Extensive subclonal mutational diversity in human colorectal cancer and its significance

Lawrence A. Loeba,b,1, Brendan F. Kohrna,2, Kaitlyn J. Loubet-Seneara,2, Yasmin J. Dunnb, Eun Hyun Ahnb, Jacintha N. O’Sullivanc, Jesse J. Salkd,e, Mary P. Bronnerf, and Robert A. Beckmanb,h,i

aDepartment of Pathology, University of Washington, Seattle, WA 98195; bDepartment of Biochemistry, University of Washington, Seattle, WA 98195; cTrinity Translational Medicine Institute, Department of Surgery, Trinity College Dublin, St. James’s Hospital, Dublin 8, Ireland; dDivision of Medical Oncology, University of Washington, Seattle, WA 98195; eTwinStrand Biosciences, Inc., Seattle, WA 98121; fDepartment of Pathology, University of Utah, Salt Lake City, UT 84112; gDepartment of Oncology, Georgetown University Medical Center, Washington, DC 20007; hDepartment of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University Medical Center, Washington, DC 20007; iLombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20007

Edited by Philip C. Hanawalt, Stanford University, Stanford, CA, and approved November 3, 2019 (received for review June 14, 2019)

Human colorectal cancers (CRCs) contain both clonal and subclonal mutations. Clonal driver mutations are positively selected, present in most cells, and drive malignant progression. Subclonal mutations are randomly dispersed throughout the genome, providing a vast reservoir of mutant cells that can expand, repopulate the tumor, and result in the rapid emergence of resistance, as well as being a major contributor to tumor heterogeneity. Here, we apply duplex sequencing (DS) methodology to quantify subclonal mutations in CRC tumor with unprecedented depth (10^4) and accuracy (10^-7). We measured mutation frequencies in genes encoding replicative DNA polymerases and in genes frequently mutated in CRC, and found an unexpectedly high effective mutation rate, 7.1 × 10^-7. The curve of subclonal mutation accumulation as a function of sequencing depth, using DNA obtained from 5 different tumors, is in accord with a neutral model of tumor evolution. We present a theoretical approach to model neutral evolution independent of the infinite-sites assumption (which states that a particular mutation arises only in one tumor cell at any given time). Our analysis indicates that the infinite-sites assumption is not applicable once the number of tumor cells exceeds the reciprocal of the mutation rate, a circumstance relevant to even the smallest clinically diagnosable tumor. Our methods allow accurate estimation of the total mutation burden in clinical cancers. Our results indicate that no DNA locus is wild type in every malignant cell within a tumor at the time of diagnosis (probability of all cells being wild type, 10^-109).

Quantification of subclonal mutations at very high depth. Both strands of single DNA molecules are sequenced, and mutations are defined as those that are present in both strands of the same molecule at the same position and are complementary (6). We conducted ultradeep sequencing of 11 microsatellite instability (MSI)-negative CRCs (T) and adjacent normal (N) tissues. We assembled 2 gene libraries, a 10-kb library encoding replicative DNA polymerase delta and epsilon active sites (5 T/N pairs), and a 13-kb library encoding genes frequently mutated in CRC (11 T/N pairs). No clonal mutations were detected in the evolutionarily conserved DNA polymerase genes, but clonal mutations were plentiful within the second library in tumors as expected based on mathematical modeling | tumor evolution | drug resistance | duplex sequencing | genetic instability

A million subclones | personalized cancer treatment

Acknowledgments and Notes

This article is a PNAS Direct Submission. This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Author contributions: L.A.L. and R.A.B. designed research; K.J.J.-S., Y.J.D., E.H.A., J.N.O., and M.P.B. performed research; B.F.K., E.H.A., J.J.S., and R.A.B. contributed new reagents/analytic tools; E.H.A. contributed GBM data for comparison; M.P.B. collected samples; B.F.K. and E.H.A. analyzed data; and L.A.L. and R.A.B. wrote the paper.

Competing interest statement: L.A.L. and the University of Washington have a license agreement with TwinStrand Biosciences for the use and development of Duplex Sequencing technology. L.A.L. and J.J.S. are founding members of TwinStrand Biosciences. L.A.L. is a member of the Scientific Advisory Board of Stratos Genomics, Inc. J.J.S. is the Chief Scientific Officer of TwinStrand Biosciences. R.A.B. consults for AstraZeneca, EMD Serono, Vertex, Zymeworks, and CStone, and is the Founder and Chief Scientific Officer of OncoMind, LLC.

Cancers evolve many mutations. Clonal driver mutations are selected early. Subsequent evolution occurs in a branching fashion, possibly without selection (“neutral evolution”). Rarer mutations occur later on smaller branches of the evolutionary tree. Using a DNA-sequencing method, duplex sequencing, with unprecedented accuracy and sensitivity, we quantified rare unique subclonal mutations in diagnostic specimens from 5 human colorectal cancers. Rarer subclones probe later evolutionary time points than previously possible. We confirm neutral evolution at later times and find many more subclonal mutations than expected. A theoretical method allowed us to extrapolate further forward in time to diagnosis. At diagnosis, every DNA base is mutated in at least one cancer cell. In particular, any therapy resistance mutation would be present.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1910301116/-/DCSupplemental.
The Cancer Genome Atlas (TCGA) data (TP53, 5/11; KRAS, 5/11; BRAF, 1/11; PIK3CA, 5/11; UMPs, 1/11) (7).

