BioRxiv-v2

Δ-d_N/d_S: A New Criteria to Distinguish among Different Selection Modes in Gene Evolution

Xun Gu[1*]

[1]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011, USA.

*Corresponding author: xgu@iastate.edu

Abstract

One of the most widely-used measures for protein evolution is the ratio of nonsynonymous distance (d_N) to synonymous distance (d_S). Under the assumption that synonymous substitutions in the coding region are selectively neutral, the d_N/d_S ratio can be used to test the adaptive evolution if $d_N/d_S>1$ statistically significantly. However, due to selective constraints imposed on amino acid sites, most encoding genes demonstrate $d_N/d_S<1$. As a result, d_N/d_S of a gene is less than 1, even some sites may have experienced positive selections. In this paper, we develop a new criterion, called Δ-d_N/d_S, for positive selection testing by introducing an index H, which is a relative measure of rate variation among sites. Under the context of strong purifying selection at some amino acid sites, our model predicts $d_N/d_S=1-H$ for the neutral evolution, $d_N/d_S<1-H$ for the nearly-neutral selection, and $d_N/d_S>1-H$ for the adaptive evolution. The potential of this new method for resolving the neutral-adaptive debates has been illustrated by case studies. For over 4000 vertebrate genes, virtually all of them showed $d_N/d_S<1-H$, indicating the dominant role of the nearly-neutral selection in molecular evolution. Moreover, we calculated the d_N/d_S ratio for cancer somatic mutations of a human gene, specifically denoted by C_N/C_S. For over 4000 human genes in cancer genomics, about 55% of genes showed $1-H<C_N/C_S<1$, about 45% of genes showed $C_N/C_S<1$, whereas less than 1% of genes showed $C_N/C_S<1-H$. Together our analysis suggested driver mutations, i.e., those initiate and facilitate carcinogenesis, confer a selective advantage on cancer cells, leading to $C_N/C_S>1$ (strong positive selection) or $1-H<C_N/C_S<1$ (weak positive selection, combined with strong purifying selection), whereas nearly neutral selection due to reduced effective clonal size is highly unlikely in cancer evolution.

Introduction

Since the proposal of neutral theory (Kimura 1968; King and Jukes 1969) as well as the more general nearly-neutral theory (Ohta 1973; Kimura 1983), the debate between selectionists and neutralists has continued to the era of genome sciences (Lynch 2007; Hahn 2008; Fay et al. 2011; Lynch et al. 2016; Kern and Hahn 2018; Jensen et al. 2019). This is partly because detection of positive selection at the genome level remains a challenge in germline mutations (Bustamante et al. 2005; Sabeti et al. 2006) or in somatic mutations in cancers (Chen et al. 2019a). One of most widely-used measure for protein sequence evolution is the ratio of nonsynonymous distance (d_N) to synonymous distance (d_S) (Li 1997). The d_N/d_S ratio has been widely used to test the underlying selection mode. Under the assumption that synonymous substitutions in the coding region are selectively neutral, the d_N/d_S ratio is a conventional estimate for the ratio of mean evolutionary $\bar{\lambda}$ to the mutation rate (v) (Kimura 1983). It appears that the ratio $d_N/d_S=1$ means a strict neutral evolution in nonsynonymous substitutions; $d_N/d_S>1$ for positive selection and $d_N/d_S<1$ for negative selection.

However, the null hypothesis of $d_N/d_S=1$ (strictly neutrality) can be confounded by some amino acid sites subject to very strong purifying selection (lethal mutations) (Ohta 1992). Therefore, without knowing the proportion of lethal mutations to make some corrections, the d_N/d_S ratio of a gene could be less than one even when some sites are indeed subject to positive selection. In other words, the application of $d_N/d_S>1$ as a criterion to detect positive selection becomes less powerful. Indeed, the observation of $d_N/d_S<1$ can be explained by either the existence of essential sites plus some sites under (weak) positive selection, or the existence of essential sites in the dominance of strictly neutral evolution, or the nearly nearly-neutral evolution. In this paper, we propose a new criterion to distinguish between these possibilities.

