Why Mutant Allele Frequencies in Oncogenes Peak Around 0.40 and Rapidly Decrease?

Kasthuri Kannan[1,2] *and Adriana Heguy*[1]

Abstract

The mutant allele frequencies in oncogenes peak around 0.40 and rapidly decrease. In this article, we explain why this is the case. Invoking a key result from mathematical analysis in our model, namely, the inverse function theorem, we estimate the selection pressures of the mutations as a function of germline allele frequencies. Under complete dominance of oncogenic mutations, this selection function is expected to be linearly correlated with the distribution of the mutant alleles. We demonstrate that this is the case by investigating the allele frequencies of mutations in oncogenes across various cancer types, validating our model for mean effective selection. Consistent with the population genetics model of fitness, the selection function fits a gamma distribution curve that accurately describes the trend of the mutant allele frequencies. While existing equations for selection explain evolution at low allele frequencies, our equations are general formulas for natural selection under complete dominance operating at all frequencies. We show that selection exhibits linear behavior at all times, favoring dominant alleles with respect to the change in recessive allele frequency. Also, these equations show, selection behaves like power-law against the recessive alleles at low dominant allele frequency.

Keywords: oncogenes, allele frequencies, natural selection, gamma distribution, inverse function theorem

1. Introduction

Cancer is an evolutionary disease with mutation acting as a unit of selection and, therefore, quantifying selection is necessary. Since the distribution of the allele frequencies of somatic mutations reflects the nature of selection underlying the mutant clones, modeling this distribution is essential. Recently, Williams et al. [1] showed neutral tumor evolution results in a power-law distribution of the mutant allele frequencies, and this law fits 303 of 904 cancers of various types. While neutral evolution remains an important aspect in several cancer types, the distribution of the allele frequencies of mutations in genes that undergo positive selection, has not been determined so far. This is difficult to model because the allele frequencies of positively selected mutations are related to their functional consequences and therefore we have to take into account the degree of dominance exhibited by these mutations. However, in the case of oncogenes, which are primarily dominant, determining the distribution of the mutant allele frequencies should be feasible. It is worth noting that in the context of human polymorphisms, the allele frequencies of new deleterious mutations has been studied and fitness distribution is shown to follow a gamma distribution [2]. Nonetheless, the fitness function and distribution of allele frequencies of oncogenic mutations in tumors have not been described.

The mutant allele frequencies in oncogenes peak around 0.40 and rapidly decrease. We built a mathematical model to describe the trend of these frequencies. We assumed complete dominance of oncogenic mutations, although we realize dominance (or partial dominance) can be modeled as a function of the functional impact of these mutations, which can be derived from algorithms such as PolyPhen and SIFT [3] [4]. The scope of this article is restricted to describe the general tendency of the frequencies rather than considering the impact of the mutations on their frequencies. By taking advantage of a key result in mathematical analysis, namely, the inverse function theorem, we estimate the mean effective selection of the mutations as a function of germline allele frequencies. Under complete dominance of oncogenic mutations, this selection function is expected to be linearly correlated with the distribution of the mutant alleles. We demonstrate that this is the case by investigating the allele frequencies of mutations in oncogenes across various cancer types, validating our model for mean effective selection. Consistent with population genetics model of fitness, the selection function fits a gamma distribution curve that accurately describes the trend of the mutant allele frequencies.

This model infers mean effective selection for oncogenic mutations without considering other alterations in the DNA that could change the dynamics of the tumor micro-environment. Moreover, this measure is an *effective selection coefficient* in the sense that it is selection coefficient with respect to reduction in germline allele frequencies. Combining this estimate with other modifications, such as copy number changes, and integrating the functional impact of the resulting protein will help in understanding the evolution of the mutant clones in various tumors.

[1]New York University - School of Medicine
[2]Corresponding author

Although the equations that we derive are currently applied in the context of oncogenes, these are general formulas for natural selection under complete dominance operating at all frequencies. While being consistent with known formulas that explain evolution at low frequencies, one of these equations show that under low frequencies of the dominant alleles, selection against recessive alleles behave in a power-law like manner, reiterating the powerful role of natural selection. This would also explain the reason some tumors undergo rapid clonal evolution. Further, at high frequencies of the dominant alleles, the linear expansion exhibited by selection with respect to change in recessive germline allele frequencies could partly be the reason behind drug resistance, under dominance.