Results

The mean subclonal mutation frequency was 8.4 \times 10^{-7} \pm 6 \times 10^{-8} per nucleotide sequenced (n = 11) in the tumors and 9.2 \times 10^{-7} \pm 1 \times 10^{-7} in the “normal” colon. The mutational burden in adjacent normal tissue is not statistically different from the uncorrected mutational burden in the tumor (Fig. 1, Top), but when the tumor mutational burden is corrected for the percentage of normal tissue present, estimated with digital imaging analysis (SI Appendix), tumors exhibit a 1.9-fold increase in the mutational burden compared to normal tissue, which is statistically significant (Fig. 1, Bottom). The presence of numerous subclonal mutations in normal colon expands on the results of Martincorena et al. (8, 9), who observed clonal mutations throughout normal esophageal tissue. There are several possible explanations for this phenomenon. First, it is difficult to compare the mutation frequencies in normal colonic tissue and relate them to relative mutation rates without knowing the corresponding proliferation histories. In normal colon, stem cells divide unequally, one daughter replacing the parental stem cell while the other differentiates into the intestinal lumen; in tumors, each cell undergoes symmetric divisions. Furthermore, it has been estimated that one-half of the mutational burden in a tumor was acquired in cells before birth of the founder due to the large number of cell divisions over many years before the founder cell is formed, which could also explain the observed results (10). Finally, with age, more and more “normal” cells may simply acquire a mutator mutation but still have an incomplete set of oncogenic driver mutations, whereas malignant cells may have a complete set (11).

We measured the mutational spectrum of the 96 possible triplets (consisting of the mutation and 3′ and 5′ bases flanking each substitution) in CRCs, adjacent normal colon, and glioblastomas (CRC data in SI Appendix, Fig. S2). CRCs cluster closely in triplet space, and the distribution does not differ statistically from normal colon. Cosine analysis indicates that the landscape of triplets (12) is different in the highly conserved polymerase sequences in glioblastomas than in CRCs (Fig. 2 and SI Appendix) in accord with Hoang et al. (13), who found mutational spectra similar between T and N within a single tumor but different between different tumor types.

The contribution of selection in tumor evolution is still debated. It is generally agreed that oncogenic driver mutations conferring critical cancer phenotypes are often clonal and positively selected. Some models assume successive purifying selective sweeps associated with acquisition of additional drivers (14), while others (11, 15) assume neutral evolution after malignant transformation, where neutral evolution refers to the idea that shortly after cellular transformation the malignant cell has acquired an assortment of driver mutations, and mutations thereafter do not confer a further fitness advantage or disadvantage in the absence of therapy. Thus, these further mutations are randomly acquired, neither enriched nor purified away during subsequent tumor evolution prior to therapy. The latter neutral evolution models featured an early mutator mutation (11, 15, 16) increasing the mutation rate and accelerating the acquisition of driver mutations. In a subsequent “Big Bang model” (17), selective sweeps also need not be invoked. Several experimental studies at varying depths supported neutral evolution (13, 17, 18). Our work sequences more deeply and omits clonal mutations, both driver and passenger, in the analysis. In this work, we have operationally defined the subclonal space as \leq 10% allelic frequency; it is these mutations that are evaluated for neutral evolution.

**Determination of Mutation Rate and Mutation Burden via Sequencing the Same Sample at Different Duplex Depths.** As it is not possible to sequence every genome present in a tumor, rare mutations (i.e., mutations present in one cell or a small number of cells) are infrequently sampled. Due to the branching nature of evolution, the earliest mutational events near the trunk of the evolutionary tree are scored in the majority of cells and can be detected at low DS depth. As we increase DS depth, additional recent mutations that are present in a smaller fraction of the cells are also detected. As the number of nucleotides sequenced at a given genomic position is nearly always less than the number of cells in the tumor, we are unlikely to detect evidence of recent mutational events late in the tumor’s evolution. Thus, the full mutation burden in the tumor cannot be directly determined, and the estimate of mutation rates is based on an incomplete dataset.

We developed a method to estimate the mutation rate and the full mutation burden in the tumor by comparing measurements at several different DS depths. Herein and in SI Appendix, we present the theoretical analysis for mutation rate estimation, and then evaluate 5 colorectal tumors by DS at a depth of up to 20,000× and an accuracy of <10^{-7}.

We define the following terms and concepts (SI Appendix, Table S2): 1) Number of effective cell divisions: \( N_e \) the net number of times a single cell in the tumor produces a new living daughter cell to become 2 living cells, both of which survive. This is defined only for growing tissues like tumors. In nonmalignant tissues, the total number of cells remains constant due to a balance of cell birth and death rates (\( b \) and \( d \), respectively). In a tumor, \( b > d \), and therefore the tumor grows over time. However, as \( d \neq 0 \), it takes more than one actual cell division to make an effective cell division. That is, following any actual cell division, the daughter or parent cell may die. In such a case, that cell division does not count as an effective cell division, which refers only to cell divisions that result in an increase in cell number by 1. This simple concept is convenient for analysis because, for any tumor of \( N \) cells that grew from a single founder cell, the number of effective cell divisions is, by definition, always \( N - 1 \), whereas the actual number of cell divisions is greater than or equal to the number of effective cell divisions, and cannot be determined due
to the unknown history of the tumor. It is also important to distinguish cell divisions, which we have equated to the net number of all cells that have divided, from cell generations or doublings, each of which involves a very large number of individual cell divisions, the more so later in the tumor’s history. 2) Relationship between number of actual cell divisions \( N_E \) and effective cell divisions \( N_A \): The rate of increase in a population due to a single cell (not including further proliferation from its daughters) is simply \( b \) in the absence of cell death, whereas the net rate of increase in a population in the presence of cell death is \( b - d \). The ratio of these 2 quantities determines the relationship between actual and effective cell divisions:

\[
N_A = \int_0^T \frac{b(t)N_E}{b(t) - d(t)} dt. \tag{1}
\]

We have expressed this as a time integral since \( b \) and \( d \) may vary over time. This relationship does not apply for adult tissues without net growth, and would be mathematically undefined when \( b = d \).

Average values of \( b \) and \( d \) have been estimated using histologic techniques as the ki-67 index and the caspase 3 index, respectively. These rates are subject to both spatial and temporal heterogeneity that characterize tumor evolution, and, therefore, precise inference of the actual number of cell divisions has been difficult to obtain.

