (2017).

[29] Blanpain, C. Tracing the cellular origin of cancer. *Nat. Cell Biol.*, **15,** 126–134 (2013).

[30] Miles, W. O., Dyson, N. J., and Walker, J. A. Modeling tumor invasion and metastasis in drosophila. *Dis. Model. Mech.*, **4,** 753–761 (2011).

[31] Kirienko, N. V., Mani, K., and Fay, D. S. Cancer models in caenorhabditis elegans. *Dev. Dynam.*, **239,** 1413–1448 (2011).

[32] Gutekunst, J., et al. Clonal genome evolution and rapid invasive spread of the marbled crayfish. *Nat. Ecol. Evol.*, **2,** 567–573 (2018).

[33] Maiakovska, O., et al. Genome analysis of the monoclonal marbled crayfish reveals genetic separation over a short evolutionary timescale. *Commun. Biol.*, **4,** 74 (2021).

[34] Lyko, F. The marbled crayfish (decapoda: Cambaridae) represents an independent new species. *Zootaxa*, **4363,** 544–552 (2017).

[35] Drummond, A., and Rambaut, A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7,** 214 (2007).

[36] Alexandrov, L. B., et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.*, **47,** 1402–1407 (2015).

[37] Petljak, M., et al. Characterizing mutational signatures in human cancer cell lines reveals episodic apobec mutagenesis. *Cell*, **176,** 1282–1294.e20 (2019).

[38] Merlo, L. M., Pepper, J. W., Reid, B. J., and Maley, C. C. Cancer as an evolutionary and ecological process. *Nat. Rev. Cancer*, **6,** 924–935 (2006).

[39] Wang, J., et al. Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, **48,** 768–776 (2016).

[40] Desai, P., et al. Somatic mutations precede acute myeloid leukemia years before diagnosis. *Nat. Med.*, **24,** 1015–1023 (2018).

[41] Abelson, S., et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature*, **559,** 400–404 (2018).

[42] Biswas, D., et al. A clonal expression biomarker associates with lung cancer mortality. *Nat. Med.*, **25,** 1540–1548 (2019).

[43] Losic, B., et al. Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Comm.*, **11**:291– (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[44] Skead, K., et al. Interacting evolutionary pressures drive mutation dynamics and health outcomes in aging blood. *Nat. Comm.*, **12,** 4921– (2021).

[45] Pirzkall, A., et al. Tumor regrowth between surgery and initiation of adjuvant therapy in patients with newly diagnosed glioblastoma. *Neuro. Oncol.*, **11,** 842–852 (2009).

[46] Okolie, O., et al. Reactive astrocytes potentiate tumor aggressiveness in a murine glioma resection and recurrence model. *Neuro. Oncol.*, **18,** 1622–1633 (2016).

[47] Monteiro, A. R., Hill, R., Pilkington, G. J., and Madureira, P. A. The role of hypoxia in glioblastoma invasion. *Cells*, **6,** 45 (2017).

[48] Sabelström, H., et al. High density is a property of slow-cycling and treatment-resistant human glioblastoma cells. *Exp. Cell Res.* **378,** 76–86 (2019).

[49] Scott, J. N., et al. Which glioblastoma multiforme patient will become a long-term survivor? a population-based study. *Ann. Neurol.*, **46,** 183–188 (1999).

[50] Bozic, I., Gerold, J. M., and Nowak, M. A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLoS Comput. Biol.*, **12,** e1004731 (2013).

[51] Ohtsuki, H., and H, I. Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. *Theor. Popul. Biol.*, **117,** 43–50 (2017).

[52] Chen, Y., Tong, D., and Wu, C.-I. A new formulation of random genetic drift and its application to the evolution of cell populations. *Mol. Biol. Evol.*, **34,** 2057–2064 (2017).

[53] Peischl, S., Kirkpatrick, M., and Excoffier, L. Expansion load and the evolutionary dynamics of a species range. *Am. Nat.*, **185,** E81–E93 (2015).

[54] Rubanova, Y., et al. Reconstructing evolutionary trajectories of mutation signature activities in cancer using tracksig. *Nat. Comm.*, **11,** 731– (2020).

[55] DeWitt, W. S., Harris, K. D., Ragsdale, A. P., and Harris, K. Nonparametric coalescent inference of mutation spectrum history and demography. *PNAS*, **118,** e2013798118 (2021).

[56] Kondo, A., Safaei, R., Mishima, M., Niedner, H., Lin, X., and Howell, S. B. Hypoxia-induced enrichment and mutagenesis of cells that have lost dna mismatch repair. *Cancer Res*. **61,** 7603, (2001).

[57] R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.10.08.463633; this version posted October 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[58] Van Rossum, G., and Drake Jr, F. L. *Python reference manual* (Centrum voor Wiskunde en Informatica Amsterdam, 1995).