



Why do mutant allele frequencies in oncogenes peak around .40 and rapidly decrease?

Kasthuri Kannan and Adriana Heguy

School of Medicine, New York University, New York, NY, USA

ABSTRACT

The mutant allele frequencies in oncogenes peak around .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 coefficients of the mutant alleles 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 behaviour at all times, favouring dominant alleles with respect to the change in recessive allele frequencies. Also, these equations show, selection behaves like power law against the recessive alleles at low dominant allele frequencies.

ARTICLE HISTORY

Received 4 May 2016
Accepted 29 July 2016


KEYWORDS

Oncogenes; allele frequencies; natural selection; gamma distribution; inverse function theorem

1. Introduction

Cancer is an evolutionary disease with mutant alleles 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, modelling this distribution is essential. Recently, Williams, Werner, Barnes, Graham, and Sottoriva (2016) showed neutral tumour 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

CONTACT Kasthuri Kannan  kasthuri.kannan@nyumc.org

 Supplemental data for this article can be accessed <http://dx.doi.org/10.1080/23737867.2016.1221328>

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 have been studied and fitness distribution is shown to follow a gamma distribution (Eyre-Walker, Woolfit, & Phelps, 2006). Nonetheless, the fitness function and distribution of allele frequencies of oncogenic mutations in tumours have not been described.

The mutant allele frequencies in oncogenes peak around .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 modelled as a function of the functional impact of these mutations, which can be derived from algorithms such as PolyPhen and SIFT (Adzhubei et al., 2010; Ng & Henikoff, 2003). 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 tumour micro-environment. Moreover, this measure is an *effective selection coefficient* in the sense that it is selection coefficient relative to the change in mutant/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 tumours. 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 (Equation (9)) shows that selection against recessive alleles behaves in a power-law-like manner, reiterating the powerful role of natural selection. This would also explain the reason some tumours undergo rapid clonal evolution. Further, at high frequencies of the dominant alleles, the linear expansion exhibited by selection could partly be the reason behind drug resistance, under dominance.

2. Determining selection coefficients

We use the standard model described in Falconer (1960) for selection under complete dominance. An illustration of this model for dominant and recessive alleles when the frequencies are not time dependent is shown in Supplementary Figure 1. Under this model, if s is the coefficient of selection, p and q are the allele frequencies with $p + q = 1$, the new

allele frequencies f , in terms of p and q are given by

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

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

At time t , let M be a population of cells consisting of mutant and germline genotypes with $p_t^2, 2p_tq_t$ and q_t^2 as 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 coefficient of selection, s , which is proportional to the reduction of the germline genotype, compared to the mutant genotypes which are favoured for the tumour growth. In reality, the selection coefficient 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 mutant allele frequency, M_{q_t} , in terms of the recessive germline allele frequency 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 mutant allele 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 tumour initiation and the time of tumour biopsy, respectively. Then,

$$M(q_T) = \int_{T_0}^T dM_{q_t} = - \int_{T_0}^T \frac{sq_t^2(1 - q_t)}{1 - sq_t^2} dt. \tag{3}$$

Let $\mathbb{C}[T_0, T]$ be the set of all continuous functions on the interval $[T_0, T]$ and let $q_t \in \mathbb{C}[T_0, T]$. Note that $\mathbb{C}[T_0, T]$ can be viewed as a set of random variables (since continuous functions on \mathbb{R} are measurable). Let this set when viewed as a set of random variables be denoted by $\mathbb{R}[T_0, T]$. Define $G: \mathbb{C} \rightarrow \mathbb{R}$ by $G(q_t) = q_t$, that is, G is an identity function from \mathbb{C} to \mathbb{R} . Since $\mathbb{C}[T_0, T]$ is a Banach space, inverse function theorem tells us that

$$\frac{dG(q_t)}{dt} = q'_t = \frac{1}{[G^{-1}]'(q_t)} \implies dt = [G^{-1}]'(q_t)dq_t$$

Denoting $F = G^{-1}$ and $u = q_t$, we see that, if $u_0 = q_{T_0}$ and $u_T = q_T$ are initial and final frequencies, then Equation (3) is

$$M(u_T) = - \int_{u_0}^{u_T} \frac{su^2(1 - u)}{1 - su^2} F'(u)du \tag{4}$$

This formula allows us to express the number of mutant alleles purely in terms of germline allele frequencies and decouples time dependence. Since G defines a function in \mathbb{C} as a

random variable, G essentially associates a cumulative distribution function (CDF) (or a probability density function) that describes the random variable. If λ is the fraction of reduction of germline alleles due to selection s at any time t , we can model the CDF associated with G either as a relative increase in mutant alleles or as an absolute increase in mutant alleles that corresponds to an absolute decrease in germline alleles. That is, an absolute decrease in germline allele frequency g_t due to λ reduction in germline alleles is given by,

$$G(g_t) = \frac{g_t - \lambda g_t}{1 - \lambda g_t}$$

and hence F is described by the CDF

$$F(g_t) = \frac{g_t}{\lambda(g_t - 1) + 1}.$$

Alternatively, a relative increase in mutant allele frequency is given by the CDF (Note: this CDF is a relative increase)

$$F(g_t) = \frac{1 - g_t}{g_t - \lambda g_t}.$$

This is a relative increase because (ignoring the quadratic and higher powers of λ)

$$F(g_t) = \frac{1 - g_t}{g_t - \lambda g_t} = \left(\frac{1 - g_t}{g_t} \right) + \lambda \left(\frac{1 - g_t}{g_t} \right).$$

