Extreme value theory as a general framework for understanding mutation frequency distribution in cancer genomes

Natsuki Tokutomi*, Kenta Nakai1,2, Sumio Sugano3,4

1 Department of Computational Biology and Medical Science, Graduate School of Frontier Science, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-0882, Japan
2 Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
3 Medical Research Institute, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyou-ku, Tokyo 113-8510, Japan
4 Future Medicine Education and Research Organization, Chiba University, 1-8-1 Inohana, Chiba, Chiba 260-0856, Japan

* evt.cancer.genome@gmail.com

Abstract

Currently, there is no recognized population genetics framework describing the population dynamics of cancer cells that is applicable to real cancer genome data. By focusing on cancer as a Darwinian evolutionary system, we formulated and analyzed the observed mutation frequency among tumors (MFaT) as a proxy for the hypothesized sequence read frequency and beneficial fitness effect of a cancer driver mutation. Analogous to intestinal crypts, we assumed that sample donor patients are separate culture tanks where proliferating cells follow certain population dynamics described by extreme value theory (EVT). To validate this, we analyzed three large-scale cancer genome datasets, each harboring >10 000 tumor samples and in total involving >177 898 observed mutation sites. We clarified the necessary premises for the application of EVT in the strong selection and weak mutation (SSWM) regime in relation to cancer genome sequences at scale. We also confirmed that the stochastic distribution of MFaT is likely of the Fréchet type, which challenges the well-known Gumbel hypothesis of beneficial fitness effects. Based on statistical data analysis, we demonstrated the potential of EVT as a population genetics framework to understand and explain the stochastic behavior of driver-mutation frequency in cancer genomes as well as its applicability in real cancer genome sequence data.

Introduction

Variant Allele Frequency in Cancer Genomics and Population Genetics of Cancer Cells

The concept of “mutation frequency” in cancer genomics in general uses the variant allele frequency (VAF), which is defined as a fraction of mutant reads to total reads at a given genomic site. This value depends on the sequenced sample and thus is different from the fraction of mutated samples in total samples.
Recent studies have shown that the neutral evolution of the cancer genome results in a power law distribution of tumor bulk-sample VAFs reported by a next-generation sequencer \[1,2\]. However, to date, no sophisticated statistical framework exists for understanding the mutation frequency distribution in the cancer genome that is consistent with population genetics and can be ultimately used to describe the population dynamics of cancer cells.

The Big Bang Model and Preclonal Evolution of Cancer

According to the “Big Bang” model of human colorectal tumor growth, multiple tumor subclones are formed with a single expansion of cell population at an early stage of tumor growth \[3\]. In this model, these cells not only acquire clonal lesions that are shared among many cells within the tumor but also acquire subclonal lesions that are observed only in a fraction of those cells at almost the same time. These subclonal lesions are not subjected to stringent selection in cancer evolution (i.e., neutral evolution), thus leading to a state of intra-tumor heterogeneity. However, it is known that cancer as a whole is a Darwinian evolutionary system, in which driver mutations undergo selective evolution and passenger mutations undergo neutral evolution \[4\]. Preclonal evolution before the expansion of the cell population is thought to be selective, with cancer cells acquiring driver mutations that extremely increase cellular fitness \[5\].

Cancer Driver Mutations in the Strong-Selection-Weak-Mutation Regime

All cancer driver mutations are beneficial for the survival of cancer cells, and acquisition of these mutations brings fitness gain to cancer cells. Among population genetics theories that are able to describe the behavior of the fitness effects of beneficial mutations, the strong-selection-weak-mutation (SSWM) regime produces a first-approximation fitness landscape of beneficial mutations \[6\]. This framework assumes that all beneficial mutations that occur in the population will increase fitness monotonically: i.e., they are on “selectively accessible” paths. Thus, their adaptive walk is typically short.

SSWM is characterized by two key assumptions: strong selection and weak mutation \[7\]. Here, “strong selection” is equivalent to excluding neutral mutations from consideration. Because all cancer driver mutations are beneficial, the strong selection assumption holds for these mutations. In contrast, “weak mutation” means excluding cases in which multiple mutations exist simultaneously within the population. In this assumption, we consider only a genotype that is made by introducing a single mutation to a single wild-type genome. We model population adaptation as a repetition of this step, in which a genotype that is selected from such genotypes becomes a new wild type in the population. Although this assumption always holds true for a microscopic landscape with a small population size, it will be violated if multiple clones interact and multiple mutations compete against each other during their fixation process (clonal interference). In the preclonal evolution of cancer cells, the weak mutation assumption holds because only a single clone in the population exists.

An expansion of the cancer cell population size generates multiple clones with multiple different beneficial mutations, thus causing clonal evolution of the newer cells. Here, the weak mutation assumption is thought to be violated and the population dynamics of the cells will shift from SSWM to greedy adaptation \[8\].
wherein mutations with greater fitness effects are certainly fixed in the population, resulting in increased repeatability of the fitness trajectory of cells in such a situation, has been confirmed by an *Escherichia coli* experiment [9]. The fact that a limited number of driver mutations has been observed repeatedly in multiple cancer samples suggests not only the repeatability of the oncogenic process, but also the underlying evolutionary structure itself.

An increase in population size not only strengthens the deterministic traits of the population by clonal interference but also enables its fitness valley crossing [6][10]. This gives rise to an escape genotype that is not on the selectively accessible paths in SSWM, thereby countering the deterministic traits of the population. Stochastic tunneling, a critical element in the valley-crossing mechanism, enables fixation of deleterious mutations as well as neutral evolution [11]. These theoretical aspects of cell population behavior have been validated by *E. coli* experiments [12]. In cancer, such population dynamics corresponds to the subclonal evolution of passenger mutations, leading to intra-tumor heterogeneity.

Among the evolutionary models mentioned above, this manuscript focuses on SSWM to explain the acquisition of driver mutations by cancer cells. According to the mutational landscape by Gillespie, the fitness of a wild type allele at a given gene locus tends to lie on the right side of a distribution because this fitness is usually high, and the fitness of beneficial mutations lie further to the right [7][13]. In this setting, the fitness values are extreme, and thus their statistical behavior is described by extreme value theory (EVT) [14].

The first application of EVT to cancer preclonal evolution examined the proliferation of stem cells in intestinal crypts [15]. Within the internal space of an individual crypt, the adaptation and evolution of the cell population will be completely independent of the states of different crypts.

In our analysis, we employ an analogy that compares an intestinal crypt and a tumor sample. Similar to a crypt, a tumor sample has multiple cells that are evolutionary players. VAF, defined by a fraction of sequence reads, is approximately proportional to the mutant cells contained in a single sample. In contrast, we define MFaT here as a frequency of mutant samples within a single dataset. We accept MFaT as a proxy for VAF and fitness effects that are calculated from the donor or sample frequency of a mutation in a large-scale dataset. This formulation is valid only in the case of driver mutations in the Big Bang model [3].

**Materials and methods**

**Datasets**

In total, we used unique doubleton protein-altering single nucleotide variations (SNVs) that numbered 68 973 from 15 285 patients in the ICGC Release 27 dataset [16], 325 244 from 24 355 tumors in the COSMIC Version 85 dataset [17], and 91 312 from 11 089 samples in the Chang dataset [18]. These groups of unique mutations respectively correspond to groups of unique genomic sites: 68 613 in the ICGC dataset, 322 962 in the COSMIC dataset, and 90 885 in the Chang dataset. These mutations were annotated as being either in the “missense variant,” “initiator codon variant,” “stop gained,” or “stop lost” mutation class. In all datasets, the reference genome version is GRCh37, and the annotation is based on Ensembl v75.

In addition, we used the RTCGA mutations dataset ([https://rtcga.github.io/RTCGA/](https://rtcga.github.io/RTCGA/)) to enable a tumor-type-specific analysis over cancer clonal mutations identified in an identical sequencing strategy. This dataset includes 178 701 unique somatic protein-altering SNVs from 2 732 samples and
corresponding 177 898 unique genomic sites. The samples are collected from eight cancer types: BLCA (127), BRCA (949), HNSC (266), LIHC (191), LUAD (225), PRAD (268), SKCM (340), and THCA (366) (sample sizes indicated in parentheses). The sequencing strategy is whole exome sequencing (WXS) performed with two platforms (Illumina GAIIx, Illumina HiSeq) at six centers (broad.mit.edu, genome.wustl.edu, hgsc.bcm.edu, ucsc.edu, mdanderson.org, and bcgsc.ca). All mutations are annotated as being in either the “missense mutation,” “nonsense mutation,” “translation start site,” or “nonstop mutation” class. We defined mutations with VAFs in the range of [0.25, 1.00] as clonal after a pioneering study in the field [1].