2. Determining Selection Coefficients

We use the standard model described in [5] for selection under complete dominance. An illustration of this model for mutations when the frequencies are not time dependent is shown in Supplementary Figure 1. Under this model, if s is the co-efficient of selection, p and q are the allele frequencies with $p + q = 1$, the new genotype frequencies f, in terms of p and q are given by,

$$f(p) = \frac{p^2 + pq}{1 - sq^2}; \quad f(q) = 1 - f(p) = \frac{q - sq^2}{1 - sq^2} \tag{1}$$

Using this model for time-dependent allele frequencies, we derive a new equation for the number of somatic mutations in terms of the germline allele frequency.

At time t, let M be a mutation with $p_t^2, 2p_tq_t$ and q_t^2 as allele frequencies of mutant homozygous, mutant heterozygous and germline genotypes of M, respectively, with $p_t + q_t = 1$. Let the strength of the selection be expressed as a co-efficient of selection, s, which is proportional to the reduction of the germline genotype, compared to the mutant genotypes which are favored for the tumor growth. In reality, the selection co-efficient should be a function of time. However, we can consider s as mean selection acting over time and so the time dependence can be omitted. Thus, in our model, we assume the frequencies are time-dependent, but the selection is time-independent. If the fitness of the homozygous and heterozygous mutant genotypes are taken to be 1 as it is likely the case in oncogenes, the fitness of the germline genotype which is selected against, is then $1 - s$. Thus, after one generation, the new mutation frequency, M_{q_t}, in terms of the recessive germline allele q_t is given by equation 1, which is,

$$M_{q_t} = \frac{q_t - sq_t^2}{1 - sq_t^2}. \tag{2}$$

Hence, change in mutation frequency, ΔM_{q_t}, resulting in a small time interval Δt of selection is,

$$\Delta M_{q_t} = [M_{q_t} - q_t]\Delta t = -\frac{sq_t^2(1 - q_t)}{1 - sq_t^2}\Delta t.$$

Let T_0 and T be the time of tumor initiation and the time of tumor biopsy, respectively. Then,

$$M(q_T) = \int_{T_0}^{T} dM_{q_t} = -\int_{T_0}^{T} \frac{sq_t^2(1 - q_t)}{1 - sq_t^2} dt.$$

Denoting $u = q_t$, and applying inverse function theorem,

$$\frac{du}{dt} = q_t' = \frac{1}{[q^{-1}]'(u)} = \frac{1}{F'(u)}$$

where $q(q_{t_i}) = q_{t_i}$ for any time point t_i and $F = q^{-1}$. We see that, if $u_0 = q_{T_0}$ and $u_T = q_T$ are initial and final frequencies, then

$$M(u_T) = -\int_{u_0}^{u_T} \frac{su^2(1 - u)}{1 - su^2} F'(u)du \tag{3}$$

This formula allows us to express the number of mutations purely in terms of allele frequencies. If λ is the fraction of reduction of germline alleles due to the selection pressure s, we expect F to follow the function,

$$F(u) = \frac{1-\lambda u}{u-\lambda u} \Rightarrow F'(u) = -\frac{1}{(1-\lambda)u^2}$$

and therefore, equation 3 is given by,

$$M(u_T) = \frac{s}{1-\lambda}\int_{u_0}^{u_T} \frac{1-u}{1-su^2}du \approx s_\lambda \int_{u_0}^{u_T}(1-u)du = s_\lambda\left(u_T - \frac{u_T^2}{2}\right) + s_\lambda C$$

where $s_\lambda = s/1-\lambda$ can be interpreted as *mean effective selection co-efficient* favoring mutant genotypes. If the co-efficient is extremely small at T_0, then $s_\lambda C \approx 0$, and hence the number of mutations in terms of the observed germline allele frequency q at time T is given by,

$$M(q) \approx s_\lambda\left(q - \frac{q^2}{2}\right) \tag{4}$$

and s_λ is therefore,

$$s_\lambda \approx M(q)\left[q - \frac{q^2}{2}\right]^{-1} \tag{5}$$

Also, from equation 1, since we have,

$$M(p) \approx p + s_\lambda p(1-p)^2, \tag{6}$$

if s_λ is very small or when p is small, the number of mutations $M(p)$ in terms of the mutant allele frequency p can be approximated by $s_\lambda p + p$. Therefore, we expect s_λ to be correlated with the distribution of the mutant allele frequencies for small values of s_λ. However, we note from equation 5 and equation 6 that extremely small values of q or $M(p) = 0$ will make s_λ unbounded or negative, and hence correlation may not be valid.