New Methods

Evolutionary rate of an encoding-gene

In the theory of molecular evolution, the evolutionary rate (λ) of a nucleotide is determined by three levels of genetic processes: mutation at the individual level, polymorphism at the population level, and fixation at the species level (Kimura 1983). Let v be the mutation rate, s the coefficient of selection and N_e the effective population size. It has been showed that the evolutionary rate is given by

$$\lambda = v\frac{4N_e s}{1-e^{-4N_e s}} \qquad (1)$$

Eq.(1) has provided a solid foundation for resolving the debate between adaptive evolution, neutral evolution, and nearly-neutral evolution: it predicts $\lambda/v>1$ for adaptive evolution ($s>0$), $\lambda/v=1$ for neutral evolution ($s=0$), or $\lambda/v<1$ for deleterious evolution ($s<0$), respectively. Usually, $S=4N_e s$ is called the selection intensity that uniquely determines the rate-mutation ratio λ/v.

In practice, the ratio of nonsynonymous distance (d_N) to synonymous distance (d_S) is widely used in molecular evolution. Under the assumption that synonymous substitutions are selectively neutral, the d_N/d_S ratio can be used as a proxy of the λ/v ratio. Based on the d_N/d_S ratio, numerous statistical methods (Li et al. 1985; Nei and Gojobori 1986; Li 1993; Ina 1995; Cameron 1995; Yang 1997; Yang 2006) were developed to test the neutrality: rejection of the null hypothesis $d_N/d_S=1$ indicates a positive selection if $d_N/d_S>1$, or a negative selection if $d_N/d_S<1$. However, the d_N/d_S ratio is the mean estimated from the nucleotide sites of an encoding gene. Because each site may have different selection intensity (S), without loss of generality we assume that S varies among sites according to a distribution $\Phi(S)$. It follows that the d_N/d_S ratio is actually expected to be

$$\frac{d_N}{d_S} \sim \frac{\bar{\lambda}}{v} = \int \left[\frac{S}{1-e^{-S}}\right] \Phi(S)dS \qquad (2)$$

While the criteria for detecting different evolutionary modes remain the same, we have realized that the interpretation of d_N/d_S ratio test depends on the

structure of $\Phi(S)$. For instance, with a few exceptions, it has been observed that $d_N/d_S<1$ holds for encoding genes of all organisms, indicating the role of selective constraints imposed on the protein sequence. There are two common interpretations of $d_N/d_S<1$. One is the nearly-neutral model, that is, the range of S in $\Phi(S)$ is negatively continuous, with a range of $(-\infty, 0)$; the other one is the neutral-lethal model, that is, $\Phi(S)$ follow a simple two-state distribution, $S=0$ (strictly neutral) and $S=-\infty$ (lethal).

The Δ-criterion on protein sequence evolution

It appears that additional information is required to develop a new d_N/d_S-based method that can distinguish between more sophisticated selection modes. Noting that the conventional d_N/d_S tests were actually focused on the mean rate-mutation ($\bar{\lambda}/v$) ratio, we consider the second moment of the evolutionary rate, that is,

$$\frac{\bar{\lambda^2}}{v^2} = \int \left[\frac{S}{1-e^{-S}}\right]^2 \Phi(S)dS \qquad (3)$$

Next, we introduce a new quantity Δ, the difference between the second-moment and the mean of rate-mutation ratio. From Eq.(2) and Eq.(3), Δ can be written as follows

$$\Delta = \frac{\bar{\lambda^2}}{v^2} - \frac{\bar{\lambda}}{v} = \int \left(\frac{S}{1-e^{-S}}\right)\left(\frac{S}{1-e^{-S}} - 1\right)\Phi(S)dS \qquad (4)$$

As shown below, the sign of Δ provides some insights for testing different selection modes.

$\Delta=0$ under the neutral-lethal selection mode

Under the classical neutral model (Kimura 1968; 1983), the distribution $\Phi(S)$ can be constructed as follows: all mutations are classified into two categories: strictly neutral mutations ($S=0$ with a probability of $1-f_L$) or lethal mutations ($S=-\infty$ with the probability of f_L). Since the evolutionary rate is $\lambda=v$ (mutation rate) for strictly neutral mutations and $\lambda=0$ for lethal mutations, respectively, the first (mean) and the second moments of evolutionary rate of a gene under this model is simply given by

$$\begin{aligned} \bar{\lambda} &= (1-f_L)v \\ \bar{\lambda^2} &= (1-f_L)v^2 \end{aligned} \quad (5)$$

respectively. We thus conclude that, under the neutral-lethal selection mode, we have *Δ=0*, regardless of the proportion of lethal mutations.