We used the following cancer driver-gene definitions in our analysis: the Mutational Driver definition in the IntOgen Cancer Drivers Database (2014.12) dataset [19], the COSMIC Cancer Gene Census [17], and the Tokheim Oncogenes and Tumor Suppressor Genes [20]. Also, we used the following data as cancer driver mutation site definitions: SNVs whose driver activity is “known” in the IntOgen 2016.5 Driver Mutations Catalog Mutation Analysis dataset [21], SNVs in the DoCM database [22], and amino acid substitution information per gene in a recent study (Table S4 in [23]).

**Data Processing**

In the case of the ICGC dataset, we used the Release 27 Summary Simple Somatic Mutation VCF file, for which the reference genome build is GRCh37 and mutation annotations are based on Ensembl version 75. First, we selected protein-altering mutations (annotated as either “missense variant,” “stop gained,” “stop lost,” or “initiator codon variant”) in the VCF file using our custom script (see S1 Script: get_IcgcProteinAltering.py). If a mutation had multiple annotated transcripts, and it had been annotated as a protein-altering mutation in any of those transcripts, we included such a mutation in the analysis because it will satisfy the strong selection assumption in cancer driver genes. Next, we selected doubleton (affected donors > 1) SNV records and then calculated and maximized their MFaTs (donor-based MFaTs). For the ease of the process, we added gene symbols as an independent column that had been a part of the VCF annotations (based on Ensembl 75)(see S1 File: database_ICGC_temp_PostMax.tsv.gz). The fields of the affected donor count (affected donors) and the total donor count (total donors, 15 285) are accepted as the mutated tumor count and the total tumor count in an MFaT calculation, respectively.

In the case of the COSMIC dataset, we used the GRCh37 Version 85 Mutant Export TSV file that includes mutations from cancer cell lines. First, we selected SNVs that were doubletons or more frequent (CNT > 1) from the coding mutation VCF (with redundancy due to annotations) based on mutation IDs. Among those mutant export TSV records that had mutation IDs, we analyzed only those obtained from genome sequencing data (“Genome-wide screen” == “n”) mapped to the GRCh37 genome build (GRCh == “37”). We summed mutation IDs based on tumor IDs and calculated the MFaTs for the respective mutations (tumor-based MFaTs). For these mutation records, we added genomic coordinates using the VCF file, formatted the records appropriately, and then removed redundancies due to annotated gene symbols. Finally, we selected only protein-altering mutations (in practice, those with recorded amino acid substitutions) and maximized the calculated MFaTs (see S2 File: database_COSMIC_temp_PostMax.tsv.gz). The fields of the affected tumor count (affected_tumors) and the total tumor count (total_tumors, 24 355) were accepted as the mutated tumor count and the total tumor count in an MFaT calculation, respectively.

In the case of the CHANG dataset, we used the PanCancer Unfiltered MAF file. Out of all mutations recorded in the file, we selected only records with mutations that were protein-altering (annotated as either “missense variant,” “initiator codon variant,” “stop gained,” or “stop lost”) SNVs and potential doubletons according to summed
sample IDs. We added total sample counts for those records based on sample IDs, calculated MFaTs totaling sample IDs for each mutation (sample-based MFaTs), and then maximized these MFaTs (see S3 File database_CHANG_temp_PostMax.tsv.gz). The fields of the affected sample count (affected samples) and the total sample count (total samples, 11,089) were accepted as the mutated tumor count and the total tumor count in the MFaT calculation, respectively.

We selected driver mutations according to gene symbols in the case of the driver-gene definitions and to genomic coordinates in the case of driver-site definitions. We generated driver-gene definition files for the IntOgen, CGC, and Tokheim datasets, respectively (for details, see Datasets section and S4 File driver_TotalGene_database_IntOgen.tsv, S5 File driver_TotalGene_database_CGC.tsv, S6 File driver_TotalGene_database_Tokheim.tsv). Also, we generated driver-site definition files for the IntOgen, DoCM, and Bailey datasets, respectively (for details, see Datasets section and S7 File driver_TotalSite_database_IntOgen.tsv, S8 File driver_TotalSite_database_DoCM.tsv, S9 File driver_TotalSite_database_Bailey.tsv).

In the case of the RTCGA mutations dataset, we used a reformatted file by selecting the necessary columns after processing the header row (i.e., we renamed a column from “Start position” to “Start Position”) (see database_RTCGA_temp_Format.tsv.gz).

Further, we select protein-altering SNV records whose reference genome build is GRCh37 (NCBI Build == “37”), sequencing strategy is whole exome sequencing (Sequence Source == “WXS”), VAF (Variant Allele Frequency) is greater than 0.25 (VAF > 0.25), and mutation class is either “Missense Mutation,” “Nonsense Mutation,” “Translation Start Site,” or “Nonstop Mutation” (see database_RTCGA_temp_ProteinAlteringSnv.tsv.gz) (for details of the RTCGA dataset at this step, see Datasets section). Finally, we calculated total sample counts (total samples) per tumor type in a reference/alternate-sensitive manner and selected mutation records that are potential doubletons (see S10 File database_RTCGA_temp_ProteinAlteringSnvDoubleton.tsv.gz). In the calculation of MFaTs, the total sample counts (total samples) and the affected sample counts (affected samples) are accepted as total tumor counts and mutated tumor counts.

In the case of RTCGA total-tumor analysis, we summed sample IDs over all tumor types (affected samples) across the RTCGA dataset, calculated MFaTs, and subsequently maximized them (see S11 File database_RTCGA_temp_PostMaxTotal.tsv). The method of extracting the intersection set of mutations about the driver-gene and driver-site definitions is similar to the cases of ICGC and other datasets.

In the case of RTCGA type-specific analysis, we summed sample IDs, calculated MFaTs, and maximized them in a tumor-type-specific manner (affected samples) (see S12 File database_RTCGA_temp_PostMaxType.tsv). We used only driver-gene definitions in extracting the intersection set of mutations (see S13 File driver_TypeGene_database_IntOgen.tsv). In this way, we initially generated a list of genes within the driver-gene definition, given a certain tumor type, and then selected mutations that occurred within the listed genes (see S14 File drivertype_RTCGA-IntOgen_temp_PreRank.tsv).

The Definition and Calculation of MFaT

We defined the MFaT value as a ratio of sequencing samples having a certain mutation to the total observed samples. To practically and approximately calculate the continuous MFaT values, we utilized multiple total-tumor cancer genome datasets. In general, observed MFaT values will be discrete in any dataset with a sample size less than 2,000. The total-tumor datasets, which typically have thousands of sequencing samples, will provide enough resolution in MFaT values to satisfy the continuity assumption. With the aims of ensuring individual mutation observations and reducing...
computational cost, we restricted our analysis with doubleton (i.e., observed at least twice in each dataset) somatic protein-altering SNVs.

To exclude mutations with smaller fitness effects, we restricted our analysis to the mutation sites with MFaT values ranked in the top 200 in the case of total-tumor analysis and top 50 in the case of type-specific analysis. Here, we assigned smaller ranks to greater MFaT values. For tied MFaT observations, we assigned the maximum rank (e.g., for MFaT observations \{0.1, 0.3, 0.3, 0.5\}, ranks \{4, 3, 3, 1\} will be assigned). The number of plotted MFaT values after the filtering is shown in the figures as “b,” and the effective threshold “th” is shown where relevant.

The Maximization of MFaT

In practice for large-scale cancer genome analysis, it is common to observe multiple mutations on a single genomic site. This is due to the phenomenon that a single site is recurrently mutated to differing bases (substitution) in a set of samples. Since we consider only simple substitutions at this time, three alternate sequences of a given genomic site are possible (e.g., \{A, T, C\} against a reference base G). This could result in, at maximum, three different MFaT values for one site (e.g., MFaT for G to A, MFaT for G to T and G to C).