3. Results

We determined the selection coefficents s_λ for mutations in oncogenes to see if they are correlated with the distribution of the mutant allele frequencies. To do this, we first identified 574 proto-oncogenes from the Uniprot database [6] out of which 236 were exclusive to *Homo sapiens*. A total of 42, 525 mutations in these genes were queried from the cBio portal [7, 8] and 25, 848 mutations for which mutant allele frequencies were available, was retained (see Supplementary Table 1). The germline allele frequencies were computed by subtracting the mutant allele frequencies from 1 and $M(q)$ was normalized with the highest value.

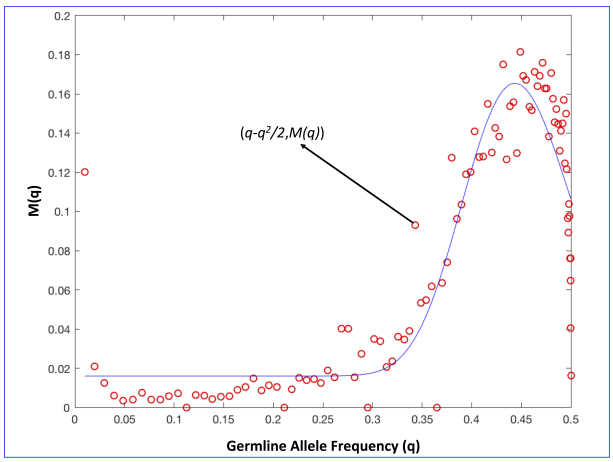

Figure 1: Scatter plot of $(q - q^2/2, M(q))$ and the gamma distribution curve fit

Equation 4 essentially states s_λ should be correlated with the random variable $M(q)$ under the random variable $q - q^2/2$. Therefore, it is natural to fit the data $(q - q^2/2, M(q))$. Since two fitness distributions, the gamma, and the exponential have been traditionally applied to model selection co-efficients [9], we employed both functions. Consistent with the fitness function of deleterious mutations in human polymorphisms [2], the gamma distribution curve fitted well with minimal residual error of 6×10^{-4} with fitting function with respect to $x = q - q^2/2$, given by,

$$g(\alpha, \beta, \rho, \delta; x) = \frac{\rho}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}} + \delta$$

where shape, scale, amplitude and offset parameters, α, β, ρ and δ were identified as 68.17, 0.006, 0.02, 0.01, respectively. MATLAB's `nlinfit` was used to fit the data with initial conditions $[1, 1, 0.5, 1]$ for the parameters.

Figure 1 shows the scatter plot of $(q - q^2/2, M(q))$ and the gamma distribution curve fit. This function allows us to compute s_λ for each mutation through equation 5, i.e., $s_\lambda = g(\alpha, \beta, \rho, \delta; x)/x$, where x is the transformed allele frequency, $q - q^2/2$. Since s_λ co-efficients are as much as the number of mutations and $M(p)$ is restricted by the frequency bin size, to find the correlation, we sub-sampled s_λ for $10,000$ iterations and considered mean correlation. Also, for the reasons discussed following equation 6, mutant allele frequencies greater than 0.90 and $M(p) = 0$ were not considered. The mean correlation was determined to be 0.79 with mean p-value $2.3e^{-12}$. We also determined the line of best fit (MATLAB's `polyfit` routine with degree 1) to infer the slope to find the optimal s_λ that would fit with the mutant allele frequencies. Figure 2 (a) shows the correlation between s_λ coefficients and $M(p)$ and figure 2 (b) shows optimal s_λ that fits the mutant allele frequencies.