In this paper, we will model F as a relative increase in the mutant alleles, and so the derivative of $F(u)$ is given by

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

and therefore, Equation (4) reduces to

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

where $s_\lambda = s/1 - \lambda$ can be interpreted as *mean effective selection coefficient* favouring mutant genotypes. If the coefficient 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

$$M(q) \approx s_\lambda \left(q - \frac{q^2}{2} \right) \quad (5)$$

and s_λ is therefore,

$$s_\lambda \approx M(q) \left[q - \frac{q^2}{2} \right]^{-1} \quad (6)$$

Also, from Equation (1), since we have

$$M(p) \approx p + s_\lambda p(1 - p)^2, \quad (7)$$

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

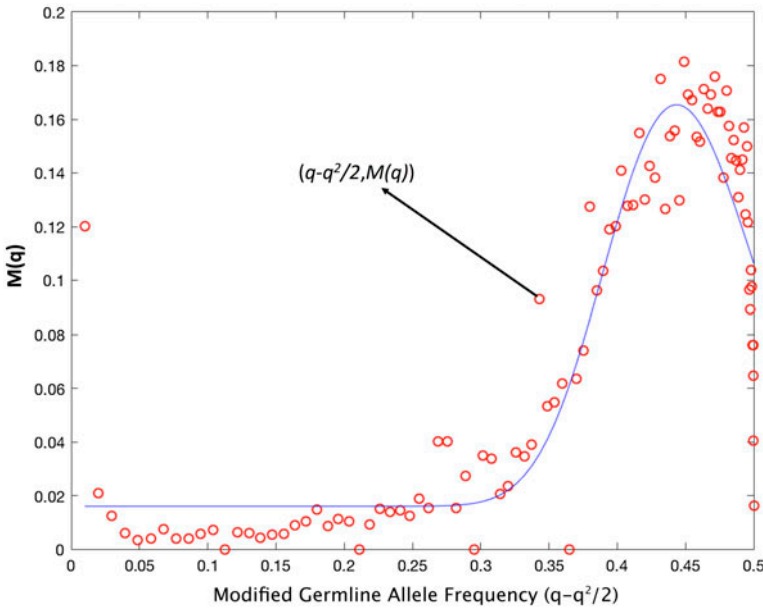


Figure 1. Scatter plot of $(q - q^2/2, M(q))$ and the gamma-distribution curve fit.

with the distribution of the mutant allele frequencies for small values of s_λ . However, we note from Equations (6) and (7) 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 coefficients 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 (The UniProt Consortium, 2015) out of which 236 were exclusive to *Homo sapiens*. A total of 42,525 mutations in these genes were queried from the cBio portal (Cerami, 2012; Gao, Aksoy, & Dogrusoz, 2013) and 25,848 mutations for which mutant allele frequencies were available, was retained (Supplementary Table 1 available upon request). The germline allele frequencies were computed by subtracting the mutant allele frequencies from 1 and $M(q)$ was normalized with its standard Euclidean norm.

Equation (5) 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 coefficients (Gillespie, 1994), we employed both functions. Consistent with the fitness function of deleterious mutations in human polymorphisms (Eyre-Walker et al., 2006), 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, .006, .02, and .01, respectively. MATLAB's¹ `nlinfit` was used to fit the data with initial conditions [1, 1, .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 (6), 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 coefficients 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 (7), mutant allele frequencies greater than .90 and $M(p) = 0$ were not considered. The mean correlation was determined to be .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.

4. Discussion

Good correlation between selection coefficients s_λ and the mutant allele frequencies explain why the frequencies are centred around .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 follows $O(p)$ and selection for/against recessive alleles grows 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 are linear. This can be seen by differentiating Equation (5):

$$\frac{dM(q)}{dq} = s_\lambda(1 - q) = s_\lambda p. \quad (8)$$

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

$$\frac{dM(p)}{dp} = -s_\theta \frac{q^2}{p}. \quad (9)$$

where $-s_\theta = -s(1 - \lambda)$ can be interpreted as effective selection against recessive alleles. These formulas, Equations (8) and (9), as relative increase and decrease, define natural selection under complete dominance. Equation (8) 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