Δ>0 under the positive selection with neutral-lethal mode

Suppose that all mutations of a gene are classified into three categories: adaptive mutations (S>0 with a probability of f_A), lethal mutations (S=-∞ with a probability of f_L), and neutral mutations (S=0 with a probability of 1-f_A-f_L). Moreover, the (positive) selection intensity of adaptive mutations follows a distribution denoted by $\Phi^+(S)$. It follows that *Δ* can be written as follows

$$\Delta = f_A \int_0^\infty \left(\frac{S}{1-e^{-S}}\right)\left(\frac{S}{1-e^{-S}}-1\right)\Phi^+(S)dS > 0 \quad (6)$$

which is always larger than 0, regardless of the existence of lethal or neutral mutations.

Δ<0 under negative selection with nearly-neutral mode with lethal mutations

Under the nearly-neutral model with lethal mutations, all mutations of a gene are classified into three categories: nearly-neutral mutations (S<0 with a probability of 1-f_L-f_0), lethal mutations (S=-∞ with a probability of f_L), and neutral mutations (S=0 with a probability of f_0). Moreover, the (negative) selection intensity of nearly-neutral mutations follows a distribution denoted by Φ^-(S). With the argument similar to Eq.(6), one can show

$$\Delta = (1-f_0-f_L)\int_{-\infty}^0 \left(\frac{S}{1-e^{-S}}\right)\left(\frac{S}{1-e^{-S}}-1\right)\Phi^-(S)dS < 0 \quad (7)$$

which is always less than 0.

A new d_N/d_S ratio test based on the *Δ* criterion

An interesting question is whether the Δ criterion is related to the conventional d_N/d_S test. To this end, we invokes the H-measure (Gu et al 1995; Gu 2007), which is defined by

$$H = 1 - (\bar{\lambda})^2/\bar{\lambda^2} \tag{8}$$

Ranging from 0 to 1, a high value of H indicates a high degree of rate variation among sites, and *vice versa*. After some rearrangements of Eq.(8), we have

$$\frac{\bar{\lambda^2}}{v^2} = \frac{(\bar{\lambda}/v)^2}{1-H} = \frac{(d_N/d_S)^2}{1-H} \tag{9}$$

where $\bar{\lambda}/v = d_N/d_S$ is applied according to Eq.(2). It follows that the relationship between Δ, H and d_N/d_S is given by

$$\Delta = \frac{d_N}{d_S}\left(\frac{d_N/d_S}{1-H} - 1\right) \tag{10}$$

Hence, the connection between Δ and d_N/d_S is straightforward: $\Delta<0$ indicates $d_N/d_S<1-H$, $\Delta=0$ indicates $d_N/d_S=1-H$ and $\Delta>0$ indicates $d_N/d_S>1-H$, respectively. Accordingly, we formulate a new d_N/d_S ratio test based on the Δ criterion (Table 1)

(i) The null hypothesis is $d_N/d_S=1-H$, expected by the strictly neutral evolution while some sites are subject to very strong purifying selections (invariable sites).

(ii) Rejection of the null hypothesis because of $d_N/d_S>1-H$ suggests a positive selection whereas some functionally important sites are virtually invariable.

(iii) Rejection of the null hypothesis because of $d_N/d_S<1-H$ provides the evidence for the nearly-neutral model instead of the strictly-neutral model with some invariable sites.

In other words, the new method has two features: (*i*) to detect signal of positive whereas some sites may be subject to very strong constraints (invariable sites); and (*ii*) to distinguish between the nearly neutral model and the strictly-neutral model with invariable sites.

Calculation of *H*

The remaining question is how one can estimate the parameter *H*, which is fundamental in the new Δ-d_N/d_S test. In the following we discuss types of datasets.