In general, the number of cases of possible cancer cell environments that dominate cancer evolution via selective pressure is unlimited. Thus, the number of possible selective coefficients (i.e., the fitness effect) at a given genomic site is well approximated by positive infinity (the infinite micro-environments assumption). This is equivalent to the case where the number of possible alternate sequences of a given genomic site is infinite.

In addition, we assume that the mutational selection coefficient of a given site is maximized among possible alternate values after evolutionary selection (the maximization assumption).

\[ S_i = \max(S_{i,1}, S_{i,2}, ..., S_{i,n}) \quad (n \to \infty) \]  

where \( S \) denotes a mutational selection coefficient, \( i \) is the index for genomic sites, and \( n \) is the number of possible values, respectively. This is consistent with the selectionist idea of the survival of the fittest, and the block maxima model in the extreme value theory.

To suffice for the maximization assumption of MFaT observations, we selected the maxima of MFaT values in each genomic site, and excluded the rest from our analysis. As a result, the counts of MFaT values, mutations, and sites will all be equal. Here, we define this process as “maximization” of MFaTs, which will enable more exact and reliable data processing in future analyses. In the case of a tumor-type-specific analysis, we performed this maximization process for the respective tumor types.

Parameter Estimation

We estimated the generalized extreme value distribution (GEV) and generalized Pareto distribution (GPD) parameters by the maximum likelihood (Nelder-Mead optimization) method using the “evd” R package (i.e., evd::fgev, evd::fpot functions). Initial values for the optimization are set in the GEV parameter estimation in the total-tumor analysis using the ICGC, COSMIC, and Chang datasets (shape: 1.25, location: 0, scale: 0), while they are not set in bootstrap simulations using these datasets. We also used different initial values (shape: 1.0, location: 0, scale: 0) in the total-tumor analysis using the RTCGA dataset.
For parameter estimation of the Pareto distribution, we defined its probability density function (PDF) and maximum likelihood estimators as the following, referring to implementations in the “VGAM” R package. Here, \( n \) is the length of the observed data vector.

\[
f_{\text{Pareto}}(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}} \quad x > x_m
\]

(3)

\[
\hat{x}_m = \min_i x_i
\]

(4)

\[
\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln x_i - n \ln \hat{x}_m}
\]

(5)

**Goodness-of-Fit Assessment by \( \chi^2 \) Goodness-of-Fit Test**

To assess goodness-of-fit of the GEV and Pareto distributions, we performed a conventional \( \chi^2 \) goodness-of-fit test over the total-tumor and tumor-type-specific cancer driver MFaTs using the “stats” R package (stats::chisq.test function; rescale.p and simulate.p.value flags were both set to TRUE). The theoretical frequencies of the MFaTs were calculated using parameters estimated by the maximum likelihood method (see Parameter Estimation section) and then compared with the observed MFaT frequencies. In this setting, the null hypothesis \( H_0 \) stated, “the observed distribution is identical with the theoretical,” and the alternative hypothesis \( H_1 \) stated, “the observed distribution is different from the theoretical.” Consequently, a higher significance level such as \( p < 0.05 \) rejected the \( H_0 \) and indicated that “the observed distribution is significantly different from the GEV,” for example. Conversely, a lower significance level such as \( p = 0.50 \) weakly supported the notion that the observed and theoretical distributions are alike.

**Bootstrap Simulation and Confidence Interval**

Bootstrap simulations over parameter estimators were performed using the “boot” R package (boot::boot function) in the total-tumor analysis. In all analyses, the iteration number is set to \( n = 1 \times 10^5 \). We did not set initial values in the cases of the ICGC, COSMIC, and Chang datasets, while they were set in the case of the RTCGA dataset (shape: 1.0, location: 0, scale: 0).

We calculated 95 \% confidence intervals for the GEV shape parameter (i.e., tail index) using simulated total-tumor bootstrap distributions and a variety of calculation methods (in boot::boot.ci function): the first-order normal approximation (Normal), the basic bootstrap interval (Basic), the studentized bootstrap interval (Student), the bootstrap percentile interval (Percentile), and the adjusted bootstrap percentile (BCa) interval.

**Density Plot**

For MFaT observations in each total-tumor case, we plotted theoretical probability density functions (PDFs) obtained from the parameter estimation of distributions and empirical PDFs obtained from observed values. We used the “evd” R package (evd::dgev and evd::dgpd functions) to plot GEV and GPD PDFs. Similarly, we defined the PDF of the Pareto distribution by referring to the “VGAM” R package (VGAM::dpareto function) (see Parameter Estimation section). We used the “stats” R package (stats::density function) to draw the empirical density function obtained from observations.
Data Normalization in Q-Q Plot

We calculated a theoretical value of an observed MFaT value \( x_k \) with \( Q(k/(N+d)) \), where \( Q(\cdot) \) is the quantile function of GEV, \( k \) is the rank of the observation when the values are ranked from the smallest, \( N \) is the number of observations (\( N = b \) in plots), and \( d \) is the numerator correction factor (both \( b \) and \( d \) are shown in plots). The quantile function is defined by an inverse function of the cumulative distribution function (CDF) of GEV. We used the “evd” R package (evd::qgev function) to calculate the quantile function values.

Formulation of Fréchet Plot

Based on the Weibull plot introduced in [40], we developed and drew a Fréchet plot (see demonstration in Supplementary Materials in S1 Appendix). After the assessment of the goodness-of-fit of the GEV distribution to the observations, we calculated slope and intercept values in a linear regression. In the cases of tumor-type-specific and gene-specific Fréchet plots, we added an additional data point MFaT = 1.0 to the observations in calculating each value of the empirical distribution function \( F_{\text{emp}}(x) \) against the observed MFaT \( x \). This satisfied the condition \( 0 < F_{\text{emp}}(x) < 1 \) so that the y-axis value \( y(x) = -\ln(-\ln F_{\text{emp}}(x)) \) is finite. In the gene-specific cases, we plotted only genes with more than two unique MFaT values. In such cases, we analyzed skin cutaneous melanoma (SKCM) tumors separately that are more enriched in mutations.

Bayesian Parameter Estimation by Markov Chain Monte Carlo Simulation

In Bayesian extreme value analysis, we estimated tumor-type-specific GEV parameters using the Markov Chain Monte Carlo (MCMC) approach. We used the “evdbayes” R implementation for this purpose. Prior distributions of respective GEV parameters (i.e., shape, scale, and location) were assumed independent and normal. This means that the variance-covariance matrix that is used to calculate prior distribution is diagonal. Also, this distribution is equal to the trivariate (i.e., with three variables) normal distribution with variables that are mutually independent. In addition, we determined the prior parameters regarding the results of the bootstrap simulation, which are dependent on observed data. This means that the prior distribution used in this step is “informative.” Specifically, the normal parameters in the informative priors were set as follows (mean ± standard deviation): shape 1.0 ± 0.5, scale \( 2.5 \times 10^{-4} \pm 3.2 \times 10^{-4} \), and location \( 1.5 \times 10^{-3} \pm 1 \times 10^{-4} \). The standard deviations of proposal distribution in MCMC were set as follows: shape 0.01, scale \( 1 \times 10^{-5} \), and location \( 1 \times 10^{-4} \). The burn-in number of MCMC was set to 5 000, whereas the number of iterations was \( 1 \times 10^5 \). The initial values of the MCMC simulation were set equal to the expectation of the normal priors. However, for scale parameters in tumor-type-specific parameter estimation, we set the next value \( \tau \) as the initial values considering the loss of genetic diversity in sequencing samples due to the shrinking sample size, which has been a problem in tumor-type-specific analyses.

\[ \tau = 1/[\text{Sample size of the tumor type}] \]  

After the MCMC simulation, we estimated tumor-type-specific GEV parameters using an expected à posteriori (EAP) estimator.
Estimation of Mutational Fitness Effects by MFaT

Consider the relative fitness $W_{i,A}$ of a cancer cell with a post-mutational sequence $A$ at a certain genomic site $i$. We define this value using the classic evolutionary selection coefficient, or fitness effect, $S_{i,A}$:

$$W_{i,A} = 1 + S_{i,A}$$ (7)

With this definition, we formulate the selection coefficient whereby the cell’s genome is mutated from a pre-mutation (or reference) sequence $R$ (with length 1 bp) to a post-mutational sequence $A$ [24],

$$S_{i,R \rightarrow A} = \frac{W_{i,A} - W_{i,R}}{W_{i,R}} = \frac{S_{i,A} - S_{i,R}}{1 + S_{i,R}}$$ (8)