Estimates of \( b \) for CRC have also been provided using labeled nucleotides, leading to an overall estimate of 0.25/6 on average (19). Comparing this estimate to the actual rate of increase of resistant subclones in CRC by noninvasive monitoring, an estimate for \( d \) of 0.18/6 on average is obtained (3). For example, using Eq. 1, if \( b \) and \( d \) have the above values, the number of actual cell divisions exceeds the number of effective cell divisions by a factor of 3.57. Note: The number of actual cell divisions is not utilized in the model described in this paper. 1) DS depth: defined as number of duplex molecules sequenced at each nucleotide position, symbolized by \( D \). 2) Number of cells in the tumor at the time of sampling: \( N \). 3) Number of cells in the tumor at an earlier timepoint \( t \): \( n(t) < N \). 4) Mutation rate per nucleotide per effective cell division: \( k_{\text{mut-eff}} \). 5) Mutation rate per nucleotide per actual cell division: \( k_{\text{mut-actual}} \). Note: This parameter is not utilized in the model because it is not directly measurable. 6) Fraction of apparently unmutated single-base loci at each position: \( F_{\text{apparent-unmutated}} \). The fraction of single-base loci sequenced for which there is no unique subclonal mutation detected when sequencing a tumor of \( N \) cells at depth \( D \). 7) Reference sequence for defining unique subclones: The consensus clonal sequence of the normal in that individual is the reference against which unique subclonal mutations are defined. Compared to the host germline, the founder cell(s) harbors: 1) clonal mutations that were selected during carcinogenesis, and 2) random drift from the germline, which occurred during the life span of the individual, during which the entire colonic epithelium was repopulated on a weekly basis, leading to a variation among normal colonic cells (10). By discarding clonal passenger and driver mutations occurring in the tumor at greater than 10% allele frequency in the analysis, the reference becomes in effect the founder cells, which may be new founder cells or “founder cells” after subsequent selective sweeps. Our analysis is independent of the history of the tumor prior to the birth of the founder cell(s).

The goal of the analysis is to infer the mutational diversity of a tumor from sequencing at a variety of depths, and to discuss the consequences for therapy. The analysis does not use the traditional input and output parameters of actual number of cell divisions \( N_E \) and mutation rate per base per actual cell division \( k_{\text{mut-actual}} \), which can only be inferred based on assumptions that are highly dependent on the unknown tumor history. Rather, it uses the input parameter of the number of effective cell divisions \( N_A \), which can be directly obtained from the tumor size and the corresponding output parameter of mutation rate per effective cell division \( k_{\text{mut-eff}} \). Bozic and Nowak (20) have previously used tumor size as an approximation for the number of cell divisions, as we have in this work. The other input parameters, DS depth \( D \), and fraction of apparent unmutated single-base loci (relative to a tumor consensus reference) are also experimentally obtained.

We can use this analysis to characterize the mutational diversity of a tumor, including the likelihood of the presence of a mutation at an arbitrarily selected site in one or more cells of a tumor, and discuss the clinical consequences. A mathematical model for optimizing targeted therapy while accounting for tumor evolution, as parameterized by the net birth rate (birth rate minus death rate) and the effective mutation rate has previously been published. Simulations using this model demonstrate the utility of this approach, which is similar to the approach described herein (21). However, the model cannot provide a definitive estimate of the tumor’s mutation rate per actual cell division \( k_{\text{mut-actual}} \) nor can it compare such a value with a comparable parameter from normal tissue to evaluate the validity of the mutator hypothesis (which states that tumors have a greater mutation rate than normal tissue). Resolution of this question requires determination of actual mutation rates, which in turn requires knowledge of the mitotic history of both the tumor and...
normal tissue. An illustration of the dependence of the inferred actual mutation frequency on the growth pattern is provided in SI Appendix. Furthermore, Eq. 1 is not designed to be informative for homeostatic tissues, where \( b = d \).

**Mathematical Approach.** In order to model the fraction of apparently unmutated single-base loci, we integrate over the entire history of the tumor, determine the fraction of apparently unmutated single-base loci for daughter cells born at different times, and obtain the average fraction of apparently unmutated single-base loci, weighted by the number of daughter cells born at different times. For mutations detected at a given DS depth, the fraction of apparently unmutated single-base loci is constant [independent of \( n(t) \)], making the average simple to calculate. At early time points, when there are few cells, it is less likely that a mutation will arise because there are fewer cells dividing at that time. However, if a mutation arises at this early time, it will be present in a larger fraction of the cells in the final tumor, because it is closer to the trunk of the evolutionary tree, and therefore will be more likely to be detectable. These two factors (the lower likelihood of a mutation at earlier times but greater likelihood of detecting a mutation that does occur at an earlier time) exactly counterbalance each other to give a constant number of expected detectable mutations arising at any time. These considerations lead to the following equation, which was used in our primary data analysis:

\[
\ln(F_{\text{apparent-unmutated}}) = -k_{\text{mut-eff}}D.
\]  

The derivation of this equation is given in SI Appendix. Plotting the natural logarithm of \( F_{\text{apparent-unmutated}} \) vs. \( D \) is thus expected to lead to a straight line, with slope of \(-k_{\text{mut-eff}}\), which parallels our data very accurately (SI Appendix, Fig. S1 A–E). The expression differs from related methods (22) due to the absence of the infinite-sites assumption (23). In SI Appendix, we show that, at sequencing depths below \( 1/k_{\text{mut-eff}} \), Eq. 2 predicts a linear relationship between number of unique subclonal mutations and sequencing depth, similar to other methods (22). Fig. 3 is plotted in this manner due to the intuitive clarity of presentation. However, at very high depth, this simple linear relationship breaks down and greatly underestimates the total mutational burden of a tumor (see Figs. 5–7). These points are shown mathematically in SI Appendix.

If we could sequence every single cell in the tumor, we could determine the total mutational burden of the tumor. In Eq. 2, we would substitute \( D = N \). Raising \( e \) to the power of each side of Eq. 2, with \( D = N \), we obtain the following:

\[
\text{Fraction of single base loci unmutated in the entire tumor} = e^{-k_{\text{mut-eff}}N}.
\]  

If there are \( R \) single-base loci in the genome, mutation of which can lead to drug resistance to a single drug, and there are \( K \) non-cross-resistant drugs administered, we modify Eq. 3 to predict the probability of no cross-resistance to all \( K \) therapies, due only to mutation:

\[
P_{\text{no simultaneous cross resistance}} = e^{-\left( R k_{\text{mut-eff}} \right)^K}.
\]  

This equation was used to calculate the values in Table 1. A detailed derivation and more results of these calculations are given in SI Appendix.