Multiple-alignment of protein sequences of a gene

Following Gu (2007), we propose a computational based on a protein sequence phylogeny that can be reliably reconstructed. Let $\bar{x}$ and $V(x)$ be the mean and variance of number of changes per amino acid site, respectively. Suppose that amino acid changes at a site follow a Poisson process along the phylogeny, whereas the evolutionary rate (λ) varies among sites. Then one may write $\bar{x}=\bar{\lambda}T$ and $V(x)=V(\lambda)T^2+\bar{\lambda}T$, respectively, where T is the total evolutionary time along the phylogeny. These two equations together directly lead to an estimate of H by

$$\hat{H} = \frac{V(\lambda)}{V(\lambda)+(\bar{\lambda})^2} = \frac{V(x)-\bar{x}}{V(x)+\bar{x}(\bar{x}-1)} \qquad (11)$$

Given the phylogeny, we implement two methods to infer the number of changes per amino acid site: the first one is the parsimony method to infer the minimum-required number of changes per site, and the second one is the method of Gu and Zhang (1997) to estimate the (bias-corrected) number of amino acid changes at each site.

Cancer somatic mutations of a gene

The new Δ-d_N/d_S test, which can be applied to study cancer evolution after some specifications, called the Δ-C_N/C_S test. An interesting question is whether the Δ criterion is related to the conventional d_N/d_S test. Let H be the relative measure for the variation of the rate of somatic mutations among sites. For cancer genomic data, it is straightforward to calculate $\bar{x}$ and $V(x)$, the mean and variance of number of somatic mutations per amino acid site, respectively, which directly lead to an estimate of H by Eq.(11).

Results and Discussion

Data availability

Vertebrate protein sequence data

The protein sequences and homology information of the human and other seven vertebrate genomes (mouse, dog, cow, chicken, Xenopus, fugu, zebrafish) were downloaded from Ensembl EnsMart (http://www.ensembl.org/Multi/martview). Though there are several annotated homology relationships between the human and other genomes in Ensembl, we only considered those pairs of genes annotated as UBRH (Unique Best Reciprocal Hit, meaning that they were unique reciprocal best hits in all-against-all BlastZ searches) to be orthologous. In our study, we used genes that have their orthologs across all eight genomes, i.e. V8 ortholog set. Protein sequences of V8 ortholog set were downloaded from Ensembl; if a gene has multiple protein sequences, usually caused by alternative splicing, only the longest sequence was kept. For each V8 ortholog set, homologous proteins were aligned by the program T-coffee using the default parameters (Notredame et al. 2000). The number of synonymous substitutions per synonymous site (*dS*) and the number of nonsynonymous substitutions per nonsynonymous site (*dN*) between human and mouse orthologs were also retrieved form Ensembl EnsMart, which were estimated by PAML package (Yang 1997) using the likelihood method.

Cancer genomics data

We extracted the cancer somatic mutations of 10,224 cancer donors from TCGA PanCanAtlas MC3 project (https://gdc.cancer.gov/about-data/publications/mc3-2017), which includes 3,517,790 somatic mutations in the coding regions, 2,035,693 of which causing amino acid changes in cancers (missense mutations). After filtering out some redundancies according to the criterion of Bailey *et al.*(2018), this left us with 9,078 samples with 1,461,387 somatic mutations in the coding regions, which consisted of 793,577 missense mutations that were used in our study. Sequences and annotations of human genes were extracted from the Ensembl database (http://www.ensembl.org,

GRCh37, Release 75). For each gene, we chose the transcript which is most commonly used in TCGA.

Next, genes that have no amino acid site with at least three recurrent missense mutations were excluded because of the lack of degrees of freedom for calculating the *H* value; the final dataset has 4,047 genes. The list of 538 known cancer genes was compiled by merging the list of 253 cancer genes with the missense mutation type from the Cancer Gene Census Tier 1 (COSMIC GRCh37, V89)(Tate et al. 2019), 127 significant mutated genes (SMGs) reported by Kandoth, et al. (2013), 260 SMGs reported by Lawrence, et al. (2013) and 299 cancer driver genes reported by Bailey, et al. (2018).