Then, we consider cancer cell environments that determine selective pressure throughout cancer evolution. We consider the set of all possible environmental states $\Theta$ and its elements $\theta$ that determine the value of a selection coefficient of a cancer cell with a given genotype. Next, we define the selection coefficient whereby the cell mutates from a pre-mutational sequence $R$ to a post-mutational sequence $A$ at a genomic site $i$ within a given environmental state $\theta$:

$$S_{i,R \rightarrow A,\theta} = \frac{S_{i,A,\theta} - S_{i,R,\theta}}{1 + S_{i,R,\theta}}$$ (9)

In addition, we assume that the fitness effect of a cancer driver mutation after preclonal evolution is maximized (the maximization assumption). In other words, the fitness effect of a given mutation is the maximum of many alternative values that are possible depending on the combinations of alternate sequence information and possible cancer cell environments. Under this formulation, the mutational selection coefficient (MSC), whose values are unique to cancer driver mutation sites, is defined by:

$$S_i = \max(\{S_{i,R \rightarrow d,\theta}\}_{d \in D, \theta \in \Theta})$$ (10)

Here, $D$ is the set of post-mutational DNA sequences that are possible at the site $i$, and $d$ is its single element. $D$ is dependent on the genomic site $i$ and the scope of consideration of mutational classes. For example, in the case considering only simple substitutions, then $D = \{A,T,C\}$ if the pre-mutational sequence $R$ was $G$. In this case, the size (i.e., the number of elements) of $D$ is $\#\{D\} = 3$. In contrast, the set $\Theta$ is independent of genomic sites and pre- and post-mutational sequences. Here, we assume the size of the set $\#\{\Theta\}$ is infinite (i.e., the number of possible micro-environmental states is numerous) (the infinite micro-environments assumption). Then, the number of possible selection coefficients $n = \#\{D\} \cdot \#\{\Theta\}$ for a given site $i$ tends to infinity, and $n \to \infty$ holds.

Let the individual values of $S_{i,R \rightarrow d,\theta}$ at respective genomic sites $i$ be continuous and independent and identically distributed (IID) about $D$ and $\Theta$ (the IID assumption and the continuity assumption). Based on the maximization, IID, and continuity assumptions, the block maxima model in extreme value theory is valid over $S_i$, demonstrating that $S_i$ converges to the generalized extreme value distribution (GEV) as $n \to \infty$. Finally, as this limit holds due to the infinite micro-environments assumption, the probability distribution of $S_i$ will be GEV over the set of genomic sites $J(i \in J)$. 

November 15, 2020 9/24
Here, we assume that MSC at a site \( i \) is proportional to the MFaT of the site \( i \) (the proportionality assumption).

\[ S_i \propto \text{MFaT}_i \]  \hspace{1cm} (11)

We generalize this by considering normalizing constants of MFaTs, \( \sigma_0 \) and \( \mu_0 \).

\[ S_i = \frac{\text{MFaT}_i - \mu_0}{\sigma_0} \]  \hspace{1cm} (12)

In general, fitness \( W \) is unobservable, and so is the relative fitness \( W_{i,A} \) by a mutation at site \( i \) in preclonal cancer cells. Thus, here we heuristically set \( \mu_0 = 0 \) and \( \sigma_0 = 1 \). Then, the estimator of MSC at site \( i \) is given by

\[ \hat{S}_i = \text{MFaT}_i \]  \hspace{1cm} (13)

**EAP Estimation of MSCs by GEV-Binomial Model**

We consider the probability distribution of the sample count \( k \) of tumor samples which have a certain post-mutational sequence at a given genomic site \( i \). Since \( 0 < \text{MFaT} < 1 \), here we consider the binomial distribution \( \text{Binom}(m, S_i) \) with \( S_i \) (the MSC at site \( i \)) as a ratio parameter and \( m \) (the number of tumor samples) as a size parameter. We update the Bayesian knowledge of the parameter \( S_i \) in our Bayesian framework using observations of \( k \) and \( m \). Specifically, we assume GEV as a prior distribution over \( S_i \), and then we calculate the posterior distribution using discrete observations at individual genomic sites.

The prior distribution of \( S_i \) is expressed using the PDF of GEV \( f_{\text{GEV}}(s) \) with three parameters (shape \( \xi \), scale \( \sigma \), and location \( \mu \)) as follows:

\[ f_{\text{GEV}}(s) = \frac{1}{\sigma} A(s)^{-1 - \frac{1}{\xi}} \exp \left[ -A(s)^{-\frac{1}{\xi}} \right] \]  \hspace{1cm} (14)

\[ \xi \neq 0 \quad ; \quad A(s) = 1 + \frac{s - \mu}{\sigma} > 0 \]  \hspace{1cm} (15)

Here, \( A(s) \) is a function of \( s \), the argument of the PDF. The likelihood at each value of tumor sample counts, \( k = 0, 1, 2, \ldots, m \), is expressed using the probability mass function (PMF) of the binomial distribution \( f_{\text{Binom}}(k) \) as follows:

\[ P(X = k|S_i = s) = f_{\text{Binom}}(k) = \binom{m}{k} s^k (1 - s)^{m-k} \]  \hspace{1cm} (16)

From Bayes theorem and appropriate assumptions over PMF and PDF, we assume the following equation over the posterior distribution of \( S_i \)

\[ P(S_i = s|X = k) = \frac{P(X = k|S_i = s)P(S_i = s)}{P(X = k)} \]  \hspace{1cm} (17)

Here, the necessary assumptions are
\[ P(S_i = s) = f_{GEV}(s) \]  

\[ P(X = k) = \sum P(X = k | S_i) \]  

For the stepwise calculation of \( P(X = k) \), we need to consider tumor sample counts at all genomics sites. However, this is easily achieved by normalization of the numerator \( P(X = k | S_i = s)P(S_i = s) \). Finally, we estimate \( S_i \) by the expected à posteriori (EAP) estimation and obtain its 95% confidence interval (EAP ± 1.96 APSD) by calculating the à posteriori standard deviation (APSD).

**Results**

**The Definition of MFaT**

The mutant allele frequency among tumors (MFaT) is the frequency of samples within a considered sequencing dataset that have a given mutation at a given genomic site. The MFaT value is defined at any genomic sites corresponding to individual mutations recorded in the dataset. In this formulation, the count of samples having a mutation is normalized by the number of total samples within the dataset, permitting a comparison of mutated sample frequencies between different datasets.

\[ \text{MFaT} = \frac{\text{#(mutated tumors)}}{\text{#(total tumors)}} \]  

In this study, we calculated the MFaTs of only protein-altering SNVs that had doubleton or higher frequency to ensure observations (for details, see Materials and Methods). A protein-altering mutation in our study is defined as a mutation that changes the protein sequence in either of the annotated transcripts. On mutations that do not satisfy this condition, or non-protein-altering mutations, one can define MFaTs although the proportionality assumption does not hold over the values (for details, see Discussion).

**The Probability Distribution of Driver MFaT Is Likely a Type II Extreme Value Distribution**

A visual inspection of the distribution shape using a density plot (Fig 1A and Fig 2A) suggested that the probability distribution of cancer driver mutation MFaTs is approximately equal to the extreme value distribution (for details, see Materials and Methods). A Q-Q plot with Sugano normalization (see Materials and Methods) confirmed that the driver MFaT distribution is approximately described by the extreme value distribution given the relationship between observed and theoretical values (Figs 1B, 2B). In addition, the arrangements of data points in Fréchet plots are approximately linear (\( R^2 > 0.92 \) in driver-gene definition cases and \( R^2 > 0.97 \) in driver-site definition cases), suggesting driver MFaTs follow the Fréchet distribution (Figs 1E, 2E). The results of a bootstrap simulation (Figs 1C, 1D, 2C, 2D) over the tail index as the shape parameter of the generalized extreme value distribution (GEV) revealed that the tail index of driver MFaTs is common among different cancer genome datasets (1.25 ± 0.62 with a 95% confidence interval), and its value is positive (Type II extreme value distribution or Fréchet distribution). Also, the result suggested that the tail index can contribute to the estimate of systematic biases within a dataset, and in
Fig 1. Exploratory plots on cancer driver mutation MFaTs in the total-tumor analysis with driver-gene definitions.
In these figures, the “ICGC-IntOgen,” for example, denotes a set of mutations as an intersection of the ICGC mutations with the IntOgen driver-gene definition (for details, see Materials and Methods).
(A) Density plot of MFaTs. The colored solid line is probability density of observations, the black solid line is the probability density function (PDF) of GEV, and the dotted lines are the PDFs of the GPD and Pareto distributions.
(B) Q-Q plot of MFaTs. Here, “b” denotes the number of genomic sites of beneficial mutations considered, “th” denotes the effective threshold against MFaTs when selecting mutations according to ranks (for details, see Materials and Methods), and “d” denotes the parameter in the normalization in the Q-Q plot. The straight line has the equation y = x. In the Q-Q plot, each point denotes a pair consisting of an observation and its corresponding theoretical value.
(C) Bootstrap distributions of the tail index. The white point indicates the median of the distribution. The black square shows the first and the third quantiles.
(D) Bootstrap confidence intervals. The dotted line shows the value that is in range of all confidence interval cases.
(E) Fréchet plot. The $R^2$ values and the equation for the linear regression lines are shown. The symbols “b” and “th” are as defined in (B).