In SI Appendix, the following additional topics are considered: application of the method to the 5 individual CRC tumors sequenced at multiple depths, the relationship of this theoretical work to earlier work with particular emphasis on the mutant allele fraction as a function of sequencing depth, discussion of assumptions and approximations of the model, illustration of the dependence of actual mutation frequency on the growth pattern (e.g., Gompertzian growth), evaluation of the likely consequences for the estimate of tumor mutational burden and for the mutator hypothesis, and simulations defining the sensitivity of the method for detecting deviations from neutral evolution.

Because of the branching nature of tumor evolution, increasingly rare private mutations are found in later, more numerous branches. Thus, our ability to sequence more deeply with DS gives us a window into evolution further forward in time from the birth of the founder cell. In order to determine whether neutral evolution continues at a constant high mutation rate at later time points, we sequenced multiple independent samples from 5 tumors to progressively increasing depths up to 20,000, using capture probes consisting of oligonucleotides complementary to exons of replicative DNA polymerases. Our mathematical analysis is similar but not identical to the approach Williams et al. (22) devised to analyze the TCGA database. Both methods support the neutral model, in which most mutations do not affect fitness, and a linear relationship is predicted between the number of unique subclonal mutations and the sequencing depth up to depths approaching the reciprocal of the effective mutation
cell with that selected mutation will be more fit by a factor of 1 + s, increasing in number by this factor relative to a cell with fitness 1 each effective cell division. For each set of sensitivity parameters, we compared the average absolute fractional residual error (i.e., absolute value of the experimental data minus that predicted by the model, all divided by the experimental data). If a model had an average absolute fractional residual error above the upper 95% confidence interval of the average absolute fractional residual error of a competing model, it was considered ruled out. Fig. 4 shows that while there was no statistically significant difference between weak selection and neutral models, selection models with $s \geq 0.23$ are ruled out. Weak selection models that fit the data must be paired with lower mutation rates. Despite having more parameters for fitting, they generally required $y$ intercepts further from the theoretical value of zero than their neutral counterparts. In addition to not being able to rule out weak selection, we cannot rule out selection occurring at $\leq 1\%$ of base loci. Weak or infrequent selection approaches neutral evolution in that the term neutral evolution is not meant to be absolute. The results suggest that neutral evolution at a high mutation rate (or weak selection at lower mutation rates if present) continues as far forward in time as we can see, i.e., until the tumor reaches 20,000 cells.

The slope of the lines plotted for individual tumors determine a mutation rate per template nucleotide locus per effective cell division, or “effective mutation rate” (Methods and SI Appendix), in a manner similar but not identical to that of Williams et al. (22). An effective cell division is defined by the addition of one new cell to the tumor, and consists of multiple actual cell divisions depending upon the differences between cell birth and death rates. Thus the “actual mutation rate” will in general be lower than the effective mutation rate. The observed effective mutation rate may change without a change in the actual mutation rate if the balance between birth and death rates changes during tumor growth, but our conclusions below remain robust for different growth patterns (SI Appendix). The number of effective cell divisions is based on tumor size, in that an effective cell division adds one net new cell to the tumor. The actual mutation rate cannot be determined without knowledge of cellular birth and death rates throughout the tumor’s history. Effective mutation rates can be used to estimate total mutational burden of a tumor (below) and to govern evolutionarily optimized therapy (21).

![Fig. 4. Sensitivity threshold for ruling out selection models. Average absolute fractional residual curves for best fit neutral (red) and selection (blue) models are plotted (solid lines) along with their 95% confidence limits (dashed lines) as a function of the strength of selection $s$. Reference data were 1 of the 5 fresh frozen CRC tumors that we sequenced at multiple depths, plotted according to the approach in SI Appendix, Methods, with the natural logarithm of the unmutated fraction of the capture set plotted against depth. For the selection model, $a_s$ was set at 0.015, indicating 1.5% of loci were positively selected (97% were neutral and 1.5% negatively selected).](https://www.pnas.org/doi/10.1073/pnas.1902630116)
Among the 5 tumors sequenced at different depths, the average effective mutation rate is $7.1 \times 10^{-5}$. This is substantially higher than previously estimated for actual mutation rates in normal tissues based on human population genetics and/or tissue culture studies ($\sim 10^{-10}$) (24). However, one cannot compare an effective mutation rate to an actual one without knowledge of the proliferation history, and thus cannot definitively state whether this is indicative of a mutator phenotype.

Given a genome length of $3.1 \times 10^9$, multiplying by the effective mutation rate per nucleotide, we estimate 2,200 new mutations per new daughter cell added to the tumor, or for 32 doublings between the founder cell and a recently born cell when the tumor is large enough to be detected radiologically, a genetic difference of $\sim 65,000$ subclonal mutations. In interpreting this very high number, we note that the vast majority of these mutations would be private mutations detectable only by single-cell or single-molecule sequencing, including mutations in noncoding DNA.

**Discussion**

Quantification of intratumoral genetic diversity and its relationship to therapy resistance began with the landmark 1965 use of combination therapy to prevent emergence of resistance in childhood leukemia (25). Numerous authors have concluded that each tumor cell is genetically distinct (18, 22, 26–28). Both Loeb et al. (29) and Sottoriva et al. (17) pointed out that a high mutation rate facilitates drug resistance, and Sottoriva et al. (17) made a similar point for neutral evolution. Preexisting mutations have been linked to resistance in preclinical experimental models (30), in single-cell analysis of subclones expanded in 3D culture (28), and in clinical cases (31). Thus, preexisting resistance to single agent therapy is likely in many cases.