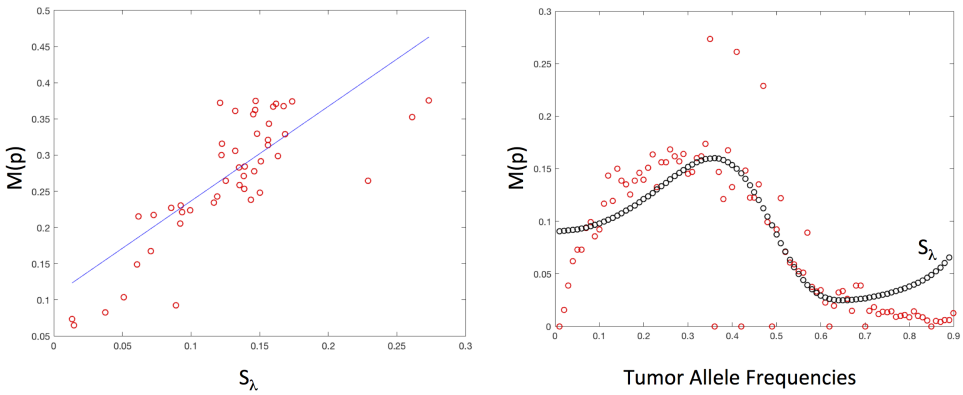

Figure 2: Correlation between s_λ and $M(p)$

4. Discussion

Good correlation between selection coefficients s_λ and the mutant allele frequencies explain why the frequencies are centered around 0.40 and rapidly decrease. Selection coefficients are maximized around this region and reduce exponentially.

In population genetics it is known that at low frequencies under dominance, selection for/against dominant alleles follow $O(p)$ and selection for/against recessive alleles grow at $O(q^2)$. This is because equation 1 tells us that,

$$\Delta f(p, q) = \pm \frac{spq^2}{1 - sq^2}.$$

While this is true when the unit of measure is generation time, integrating the selection equation with respect to time establishes that at all frequencies selection for the dominant alleles with respect to change in recessive allele frequency is linear. This can be seen by differentiating equation 4:

$$\frac{dM(q)}{dq} = s_\lambda(1 - q) = s_\lambda p. \tag{7}$$

Similar analysis on selection against recessive allele with respect to dominant allele frequency would reveal (See Supplementary Method 1),

$$\frac{dM(p)}{dp} = -s_\theta \frac{q^2}{p}. \tag{8}$$

where $-s_\theta = -s(1-\lambda)$ can be interpreted as effective selection against recessive alleles. These formulas, equations 7 and 8, as relative increase and decrease, define natural selection under complete dominance. Equation 7 is natural after all - for a given dominant allele frequency, the rate of change of dominant alleles in terms of (the loss of) germline alleles, should be proportional to the amount of selection that takes place. While both equations are consistent with known observation under low frequencies, equation 8 suggests that at low frequencies of the dominant alleles, selection against the recessive alleles act in a power-law like manner, demonstrating why positive

(natural) selection is a potent force. Under complete dominance, selection would maximize the dominant alleles at all costs.

In the context of cancer, especially in the evolution of mutant clones in oncogenes when mutant allele frequencies are small, the normal linear growth of the mutant alleles along with the power-law like loss of recessive germline alleles will doubly accelerate the progression of the mutant clones, possibly contributing to heterogeneity in the presence of competing mutations. Similarly, when recessive germline allele frequencies are small, and they grow quadratically, the progression of the mutant alleles will proceed linearly, still dominating and possibly conferring more resistance to therapy and giving rise to metastatic clones. Therefore, incorporating selection measures in evaluating functional impact of the mutations and assessing the aggressiveness of the tumors by taking the degree of selection into account will lead to better understanding of this complex evolutionary disease.

Acknowledgements

The author likes to acknowledge Dr. Friedrich Philipp who suggested the use of inverse function theorem in an online forum.

Notes

The MATLAB routine used to generate the results is available upon request.

References

1. Mark J Williams, Benjamin Werner, Chris P Barnes, Trevor A Graham, & Andrea Sottoriva. Identification of neutral tumor evolution across cancer types. *Nature Genetics* **48,** 238–244 (2016).
2. Eyre-Walker, Adam, Megan Woolfit, and Ted Phelps. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans. *Genetics* **173.2,** 891–900 (2006).
3. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Methods* **7(4),** 248–249 (2010).
4. Pauline C. Ng and Steven Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucl. Acids Res.* **31 (13),** 3812–3814 (2003).
5. D.S.Falconer. *Introduction to Quantitative Genetics* (The Roland Press Company, 1960).
6. The UniProt Consortium. UniProt: a hub for protein information. *Nucl. Acids Res.* **43 (D1),** D204–D212 (2015).
7. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science signaling* **6(269),** pl1 (2013).
8. Ethan Cerami et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2(401).** doi:doi:10.1158/2159-8290.CD-12-0095 (2012).
9. Gillespie, J. H. *The Causes of Molecular Evolution* (Oxford Series in Ecology and Evolution, 1994).

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).