Fig 2. Exploratory plots on cancer driver mutation MFaTs in the total-tumor analysis with driver-site definitions.
In these figures, symbols such as “ICGC-DoCM” are similar to Fig 1. However, the symbol “IntOgen,” for example, denotes a driver-site definition (for details, see Materials and Methods).
(A) Density plot of MFaTs. The colored solid line is the probability density of observations, the black solid line is the probability density function (PDF) of GEV, and the dotted lines are the PDFs of the GPD and Pareto distributions. Disagreement between the observed and theoretical densities in the case of ICGC-Bailey is likely due to the lack of data points ($b = 55$).
(B) Q-Q plot of MFaTs. Here, “b” denotes the number of genomic sites of beneficial mutations considered, and “d” denotes the parameter for normalization in the Q-Q plot. The straight line has the equation $y = x$. In the Q-Q plot, each pair of an observation and a theoretical value corresponds to the respective points.
(C) Bootstrap distributions of the tail index. The white point indicates the median of the distribution. The black square shows the first and the third quantiles. We excluded the cases of the Bailey driver-site definition from the process due to their lack of data points.
(D) Bootstrap confidence intervals. The dotted line shows the value that is in range of all confidence interval cases. Cases with the ICGC mutation frequency data and Bailey driver-site definition were omitted from the calculation.
(E) Fréchet plot. The $R^2$ values and the equation for the linear regression lines are shown. The symbol “b” is defined as in (B).

this case, the ICGC dataset harbors greater noise compared to other datasets (Figs 1C, 2C).

Finally, the results of a $\chi^2$ goodness-of-fit test did not exclude the possibility that the actual distribution of driver MFaTs is GEV (with a significance level $\alpha = 0.05$) (Tables I and II). In this statistical test, the null hypothesis $H_0$ that may be rejected is “the observed distribution is identical with the theoretical,” and the alternative hypothesis $H_1$ that may be accepted is “the observed distribution is different from the...
theoretical.” Thus, the former null hypothesis ($H_0$) is our claim. For example, in the case of ICGC-DoCM in the driver-site analyses, we cannot conclude that the two distributions are different because the null hypothesis is not rejected at a significance level $\alpha = 0.05$ ($p = 0.51 > 0.05$). This does not state that the two distributions are identical but weakly supports the claim that the distribution of driver MFaT is GEV.

Table 1. The $\chi^2$ goodness-of-fit test $p$-values of total tumors with driver-gene definition (rank < 200).

| Combination Case | MFaT       | Driver Definition | GEV $p$-value | Pareto $p$-value |
|------------------|------------|-------------------|---------------|-----------------|
| ICGC-IntOgen     | ICGC       | IntOgen           | 0.280         | 0.005           |
| ICGC-CGC         | ICGC       | CGC               | 0.301         | 0.009           |
| ICGC-Tokheim     | ICGC       | Tokheim           | 0.269         | 0.081           |
| COSMIC-IntOgen   | COSMIC     | IntOgen           | 1.000         | 0.758           |
| COSMIC-CGC       | COSMIC     | CGC               | 1.000         | 0.767           |
| COSMIC-Tokheim   | COSMIC     | Tokheim           | 1.000         | 1.000           |
| CHANG-IntOgen    | CHANG      | IntOgen           | 1.000         | 1.000           |
| CHANG-CGC        | CHANG      | CGC               | 1.000         | 1.000           |
| CHANG-Tokheim    | CHANG      | Tokheim           | 1.000         | 1.000           |

Note that in this setting, a higher significance level such as $p < 0.05$ indicates that the observed and theoretical distributions are different, and a lower significance level such as $p = 0.50$ weakly supports a notion that the two distributions are alike. In this table, the GEV $p$-values are all above 0.25 maintaining the null hypothesis ($H_0$) that the two distributions are identical, while some of the Pareto $p$-values has fallen below 0.05 accepting the alternative hypothesis ($H_1$) that the two distributions are different.

Table 2. The $\chi^2$ goodness-of-fit test $p$-values of total tumors with driver-site definition.

| Combination Case | MFaT       | Driver Definition | GEV $p$-value | Pareto $p$-value |
|------------------|------------|-------------------|---------------|-----------------|
| ICGC-IntOgen     | ICGC       | IntOgen           | 0.405         | 0.336           |
| ICGC-DoCM        | ICGC       | DoCM              | 0.529         | 0.316           |
| ICGC-Tokheim     | ICGC       | Bailey            | 0.244         | 0.203           |
| COSMIC-IntOgen   | COSMIC     | IntOgen           | 1.000         | 1.000           |
| COSMIC-DoCM      | COSMIC     | DoCM              | 1.000         | 1.000           |
| COSMIC-Bailey    | COSMIC     | Bailey            | 1.000         | 0.672           |
| CHANG-IntOgen    | CHANG      | IntOgen           | 1.000         | 1.000           |
| CHANG-DoCM       | CHANG      | DoCM              | 1.000         | 1.000           |
| CHANG-Bailey     | CHANG      | Bailey            | 0.632         | 0.625           |

Note that in this setting, a lower significance level with higher $p$-value such as $p = 0.50$ weakly supports a notion that the observed and theoretical distributions are alike. In this table, $p$-values are equal or higher in the GEV than the Pareto distribution.

Collectively, the total-tumor analysis, using both driver-gene and driver-site definitions, strongly suggested that cancer driver mutation MFaTs in total-tumor datasets are appropriately modeled by the Type II extreme value (Fréchet) distribution.

Bayesian MCMC Approach Confirms That Tumor-Type-Specific Driver MFaT Distributions Are Also Fréchet Type

We then asked whether driver MFaT distributions were also of the Fréchet type in each tumor type. We used the RTCGA dataset for this analysis regarding data availability.
The results of the “total”-tumor analysis using the RTCGA dataset (Figs 3A and 3B) confirmed that estimated values of GEV parameters in the case of the RTCGA dataset are reproducible and independent of differences in driver-gene definitions, as shown with the ICGC, COSMIC and CHANG datasets (Figs 1C and 2C). Tumor-type-specific analyses of eight tumor types (Figs 3C and 3D) confirmed that the results were similar to the results of the total-tumor analysis. Parameter estimation by Bayesian MCMC (Fig 3C) showed the MFaT distributions belonged to the Fréchet type, although some degree of variability in GEV parameters according to the differences in tumor types were observed. In addition, the comparison of histograms of observations to estimated densities (Fig 3D) confirmed that the probability distribution of GEV approximately describes the actual distributions of type-specific driver MFaTs. The arrangements of data points in Fréchet plots are approximately linear ($R^2 > 0.90$ in any of eight tumor types), suggesting that driver MFaTs follow the Fréchet distribution (Fig 3E). Finally, the results of the $\chi^2$ goodness-of-fit test did not reject the null hypothesis that the distribution is GEV, except for the case of THCA (Table 3).