We estimate the total burden of unique subclones in a tumor from an analysis of the unique subclones detectable at different depths by extrapolation of our data to single-cell depth in tumors of different sizes (Methods and SI Appendix). At very high depth (greater than the reciprocal of the effective mutation rate, $\sim 1.5$ million), the Williams et al. (22) model and related stochastic analyses (26, 32) substantially underestimate the tumor diversity (Figs. 5–7, Methods, and SI Appendix). These approaches utilize the “infinite-sites assumption” (23) that any mutation is unique when first formed. This assumption requires that the number of cells in the dividing population is less than the reciprocal of the effective mutation rate (i.e., 1 million cells dividing with an effective mutation rate of $10^{-10}$). In these cases, there will be no mutations at the majority of sites in the population, and more than one cell with a mutation in the same site is unlikely. We have, however, inferred a higher effective mutation rate than previous authors with the exception of Williams et al. (22). What will be the total tumor diversity if this high mutation rate continues until the tumor reaches its minimum radiologically diagnosable size of $10^6$ cells or even beyond to the terminal phase of the disease with many more cells present in the total cancer? The infinite-sites assumption is no longer applicable due to the large number of individual cell divisions per cell generation, and at each nucleotide locus we expect multiple cells will acquire the same mutation simultaneously (expected number of new cells with the given mutation $= \text{mutation rate} \times \text{nucleotide locus per cell} \times \text{number of cells dividing} > > 1$). Cheek and Antal (33) have recently objected to the infinite-sites assumption on identical grounds. A mathematical approach independent of the infinite-sites assumption, suitable for accurate estimation of tumor mutation burden, is given in SI Appendix, and the predictions are confirmed in each of the 5 tumors sequenced at multiple differing depths (Fig. 3 and SI Appendix, Fig. S1 A–E). We note that stochastic models (26, 32) and deterministic models with (22) and without (this manuscript) the infinite-sites assumption, despite their different mathematical forms, give similar results over a wide range of typical experimental conditions (Figs. 5 and 6; see mathematical proof in SI Appendix).

Stochastic models are more accurate than others at low cell number, and models without the infinite-sites assumption are more accurate for cancers large enough to be clinically diagnosed (Figs. 5 and 6 and SI Appendix). The difference in predicted diversity is highly significant at diagnosis and increases dramatically as the total cancer burden increases (Figs. 6 and 7).

In essence, the cancer gradually approaches and then enters a new quantitative phase when it grows to a total number of cells beyond the reciprocal of the effective mutation rate, acquiring substantially greater diversity. The approach is governed by SI Appendix, Eq. S16 and illustrated in Fig. 6. This equation is derived in SI Appendix, and it is shown that at depths well below the sequencing of all tumor cells (i.e., all experimental studies to date), allele frequency will be inversely proportional to sequencing depth, as claimed by Williams et al. (22). However, unlike the Williams et al. (22) model, this will not continue indefinitely, resulting in the allele frequency approaching zero. Rather, according to SI Appendix, Eq. S16, the allele frequency will smoothly approach a minimum of $k_{\text{mut-eff}}$ (Fig. 6). The underlying molecular biology may not have changed, but based on altered probabilities there are important clinical and biological considerations. Simulation studies of tumor evolution often focus on $10^9$ cells or less due to computational limitations. Based on our results, we do not believe these studies can be simply “scaled up” to larger tumors without taking into account the progressive violation of the infinite-sites assumption. Experimental small animal studies are limited in tumor size due to ethics considerations. Patients, at current levels of diagnostic sensitivity, are diagnosed in the unexplored region of larger tumors, and move further into this region as they move through their clinical course. Under neutral evolution, the result has limited or no dependence on whether the cells are in one or numerous lesions. Subject to the assumptions
that neutral evolution and a high mutation rate continue, we conclude from this analysis that every nucleotide locus is mutated in one or more cells in a clinically detectable tumor. The chance that every cell is “wild type” at any given neutral locus is 10^{-308}. This conclusion differs from the idea that all tumor cells are genetically distinct (which could result from variation at a limited number of sites) or that preexisting resistance is frequent or likely. Our work suggests that preexisting resistance to single-agent therapy is universal and inevitable.

[Note that, in this work, the reference sequence is the normal appearance kinetics does not account for host factors such as acts, in parallel redundant pathways, or in feedback loops, and also can affect protein coding regions, transcription factor binding sites, microRNA, and so on. Often new resistance mechanisms still remain to be discovered. A recent description of clinical resistance mechanisms in BRAF directed melanoma therapy is illustrative (34). The scenarios illustrate that the greater the tumor burden and the greater the number of resistance mechanisms, the more non-cross-resistant therapies would be required to make it likely that no single cell will have simultaneous resistance. Optimal therapy must not only consider preexisting resistance but also the risk of new multiply resistant subclones emerging during tumor growth (21, 35).]

Our work has limitations (see SI Appendix for additional discussion). It assumes that the mutation frequency in the purified gene fragments is representative of the whole genome. Mutational hot spots were not evident within our capture set, which was selected because it is tightly conserved in evolution.

The data do not rule out weak selection with a selection coefficient less than 0.23 (i.e., a 23% net growth advantage per cell generation), nor can it rule out the presence of a very small number of selected sites (<1%) (SI Appendix). This is in accord with sensitivity limits reported by Williams et al. (22).

Weak selection by many “minidriver” genes is an alternative theory of interest to many. Bozic et al. (36) have modeled the selection coefficient based on evaluation of putative passenger and driver mutations in astrocytic glioblastoma and pancreatic adenocarcinoma sequences, and validated their conclusions by predicting the kinetics of appearance and growth of polyps in familial adenomatous polyposis in 2 of 3 datasets. They reached the conclusion that, even for APC, s is very small: 0.004. However, the key equation in their analysis is Eq. 2, in which s appears in a ratio with the mutation rate. They then use a mutation rate 3 orders of magnitude lower than what we have determined. Using our mutation rate, their Eq. 2 would give a value of s of nearly 4, which is in accord with the value of 2.76 more recently measured by direct observation of colonic stem cell crypt evolution using fluorescently labeled cells in genetically engineered mice (37).