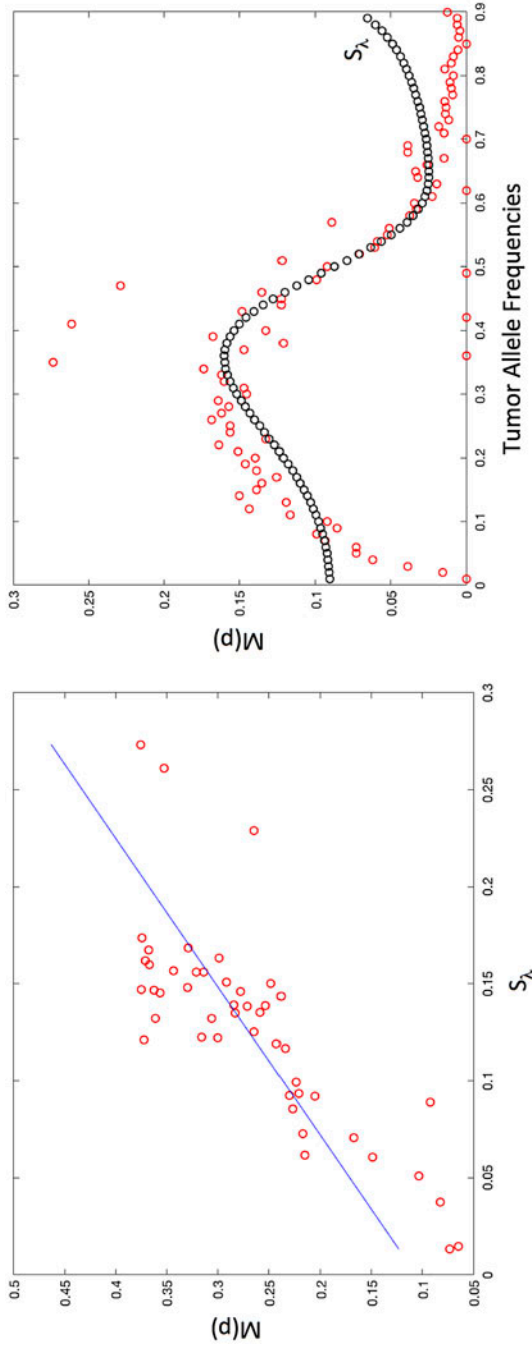


Figure 2. Correlation between s_λ and $M(p)$.

equations are consistent with known observation under low frequencies, Equation (9) suggests that at low frequencies of the dominant alleles, selection against the recessive alleles acts in a power-law-like manner, demonstrating why positive (natural) selection is a potent force. It is worth noting that this power-law growth of relative fitness has been observed in long-term evolution experiments (Lenski, 2015; Wisler, Ribeck, & Lenski, 2013). Under complete dominance, selection would maximize the dominant alleles at all costs.

Equations (8) and (9) also allow us to compute the rate of change of gene/mutation frequencies with respect to time. If α and β are the rate of growth of dominant and recessive alleles, respectively, we can expect an exponential growth. Therefore,

$$\frac{dp}{dt} = \alpha p; \frac{dq}{dt} = \beta q.$$

Hence, denoting $M(q)$ by M_q and $M(p)$ by M_p , we see

$$\frac{dM_q}{dt} = \frac{dM_q}{dq} \cdot \frac{dq}{dt} = s_\lambda \beta p q, \quad (10)$$

and

$$\frac{dM_p}{dt} = \frac{dM_p}{dp} \cdot \frac{dp}{dt} = -s_\theta \alpha q^2. \quad (11)$$

Thus, if the genes/mutations don't directly depend on time, the total derivative, i.e. selection acting over time, is given by

$$\frac{dM}{dt} = s_\lambda \beta p q - s_\theta \alpha q^2.$$

Further, Equations (10) and (11), help us write the general equation for evolution through natural selection for diploid genomes under complete dominance

$$\alpha^2 s_\theta dq dM_q + \beta^2 s_\lambda dp dM_p = 0.$$

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 tumours by taking the degree of selection into account will lead to better understanding of this complex evolutionary disease.

Note

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

Acknowledgements

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

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7, 248–249.
- Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., ... Schultz, N. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2, 401–404. doi:10.1158/2159-8290.CD-12-0095
- Eyre-Walker, A., Woolfit, M., & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173, 891–900.
- Falconer, D. S. (1960). *Introduction to quantitative genetics*. New York, NY: The Roland Press Company.
- Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., ... Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, 6, p11. doi:10.1126/scisignal.2004088
- Gillespie, J. H. (1994). *The causes of molecular evolution*. Oxford Series in Ecology and Evolution. New York, NY: Oxford University Press.
- Lenski, R. E., Wisner, M. J., Ribeck, N., Blount, Z. D., Nahum, J. R., Morris, J. J., ... Hajela, N. (2015). Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *Proceedings of the Royal Society B*, 282, doi:10.1098/rspb.2015.2292
- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31, 3812–3814.
- The UniProt Consortium. (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, 43, D204–D212.
- Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., & Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nature Genetics*, 48, 238–244.
- Wisner, M. J., Ribeck, N., & Lenski, R. E. (2013). Long-term dynamics of adaptation in asexual populations. *Science*, 342, 1364–1367.