Table 3. The $\chi^2$ goodness-of-fit test $p$-values of the RTCGA dataset with tumor-type-specific driver-genes definition (rank < 50).

| Dataset | Tumor Type | GEV $p$-value |
|---------|------------|---------------|
| RTCGA   | BLCA       | 1.000         |
| RTCGA   | BRCA       | 0.434         |
| RTCGA   | HNSC       | 1.000         |
| RTCGA   | LIHC       | 1.000         |
| RTCGA   | LUAD       | 0.506         |
| RTCGA   | PRAD       | 1.000         |
| RTCGA   | SKCM       | 0.381         |
| RTCGA   | THCA       | 0.002         |

Note that in this setting, a higher $p$-value such as $p = 0.50$ weakly supports a notion that the observed and theoretical distributions are alike. In this table, $p$-value are above 0.30 except for the case of THCA.

Collectively, the parameter estimation by the Bayesian MCMC approach confirmed that the probability distributions of tumor-type-specific driver MFaTs are also of the Fréchet type, as shown for the total tumors (Figs 1, 2, 3A, and 3B).

GEV-Binomial Model Estimates Mutational Selection Coefficients (MSCs) of Cancer Driver Protein Mutations

Here, we assessed, through the estimation of MSCs, the degree of contribution of each amino acid mutation in driver proteins to oncogenesis by applying the Bayesian approach in the framework of EVT using the RTCGA mutations dataset. We recognize that a set of assumptions are required for estimating MSCs through MFaTs in the framework of EVT. The details of such assumptions are provided in the Discussion section.

In this series of analyses, we were able to compare these mutational fitness effects estimated by mutation frequencies between tumor types, because the estimators utilize MFaTs normalized by respective sample counts. This normalization thus far is independent of the classes of genes and other features of interest. Thus, EVT as a field of population genetics is consistent with the quantitative comparison of mutational fitness effects among tumor types involving both oncogenes and tumor suppressor genes.

For example, we estimated the MSCs of BRAF mutations among various tumor types (Fig 4A). The estimated MSC was the highest for BRAF V600E in thyroid cancer (THCA), followed by BRAF V600E in skin cutaneous melanoma (SKCM). The BRAF
Fig 3. Bayesian MCMC approach for estimating tumor-type-specific GEV parameters.

In these analyses, only mutations extracted with driver-gene definitions are analyzed. In the total-tumor analysis, we used the IntOgen, CGC, and Tokheim driver-gene definitions, and in the case of tumor-type-specific analyses, we used only the IntOgen definition.

(A) Density plot of MFaTs in the cases of the RTCGA total-tumor driver-gene definitions. The colored solid line is the probability density of observations, the black solid line is the probability density function (PDF) of GEV, and the dotted lines are the PDFs of the GPD and Pareto distributions, respectively. Here, “b” denotes the number of genomic sites of beneficial mutations considered, and “th” denotes the effective threshold against MFaTs when selecting mutations according to ranks (for details, see Materials and Methods).

(B) Bootstrap distributions of GEV parameters in the total-tumor analysis. We used the maximum likelihood method for the parameter estimation. The white point indicates the median of the distribution. The black square shows the first and the third quantiles.

(C) Bootstrap distributions of GEV parameters in the tumor-type-specific analysis. We used the Bayesian MCMC method for the parameter estimation. The white point indicates the median of the distribution. The gray square shows the first and the third quantiles.

(D) MFaT histograms with tumor-type-specific GEV densities. Black bars show frequencies in the histogram. The colored lines are estimated densities.

(E) Type-specific Fréchet plots. The $R^2$ values and the equations of the linear regression lines are shown. Here, “b” denotes the number of genomic sites of beneficial mutations considered.

V600E mutation in lung adenocarcinoma (LUAD) or other amino acid changes in BRAF showed relatively low MSCs.

Fig 4. The EAP Estimates of Driver Mutation MSCs.

Tumor-type-specific posterior distributions of mutational selection coefficients (MSCs) and EAP (expected à posteriori) estimates in driver-protein mutations calculated by the GEV-binomial model (for details, see Materials and Methods). For ease of plotting, we discretized probability densities of posteriors to obtain probability masses. Among the probability masses that count the total number of tumor samples (i.e., the parameter $m$ in the GEV-binomial model), we omitted those tail probabilities that are smaller than 1/1000 in the plot. The white points at the center of each violin in the plots represent EAP estimates given the distribution. The shapes of plotted distributions have information of MFaT tails that cannot be modeled by a simple binomial model.

(A) Violin plots of tumor-type-specific posterior distributions of MSCs in genes. Violins are colored according to each amino acid substitution to enable a comparison among MSC distributions of identically coordinated mutations observed in different tumor types. Genes that have four “mutation-tumor type” combinations or more in the RTCGA dataset are shown.

(B) Violin plots of tumor-type-specific posterior distributions of MSCs in the TP53 protein sequences. The violins are arranged according to the amino acid coordinates of the mutations and are colored according to the amino acid residue after the substitution introduced by the mutation.

In the literature, the impact on the fitness effect of the BRAF V600E mutation is likely different among tumor types, including thyroid cancer, skin cutaneous melanoma, and lung adenocarcinoma. The mutation has been associated with poor prognosis and
mortality in patients with papillary thyroid cancer\textsuperscript{[25–27]}. This is likely the strongest association between these three tumor types. The prevalence of this mutation is reported also in melanoma\textsuperscript{[28]} and is shown to induce metastasis of melanoma in mice\textsuperscript{[29]}. However, this is conditional to PTEN loss, suggesting weaker association compared to the case of thyroid cancer. The contribution of this mutation is even weaker in lung adenocarcinoma\textsuperscript{[30]}. The estimated MSCs of BRAF mutations among tumor types are consistent with these known facts.

Although not much is known for the impact of each mutations on the fitness effect among the tumor types for other genes in Fig 4A, oncogenes (i.e., HRAS and PIK3CA) tend to have a few “mutation-tumor type” combinations that show relatively high MSCs. In contrast, in the case of tumor suppressor genes (i.e., CDKN2A, CTNNB1, and SPOP), MSCs were small and the differences between “mutation-tumor type” combinations are also small. This tendency is evident in the case of TP53 (Fig 4B).

Discussion

Driver MFaTs Are Expectations of VAFs in the Big Bang Model of Cancer Evolution

According to the Big Bang model of the cancer genome\textsuperscript{[3]}, individual tumor samples have independent and different oncogenic trajectories. Moreover, a driver mutation that has had an impact on carcinogenesis within a single sample is a “public” mutation shared by all cells in the tumor. In one sample, a class of driver mutations may have an impact on carcinogenesis, while in another sample, it does not. If a mutation at a certain driver-site within a tumor sample has an impact on carcinogenesis, then all of the tumor cells have that mutation, and about that driver site, VAF = 1.0 holds if the tumor sample purity is ideal (i.e., 100 %). Similarly, if a mutation at a certain driver site within a tumor sample did not impact the oncogenic process, that mutation should not be observed in any of the tumor cells, and about that driver site, VAF = 0.0 holds regardless of tumor sample purity.

If these two important aspects of cancer driver mutation VAFs are considered, the value of aggregated tumor VAF (i.e., VAF that is aggregated across tumor samples in a given dataset; for details, see Supplementary Materials in S1 Appendix) will be equal to the value of mutant allele frequency among tumors (MFaT), which is given by the ratio of mutated samples to total samples. The use of MFaT will thus be a powerful approach in normalizing, investigating, and deciphering the records of preclonal evolution in large-scale cancer genome data.

The Nine Assumptions in Extreme Value Cancer Genomics

With the framework of SSWM (strong selection and weak mutation) in population genetics\textsuperscript{[6,7]}, we were able to mechanistically and stochastically describe the preclonal evolution of cancer. To achieve this and perform valid extreme value analysis over cancer driver MFaTs, we propose the following nine assumptions. These assumptions will specify the scope of the application of the theory and will enable precise interpretation of the results.

1. The Single Macro-Environment Assumption: Cancer evolution is an evolutionary process in which cancer evolves and adapts to a single macro-environment. This is based on a notion that, from a viewpoint that considers multiple sets of patients and multiple tissue types, the oncogenic and progressive processes are repeatable (the macro-repeatability assumption). This
also is a prerequisite for two other assumptions (i.e., the proportionality assumption and the maximization assumption) stated below.

2. **The Infinite Micro-Environments Assumption:** In a microscopic point of view that focuses on patients’ genetic background, physical condition, tissue type, and tissue micro-structure, as well as the genetic diversity of cancer itself, the uncertainty of evolutionary processes, and many other critical aspects of cancer evolution, we have infinite cases of possible cancer micro-environments. This is a prerequisite of the independent and identical distribution assumption and the maximization assumption stated below.