Moreover, in our view, the modeling of polyp appearance kinetics does not account for host factors such as
success or failure in establishing the tumor vasculature, or the possible elimination of most nascent polyps by immune surveillance. These and other factors may confound comparisons of tumor initiation kinetics with clinical observations (11). Williams et al. (38) have examined bulk sequencing data from multiple sources, looking for deviations from neutral evolution in the curve of variant allele frequency vs. total mutation burden (as a surrogate for time). They have not detected evidence of a value of \( a \) below 0.2, which they also state as their limit of sensitivity. While we believe neutral evolution is the most straightforward interpretation of our data, weak selection also remains a possibility, and the extent and biological significance of weak selection remain unknown.

Our calculations are based on the use of the zero term of the Poisson distribution given the mean expected number of mutations. However, the scenario we are modeling has similarities to the work of Luria and Delbrück (39) (L–D), who evaluated the incidence of acquired mutation in bacterial cultures grown from a single cell. A mutation that occurs early is considered a "jackpot" in that it will be present in a large number of daughter cells despite initially being only one mutational event. The probability distribution function giving the probabilities of detecting a particular number of mutations in an (L–D) experiment has been, and continues to be, a subject of advanced mathematical research (33, 40, 41). However, the probability of observing zero mutations is agreed to correspond to the zero term of the Poisson distribution in continuous models (39), or its related binomial equivalent in stochastic models (41) if the experimental protocol is free of certain biases that cause the mean to be underestimated in (L–D) experiments. Specifically, in the original (L–D) experiment, jackpot mutations that occur early on in the experiment are rare events due to the low mutation frequency. Unless a very large number of observations are made, the expectation value for these jackpot mutations is a very small real number between 0 and 1. Since the experiment can only read out whole numbers, the most common readout is zero, leading to an underestimate of the average mutation rate (on rare occasions, the readout will be 1 or a higher integer, leading to an overestimate). However, in our experiment, we are observing \( 10^5 \) bases in parallel at a depth of up to \( 2 \times 10^6 \) for a total of \( 2 \times 10^8 \) observations. Luria and Delbrück (39) state that the complex (L–D) distribution applies to a small number of sites. The number of observations must be greater than the reciprocal of the apparent mutation rate to avoid these complexities. Based on the apparent mutation rate we observed, our number of observations exceeds this criterion 70-fold. Luria and Delbrück (39) further point out that a hallmark of the (L–D) distribution is a variance substantially exceeding the mean, and in their paper the ratio of variance to mean ranged from 2 to 3 to over 600, whereas the Poisson has a variance equaling the mean. They report a 5-fold overestimate of the mutation rate under these conditions when estimating based on the zero term of the Poisson distribution. In Table 2, we report that the variance/mean ratio in our experimental data for observed mutations across the 10,000 bases we have sequenced is close to 1 for all our experimental data points. This is not surprising given our large number of observations. In addition, our mutation rate is estimated using multiple time points and the slope between them, rather than a single time point as in the L–D experiment. Since the linear regression slope is influenced by the difference between successive points, any jackpots at early time points may influence the y-intercept more than the slope, further mitigating any artifacts. Finally, we discard any mutations with a 10% or greater allele fraction, minimizing jackpots.

Our major analyses in this paper, including the mutation rate estimation, the estimation of the probability that no cell will be mutated at a given site with respect to the founder cell reference, and the estimation of the probability of resistance to multiple cross-resistant therapies, rely only on the average and the zero term of the Poisson distribution, which should be accurate given the data in Table 2 and other considerations listed above. If the mutation rate were slightly higher than we have estimated, it would only strengthen the major qualitative conclusions of this paper: that diversity of subclonal mutations is extensive and that every site in the genome differs from the founder cell in at least one tumor cell once the tumor is large enough to be clinically diagnosed.

Our model neglects reversions, in contrast to the models of Jukes and Cantor (42) and Kimura (43), both of which use the infinite-sites assumption, and the more recent study by Cheek and Antal (33), which, like ours, does not employ the infinite-sites assumption. The Jukes and Cantor (42) and Kimura (43) models also differ from our model in that they describe an approach to a steady state in a large fixed size population. When steady state is achieved, the reversion of preexisting mutations exactly counterbalances the formation of new mutations. The equilibrium state is approached over millions of generations, on the order of the reciprocal of the mutation rate. Under these conditions, a large number of mutations are preexisting in a large population.

In contrast, in our case, we are interested in the divergence from the reference sequence in a single founder cell. Since the founder cell is the reference, it has no mutations by definition. Thus, we cannot have a reversion at a site in a given cell without first having a mutation. We have considered the Jukes and Cantor (42) and Kimura (43) models, and while the mutation rate observed in our work is high enough to violate the infinite-sites assumption that mutations will not arise at the same base in more than one cell in a large population [requires \( k_{\text{mut}} \geq 1/N(t) \)], it is not high enough for a large number of mutated cells to revert [requires \( k_{\text{mut}}^2 \geq 1/N(t) \) since it would have to be likely that the same base in the same cell mutates and back mutates]. Moreover, a second mutation at the same base has only 1 chance in 3 to revert back to the reference sequence. Looked at another way, a cell that has just had a mutational event at a given site has \(~40\) cell doublings before that single cell has \(~10^{12}\) progeny, the largest total cancer cell burden we consider in the paper. This is much fewer than the millions of generations required to reach equilibrium in the Jukes and Cantor (42) and Kimura (43) models. As the effective mutation rate is calibrated to cell doublings, the expected fraction of the progeny of the mutated cell reverting is at most \(~40\times\) the effective mutation rate or \(~3 \times 10^{-5}\) (if the mutation arises very early in tumorigenesis). This means that only a very small proportion of mutated cells will

### Table 2. Mean number of mutations detected per nucleotide locus and associated variance

| No. | Mean depth | Mean mutations per nucleotide | Variance in mutations per nucleotide | Variance/mean |
|-----|------------|-------------------------------|--------------------------------------|--------------|
| 1   | 1,783.71   | 1.6449E-03                    | 1.6224E-03                           | 0.9984       |
| 2   | 1,134.24   | 1.9355E-03                    | 1.9318E-03                           | 0.9981       |
| 3   | 2,196.85   | 1.8484E-02                    | 3.4788E-02                           | 1.8820       |
| 4   | 2,131.11   | 1.2575E-03                    | 1.2559E-03                           | 0.9987       |
| 5   | 20,897.13  | 1.2091E-02                    | 1.4073E-02                           | 1.1639       |
| 6   | 890.51     | 1.0567E-03                    | 1.2571E-03                           | 1.1808       |
| 7   | 11,902.35  | 1.7809E-02                    | 2.8525E-02                           | 1.6018       |
| 8   | 1,285.00   | 1.2581E-03                    | 1.2565E-03                           | 0.9987       |
| 9   | 16,433.63  | 9.5810E-03                    | 1.6070E-02                           | 1.6737       |
revert. Our system is far from equilibrium, and the correction to our calculations due to reversions is expected to be very small.

We further assume that the effective mutation rate is a population-weighted average of different subclones and the population weighting is stable, an expected consequence of neutral evolution.

Our theoretical treatment does not consider spatial effects. However, our experimental data use pooled DNA from 5 widely separated locations. Sottoriva et al. (17) concluded from multiple measurements that carcinomas are well mixed. Mixing may not affect growth and diversity under neutral evolution.

Drug resistance calculations assume that neutral evolution and a high mutation rate continue throughout a cancer's lifetime. A continued high mutation rate is supported by high effective mutation frequencies in cell lines derived from mature cancers, and by experiments showing mutators are stably selected in mixed bacterial populations (44). Genetic studies find that only a small minority of mutations are selected, consistent with neutral evolution throughout (SI Appendix).

Drug resistance calculations are also limited to point mutations and do not include other genetic mechanisms and non-genetic resistance mechanisms, nor do they consider that hyperploidy and copy number variation may provide additional sites for mutation. These additional mechanisms further strengthen our conclusion about the inevitability of preexisting resistance to single-agent therapy, while potentially increasing the number of non-cross-resistant agents necessary to reliably eliminate a cancer compared to our estimate.

In summary, we have utilized DS to explore rare subclonal mutations at very high depth and accuracy in fresh frozen MSI-negative CRC and neighboring normal tissue, allowing us to see further in time from tumor initiation than previously and revealing profound subclonal diversity. Our experimental data confirm a high mutation rate and neutral evolution (or weak selection with a lower mutation rate) as far forward in time as we can see. Using a theoretical method not confounded by the infinite-sites assumption, we calculate the total tumor mutational burden assuming the high mutation rate and neutral evolution continue throughout tumor progression. We conclude, subject to this assumption, that preexisting resistance in at least one cell to any single therapeutic agent is inevitable in any colorectal tumor large enough to be detected radiologically.

**Methods**

**Specimens and DNA Isolation.** All participating patients gave informed consent, and all ethical guidelines for tissue acquisition from human participants were followed. Guidelines for study procedures were provided by the University of Utah and the Cleveland Clinic. Following surgery, tumor tissue was snap-frozen and stored at −80 °C. Complete clinical, surgical, and pathological data were available for all cases. Pathology was classified according to the World Health Organization classification (https://www.ucc.org/resource/tmnn) of colorectal tumors. All tumors were grade II. Genomic DNA was extracted using the Qiagen DNAeasy Blood and Tissue Kit (69504). Histologic annotation to confirm diagnosis and cellularity of the extracted tumor was performed by frozen section microscopy of immediately adjacent tissue located within 5 μm. Tumor cellularity was estimated microscopically on the frozen section slides, yielding an mean adenocarcinoma cell percentage of greater than 50%. MSI status was determined with BAT26 size analysis and immunohistochemical staining of MLH1, MSH2, MSH6, and PMS2 (45).

All samples were deidentified prior to use in this study. Deidentified samples were collected under an institutional review board on file with the University of Utah, in compliance with US Code of Federal Regulations, 45 CFR Part 46. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

**DS.** Sequencing library preparation was carried out as previously described with minor modifications (6, 46, 47). See SI Appendix for further details, including capture sets, adaptor sequences, capture probes, data processing, and code availability.

**Subsampling of Tumor Libraries.** A tumor library may be computationally subsampled in which all of the DNA molecules in the library are randomly divided into sublibraries of desired sizes, never reusing the same DNA molecule sample without replacement. A single library may be used to generate points along the linear curves here rather than performing separate experiments. Although we have performed separate experiments in this study, we have also shown for the future that subsampling gives valid results. See SI Appendix, Fig. S1 A–E for details.

**Evaluation of Signatures.** We created signatures for 5 CRC tumors and associated normal samples sequenced using the polymerase delta and epsilon capture set, as well as for 5 glioblastoma multiforme samples sequenced using a capture set of glioblastoma multiforme, by comparing the number of mutations of different types (C>A, C>G, C>T, T>A, T>G, T>C) at each different trinucleotide context (NCN or NTN). Detailed analytic methods are presented in SI Appendix.

**Data Availability.** Data from this paper have been deposited in the NCBI Sequence Read Archive under accession no. SRP135906 (48). GBM data have been deposited in the NCBI database (BioProject PRJNA590549) (49).

**ACKNOWLEDGMENTS.** Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Awards NCI P01-CA077852 and NCI RO1-CA160674 (L.A.L.), and NCI R21-CA220111 (E.H.A.). Initial studies on mutation frequency as a function of sequencing depth were carried out by Edward Fox. We thank Michael Schmitz, Jared Roach, and James Shen, and Jesse Salt for comments, and Tom Walsh and Ming Lee for assistance with sequencing.

1. M. R. Stratton, P. J. Campbell, P. A. Futreal, The cancer genome. *Nature* 458, 719–724 (2009).

2. P. J. Campbell et al., Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13081–13086 (2008).

3. L. A. Diaz, Jr et al., The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* 486, 537–550 (2012).

4. T. J. Ley et al., DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature* 456, 66–72 (2008).

5. J. Salk, M. Schmitt, L. Loeb, Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.* 19, 269–285 (2018).

6. M. W. Schmitt et al., Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14508–14513 (2012).

7. Cancer Genome Atlas Research Network, Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068 (2008).

8. I. Martincorena et al., Somatic mutant clones colonize the human esophagus with age. *Science* 362, 911–917 (2018).

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

10. C. Tomasetti, B. Vogelstein, G. Pargamian, Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1999–2004 (2013).

11. R. A. Beckman, L. A. Loeb, Efficiency of carcinogenesis with and without a mutator mutation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14140–14145 (2006).