3. **The Additivity Assumption:** Cancer is a diploid neoplasm that proliferates by cell division. In the field of population genetics, a lot of effort has been devoted solely to monoploid organisms. To translate this knowledge to population genetics of cancer cells, we need an additional assumption for the fitness of heterozygotes: the fitness of a heterozygote is exactly intermediate between the two associated homozygotes \[13\]. This assumption allows us to apply the genetics of cell-dividing monoploids to cancer.

4. **The Strong Selection Assumption:** The mutations in the scope are all either beneficial or deleterious, and no neutral mutations are considered \[7\]. Because we presume driver mutations are all beneficial in cancer evolution \[4\], we can safely accept this assumption over driver mutations. In the analysis, the validity of this assumption is ensured by removing mutations with lower observed fitness gains. For the case of total-tumor analysis with driver-gene definitions, we focused on mutation sites with the top 200 MFaT values, and for type-specific analysis, we focused on the top 50.

5. **The Weak Mutation Assumption:** The beneficial mutations are fixed independently in the population \[6\]. This is equivalent to the exclusion of clonal interference from the scope. In the stage of preclonal evolution of cancer, it is plausible to assume that the probability of acquiring multiple mutations within a single generation is sufficiently low. This leads to an approximation of the presence of a single clone at any time. In combination with the strong selection assumption, we obtain the SSWM fitness landscape in which the population adapts to the environment following a series of selective sweep.

6. **The Proportionality Assumption:** From the above-stated single macro-environment assumption, cancer adapts and evolves in a single macro-environment. Here, we assume that MFaT is proportional to the fitness gain that a mutation yields in the single macro-environment. This enables quantification of mutational fitness gains in the cancer macro-environment using MFaT.

7. **The Continuous Fitness Effect Assumption (Continuity Assumption):** So far, the fitness effect that a mutation yields in cancer evolution is only indirectly observable. Thus, the precise formulation of the probability distribution of that variable is unknown as well as its existence. However, the repeatability of cancer evolution implies that such fitness effects by a mutation have a certain probability distribution, and the complexity suggests that the variable is approximately continuous. From the above, we assume that the mutational fitness gains in cancer evolution have a certain continuous probability distribution.

8. **The Independent and Identical Distribution Assumption (IID Assumption):** In general, if two mutations had different genomic coordinates,
then the phenotypic effects of the two mutations also vary. This is because different genomic sites encode different structures and functions of the organism. For example, mutations in the first and third letters of triplets in the codon table will yield different amino acid substitutions (i.e., the first- and third-letter substitutions are independent). In contrast, the fitness effect that a phenotypic effect of a given mutation confers on the organism is dependent on the environment in which the organism adapts and evolves. From the above-mentioned infinite micro-environments assumption, we have numerous cases of such environments in cancer evolution. Under these possible environments, we assume that the fitness effect of a given, single mutation has a certain probability distribution that is independent of a genomic site of the mutation (i.e., any given mutation have identical probability distribution). Then, the value of the fitness effect of a mutation is independent and identically distributed (i.i.d.) across genomic sites. This is equivalent to excluding cases of interaction of mutation effects (i.e., epistasis) from the scope.

9. The Maximization Assumption: From the infinite micro-environments assumption, the number of cases of possible cancer micro-environments is infinite. Fitness effect caused by a mutation at a given genomic site have a different value in a different micro-environment. It is known that, in the preclonal evolution of cancer, such micro-environments play critical roles in the evolution of cancer cells. Here, from the single macro-environment assumption, we consider adaptation of cancer cells to the single cancer macro-environment in the preclonal evolution step. We consider that, under such a macro-environment, cancer cells are selected based on combinations of different micro-environments and different fitness effects of cancer driver mutations. Thus, we assume that mutational fitness gain at a given cancer driver site is maximized across possible values (the block maxima model) after selection in the preclonal evolution. This is consistent with the idea of “survival of the fittest” in the theory of natural selection.

\[ S_i = \max(S_{i,1}, S_{i,2}, \ldots, S_{i,n}) \quad (n \to \infty) \] (21)

Here, \( S \) is a selection coefficient, \( i \) is an index for genomic sites, and \( n \) is the number of possible micro-environments.

Although some of the above assumptions may not fit with our current knowledge of cancer biology, the results of our analysis suggest that they may hold at least for the first approximation.

The Probability Distribution of Cancer Driver MFaTs Is Likely the Fréchet Type

At the current level of our knowledge, statistical properties of fitness effect distributions of cancer driver mutations are to a large extent unknown. A primary reason for this is the lack of a method that directly measures such fitness effects. However, the population genetics of cancer cells based on SSWM and extreme value theory has the flexibility to deal with the behavior of cancer cell populations while avoiding this problem. With these frameworks and several appropriate assumptions, we are able to discuss at least some of the properties of the fitness effect distributions from observed mutation frequencies as a result of adaptation and evolution of cells, without knowing the exact fitness effect distribution in each mutation in each cell. In other words, if the distribution of interest is continuous, the distribution of maxima drawn from samples from that distribution will be one of these three types: Gumbel, Fréchet, or Weibull [31].
Many “ordinary” probability distributions, such as normal, exponential, and gamma, belong to the Gumbel maximum domain of attraction. Based on this fact and discussion that Fréchet-type and Weibull-type distributions are not “biological,” Gumbel-type distributions have been justified as distributions of fitness effects of beneficial mutations. In addition, a historical background in which such a fitness effect distribution has been considered to be exponential also supported this preconception (Fisher’s geometric model; [24,33]). However, recent theoretical advances clarified that distributions that belong to the Fréchet and Weibull domains are also possible [14].

Biological experiments roughly supported this non-Gumbel hypothesis of fitness effects of beneficial mutations. A recent experiment involving two virus strains showed that fitness effects practically yielded by a beneficial mutation do not follow an exponential distribution [34–36]. The mathematical background of this experiment is that, if values of the fitness effects have a right-truncated distribution due to their upper limit being characteristic of a given experimental setting, then the maxima of values drawn from that distribution will follow the Weibull distribution. In this experiment, a fitness effect value of a mutation is quantified as a count of formed plaques and is a proxy for virus particles [37,38]. A quantity based on a count of biological entities is among the most powerful candidates for a variable to quantify the fitness effects of a mutation.

Also, an Escherichia coli experiment designed as an application of EVT empirically confirmed that the fitness effects of fixed beneficial mutations follow a distribution with a positive mode [39]. Although experimental settings including the method to quantify mutant fitness are greatly different from this study, the Fréchet distribution as a statistical distribution that describes the behavior of fitness effects of fixed driver mutations in tumor samples also has a positive mode in its mathematics.

Our study suggested that the distribution of fitness effects of driver mutations calculated from a sample frequency in a large-scale sequence dataset is of the Fréchet type (Figs 1-3), while it also allows distributions of the fitness effects of the individual mutations to remain unknown. The results of goodness-of-fit tests (Tables 1-3) did not reject the null hypothesis that the given two distributions are identical, supporting the possibility that the distribution of MFaTs as estimates of mutational selection coefficients (MSCs) is Fréchet distribution. Mathematically, a zero value of the shape parameter of the generalized extreme value (GEV) distribution means the Gumbel type, and a positive value means the Fréchet type. Also, the Fréchet distribution itself belongs to the attraction domain of the Fréchet distribution. These results not only present a problem to the previously held Gumbel hypothesis [14] from a practical point of view but also suggest the applicability of the Fréchet distribution in cancer genomics (Fig 4). In the case of THCA (Table 3), the null hypothesis was rejected in the goodness-of-fit test and it did not reproduce this result. It is obvious from the graph that this irreproducibility is due to the lack of mutations used in the analysis (Fig 3E).

The Applicability of Extreme Value Theory in Cancer Genomics

The posterior distributions of tumor-type-specific mutational selection coefficients (MSCs) of driver mutations calculated by the GEV-binomial model (Fig 4) contain information of distribution tails described by EVT. In the violin plots, because the EAP estimates drawn as white dots contain information of the tails that cannot be handled by a simple binomial model, those estimates have shifted to the right from the central point, as suggested by shapes of the posterior distributions. Such shifts are significant in posterior distributions of mutations, such as the S33P mutation in the CTNNB1 gene in the LIHC tumor type and the Q61R mutation in the HRAS gene in the THCA tumor type (Fig 4A). Similarly, while EAP estimates of driver mutation MSCs in the TP53 gene strongly reflect MFaTs that are mutant sample frequencies, these values also
reflect information of the tails so as to be more continuous (Fig 4B). The shifts in these estimates suggest the applicability of EVT in cancer genomic analyses that entail estimation of fitness effects of beneficial mutations.