12. L. B. Alexandrov et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain, Signatures of mutational processes in human cancer. *Nature* 500, 415–421 (2013). Correction in: *Nature* 502, 258 (2013).

13. M. L. Hoang et al., Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 113, 9846–9851 (2016).

14. P. C. Newell, The clonal evolution of tumor cell populations. *Science* 194, 23–28 (1976).

15. R. A. Beckman, Mutator mutations enhance tumorigenic efficiency across fitness landscapes. *PloS One* 4, e5860 (2009).

16. L. A. Loeb, C. F. Springgate, N. Battula, Errors in DNA replication as a basis of malignant change. *Cancer Res.* 34, 2311–2317 (1974).

17. A. Sottoriva et al., A Big Bang model of human colorectal tumor growth. *Nat. Genet.* 47, 209–216 (2015).

18. S. Ling et al., Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl. Acad. Sci. U.S.A.* 112, E6496-E6605 (2015).

19. S. Jones et al., Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4283–4288 (2008).

20. I. Bozic, M. A. Nowak, Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15964–15968 (2014).

21. R. A. Beckmann, G. S. Schermerhorn, C. H. Yang, Impact of genetic dynamics and single-cell heterogeneity on development of nonstandard personalized medicine strategies for cancer. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14586–14591 (2012).
22. M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, A. Sottoriva, Identification of neutral tumor evolution across cancer types. Nat. Genet. 48, 238–244 (2016).
23. M. Kimura, The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61, 893–903 (1969).
24. R. A. Beckman, L. A. Loeb, Genetic instability in cancer: Theory and experiment. Semin. Cancer Biol. 15, 423–435 (2005).
25. E. Frei, 3rd et al., The effectiveness of combinations of antileukemic agents in inducing and maintaining remission in children with acute leukemia. Blood 26, 642–656 (1965).
26. I. Bozic, J. M. Gerold, M. A. Nowak, Quantifying clonal and subclonal passenger mutations in cancer evolution. PLoS Comput. Biol. 12, e1004731 (2016).
27. N. L. Komarova, D. Wodarz, Combination therapies against chronic myeloid leukemia: Short-term versus long-term strategies. Cancer Res. 69, 4904–4910 (2009).
28. S. F. Roerink et al., Intra-tumour diversification in colorectal cancer at the single-cell level. Nature 556, 457–462 (2018).
29. L. A. Loeb, J. H. Bielas, R. A. Beckman, Cancers exhibit a mutator phenotype: Clinical implications. Cancer Res. 68, 3551–3557 (2008).
30. H. E. Bhang et al., Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. Nat. Med. 21, 440–448 (2015).
31. M. W. Schmitt, L. A. Loeb, J. J. Salk, The influence of subclonal resistance mutations on targeted cancer therapy. Nat. Rev. Clin. Oncol. 13, 335–347 (2016).
32. I. Bozic et al., Evolutionary dynamics of cancer in response to targeted combination therapy. eLife 2, e00747 (2013).
33. D. Cheek, T. Antal, Mutation frequencies in a birth-death branching process. Ann. Appl. Probab. 28, 3922–3947 (2018).
34. E. M. Van Allen et al.; Dermatologic Cooperative Oncology Group of Germany (DeCOG), The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. Cancer Discov. 4, 94–109 (2014).
35. C. H. Yeang, R. A. Beckman, Long range personalized cancer treatment strategies incorporating evolutionary dynamics. Biol. Direct 11, 56 (2016).
36. I. Bozic et al., Accumulation of driver and passenger mutations during tumor progression. Proc. Natl. Acad. Sci. U.S.A. 107, 18545–18550 (2010).
37. L. Vermeulen et al., Defining stem cell dynamics in models of intestinal tumor initiation. Science 342, 995–998 (2013).
38. M. J. Williams et al., Quantification of subclonal selection in cancer from bulk sequencing data. Nat. Genet. 50, 895–903 (2018).
39. S. E. Luria, M. Delbrück, Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28, 491–511 (1943).
40. D. E. Lea, C. A. Coulson, The distribution of the numbers of mutants in bacterial populations. J. Genet. 49, 264–285 (1949).
41. D. A. Kessler, H. Levine, Large population solution of the stochastic Luria-Delbrück evolution model. Proc. Natl. Acad. Sci. U.S.A. 110, 11682–11687 (2013).
42. T. H. Jukes, C. R. Cantor, “Evolution of protein molecules” in Mammalian Protein Metabolism, H. N. Munro, Ed. (Academic Press, New York, 1969), pp. 21–132.
43. M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120 (1980).
44. E. Loh, J. J. Salk, L. A. Loeb, Optimization of DNA polymerase mutation rates during bacterial evolution. Proc. Natl. Acad. Sci. U.S.A. 107, 1154–1159 (2010).
45. E. J. Fox et al., Mutually exclusive promoter hypermethylation patterns of hMLH1 and O6-methylguanine DNA methyltransferase in colorectal cancer. J. Mol. Diagn. 8, 68–75 (2006).
46. S. R. Kennedy et al., Detecting ultralow-frequency mutations by Duplex Sequencing. Nat. Protoc. 9, 2586–2606 (2014).
47. M. W. Schmitt et al., Sequencing small genomic targets with high efficiency and extreme accuracy. Nat. Methods 12, 423–425 (2015).
48. Loeb L et al., Duplex sequencing of human colorectal cancer patients. NCBI Sequencing Reads Archive. https://www.ncbi.nlm.nih.gov/spot/SPR135906. Deposited 5 February 2018.
49. E. H. Ahn et al., Mutations present in five GBM samples were determined using duplex sequencing. NCBI Sequencing Reads Archive. https://www.ncbi.nlm.nih.gov/spot/PRJNA590549. Deposited 19 November 2019.