Conclusion

Based on statistical data analysis involving multiple tumor types and multiple definitions of cancer driver mutations, this study not only demonstrates that EVT helps us to understand the statistical distribution of driver-mutation frequencies in the cancer genome, which is a critical aspect in cancer genetics, but also suggests its applicability in cancer genomics based on its potential to model the tail behavior of mutation frequency distributions.

Supporting information

S1 Appendix. Supplementary Materials. (PDF)
S1 Script. get_IcgcProteinAltering.py. (PY)
S1 File. database_ICGC_temp_PostMax.tsv.gz. (GZ)
S2 File. database_COSMIC_temp_PostMax.tsv.gz. (GZ)
S3 File. database_CHANG_temp_PostMax.tsv.gz. (GZ)
S4 File. driver_TotalGene_database_IntOGen.tsv. (TSV)
S5 File. driver_TotalGene_database_CGC.tsv. (TSV)
S6 File. driver_TotalGene_database_Tokheim.tsv. (TSV)
S7 File. driver_TotalSite_database_IntOGen.tsv. (TSV)
S8 File. driver_TotalSite_database_DoCM.tsv. (TSV)
S9 File. driver_TotalSite_database_Bailey.tsv. (TSV)
S10 File. database_RTCGA_temp_ProteinAlteringSnvDoubleton.tsv.gz. (GZ)
S11 File. database_RTCGA_temp_PostMaxTotal.tsv. (TSV)
S12 File. database_RTCGA_temp_PostMaxType.tsv. (TSV)
S13 File. driver_TypeGene_database_IntOGen.tsv. (TSV)
S14 File. drivertype_RTCGA-IntOGen_temp_PreRank.tsv. (TSV)
Acknowledgments

NT thanks Takao Suetake for his support and encouragement throughout this research. We thank the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST) for its financial support.

Author contributions

Conceptualization: Natsuki Tokutomi.

Formal analysis: Natsuki Tokutomi.

Funding acquisition: Sumio Sugano.

Investigation: Natsuki Tokutomi.

Methodology: Natsuki Tokutomi.

Project administration: Kenta Nakai and Sumio Sugano.

Resources: Natsuki Tokutomi.

Supervision: Kenta Nakai and Sumio Sugano.

Validation: Natsuki Tokutomi.

Visualization: Natsuki Tokutomi.

Writing – original draft: Natsuki Tokutomi.

Writing – review: Sumio Sugano.

References

1. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature genetics. 2016;48(3):238–244.

2. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature genetics. 2018;50(6):895–903.

3. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nature genetics. 2015;47(3):209–216.

4. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306–313.

5. Tomasetti C, Marchionni L, Nowak MA, Parmigiani G, Vogelstein B. Only three driver gene mutations are required for the development of lung and colorectal cancers. Proceedings of the National Academy of Sciences. 2015;112(1):118–123.

6. De Visser JAG, Krug J. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics. 2014;15(7):480–490.

7. Orr HA. The genetic theory of adaptation: a brief history. Nature Reviews Genetics. 2005;6(2):119–127.
8. Jain K, Krug J, Park SC. Evolutionary advantage of small populations on complex fitness landscapes. Evolution: International Journal of Organic Evolution. 2011;65(7):1945–1955.

9. Rozen DE, Habets MG, Handel A, de Visser JAG. Heterogeneous adaptive trajectories of small populations on complex fitness landscapes. PLoS One. 2008;3(3):e1715.

10. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. Theoretical population biology. 2009;75(4):286–300.

11. Iwasa Y, Michor F, Nowak MA. Stochastic tunnels in evolutionary dynamics. Genetics. 2004;166(3):1571–1579.

12. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-order selection for evolvability in a large Escherichia coli population. Science. 2011;331(6023):1433–1436.

13. Gillespie JH. A simple stochastic gene substitution model. Theoretical population biology. 1983;23(2):202–215.

14. Joyce P, Rokyta DR, Beisel CJ, Orr HA. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. Genetics. 2008;180(3):1627–1643.

15. Calabrese P, Tavaré S, Shibata D. Pretumor progression: clonal evolution of human stem cell populations. The American journal of pathology. 2004;164(4):1337–1346.

16. Consortium ICG, et al. International network of cancer genome projects. Nature. 2010;464(7291):993.

17. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic acids research. 2017;45(D1):D777–D783.

18. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature biotechnology. 2016;34(2):155–163.

19. Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer cell. 2015;27(3):382–396.

20. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proceedings of the National Academy of Sciences. 2016;113(50):14330–14335.

21. Tamborero D, Rubio-Perez C, Deu-Pons J, Schroeder MP, Vivancos A, Rovira A, et al. Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. Genome medicine. 2018;10(1):25.

22. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nature methods. 2016;13(10):806–807.
23. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–385.

24. Orr HA. The population genetics of adaptation: the adaptation of DNA sequences. Evolution. 2002;56(7):1317–1330.

25. Kebebew E, Weng J, Bauer J, Ranvier G, Clark OH, Duh QY, et al. The prevalence and prognostic value of BRAF mutation in thyroid cancer. Annals of surgery. 2007;246(3):466.

26. Kim TH, Park YJ, Lim JA, Ahn HY, Lee EK, Lee YJ, et al. The association of the BRAFV600E mutation with prognostic factors and poor clinical outcome in papillary thyroid cancer: A meta-analysis. Cancer. 2012;118(7):1764–1773.

27. Xing M, Alzahrani AS, Carson KA, Viola D, Elisei R, Bendlova B, et al. Association between BRAF V600E mutation and mortality in patients with papillary thyroid cancer. Jama. 2013;309(14):1493–1501.

28. Si L, Kong Y, Xu X, Flaherty KT, Sheng X, Cui C, et al. Prevalence of BRAF V600E mutation in Chinese melanoma patients: large scale analysis of BRAF and NRAS mutations in a 432-case cohort. European journal of cancer. 2012;48(1):94–100.

29. Dankort D, Curley DP, Cartlidge RA, Nelson B, Karnezis AN, Damsky Jr WE, et al. Braf V600E cooperates with Pten loss to induce metastatic melanoma. Nature genetics. 2009;41(5):544–552.

30. Paik PK, Arcila ME, Fara M, Sima CS, Miller VA, Kris MG, et al. Clinical characteristics of patients with lung adenocarcinomas harboring BRAF mutations. Journal of clinical oncology. 2011;29(15):2046.

31. Beisel CJ, Rokyta DR, Wichman HA, Joyce P. Testing the extreme value domain of attraction for distributions of beneficial fitness effects. Genetics. 2007;176(4):2441–2449.

32. Orr HA. The distribution of fitness effects among beneficial mutations. Genetics. 2003;163(4):1519–1526.

33. Orr HA. The distribution of fitness effects among beneficial mutations in Fisher’s geometric model of adaptation. Journal of theoretical biology. 2006;238(2):279–285.

34. Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA. Beneficial fitness effects are not exponential for two viruses. Journal of molecular evolution. 2008;67(4):368.

35. Rokyta DR, Joyce P, Caudle SB, Wichman HA. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nature genetics. 2005;37(4):441–444.

36. Ferris MT, Joyce P, Burch CL. High frequency of mutations that expand the host range of an RNA virus. Genetics. 2007;176(2):1013–1022.

37. Chao L. Fitness of RNA virus decreased by Muller’s ratchet. Nature. 1990;348(6300):454–455.
38. Rokyta D, Badgett M, Molineux I, Bull J. Experimental genomic evolution: extensive compensation for loss of DNA ligase activity in a virus. Molecular Biology and Evolution. 2002;19(3):230–238.

39. Rozen DE, De Visser JAG, Gerrish PJ. Fitness effects of fixed beneficial mutations in microbial populations. Current biology. 2002;12(12):1040–1045.

40. Cherkasov A, Jones SJ. Structural characterization of genomes by large scale sequence-structure threading. BMC bioinformatics. 2004;5(1):1–16.

41. Orr HA. The population genetics of beneficial mutations. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1544):1195–1201.