Nearly-neutral selection dominates protein evolution in vertebrates

Briefly, the estimation procedure was as follows. (*i*) Infer the phylogenetic tree from a multiple alignment of homologous protein sequences. Although there is no methodological preference, we require that the inferred tree topology should be roughly the same over methods. (*ii*) Estimate the nonsynonymous to synonymous ratio (d_N/d_S) from closely related coding sequences that satisfy d_N/d_S <1. (*iii*). Estimate the *H*-measure for rate variation among sites according to Eq.(11).

We analyzed three independent datasets (Su et al. 2009): (*a*) 342 vertebrate genes from eight species; (*b*) 400 Drosophila genes from 12 species; and (*c*) 380 yeast genes from five species. Impressively, in each dataset, almost all genes satisfy the condition of d_N/d_S<1-*H*, suggesting that nearly-neutral evolution, rather than neutral evolution plus strong negative selection, explains the pattern of protein evolution (**Table 1**). Further, we analyzed 4336 vertebrate genes, while the d_N/d_S ~ (1-*H*) scatter plotting was shown in **Fig.1**. Almost all of genes are below the line of d_N/d_S=1-*H*, or d_N/d_S<1-*H*. We thus conclude that nearly-neutral model is sufficient to explain the genome-wide pattern of protein evolution in vertebrates.

Evolution of cancer somatic mutations: Dominant positive selection under

strong functional constraints

The theory of clonal evolution in cancer biology claims that cancer cells emerge through random somatic mutations from a single cell and genetically diverge to disparate cell subclones and successive clones in cancer cell replication (Stratton et al. 2009; Vogelstein et al. 2013). Ultimately, cells alter one or few crucial pathways and acquire the hallmarks of cancer through somatic mutations followed by cancer-specific positive selections (Martincorena and Campbell 2015; Bailey et al. 2018). The argument that carcinogenesis is a form of evolution at the level of somatic cells suggests that our understanding of cancer initiation and progression can be benefited by the molecular evolutionary approaches. One well-known example is to estimate the rate ratio of somatic nonsynonymous to synonymous substitutions (C_N/C_S) of a protein-encoding gene in cancers, but resulted in inconsistent conclusions (Dees et al. 2012; Lawrence et al. 2013; Reimand and Bader 2013; Schroeder et al. 2014; Porta-Pardo et al. 2014; Mularoni et al. 2016; Martincorena et al. 2017; Zhou et al. 2017; Weghorn and Sunyaev 2017).

Thanks to the advances in cancer genomics, The Cancer Genome Atlas (TCGA) has accumulated millions cancer somatic mutations from nearly ten thousand tumor-normal pairs. We applied the new Δ-C_N/C_S analysis to explore the major pattern of selection modes in cancer evolution, where C_N/C_S of each gene is calculated by the method of Zhou et al. (2917). First we consider 294 cancer genes. About 70% of them showed C_N/C_S >1, about 30% of them showed 1-H<C_N/C_S <1, whereas only one gene showed C_N/C_S<1-H (Table 2). Together we propose the following scenario. Driver mutations, i.e., those initiate and facilitate carcinogenesis, confer a selective advantage on cancer cells (positive selection), leading to C_N/C_S>1 for the majority of cancer-related genes. Impressively, up to 30% cancer genes revealed the pattern of mutations under weak positive selection, and mutations under strong purifying selection that were ultimately eliminated from the cancer cell population (Zhou et al. 2017; Weghorn and Sunyaev 2017). Moreover, our analysis suggests that passenger somatic mutations, those unrelated to the process of carcinogenesis, are likely be selectively neutral, rather than

nearly neutral due to reduced effective clonal size (Kandoth et al. 2013), because $C_N/C_S<1-H$ has been shown highly unlikely. Our analysis in cancer genes has been confirmed by a genome-wide analysis (**Fig.2**).

Further extensions of the Δ-d_N/d_S test

We anticipate that the new criterion as the null hypothesis of neutrality with strong functional constraints ($d_N/d_S=1-H$) may have many potential applications as it can be implemented into a variety of existed d_N/d_S-related analyses. The challenge is how to estimate the H measure accurately when the observed variation of nonsynonymous mutations or substitutions is not statistically sufficient. While it is ideal to estimate d_N/d_S and H simultaneously, we found that the mathematical modeling can be sophisticated, leading to computationally difficulty in practice. Technically, one may treat H as a known constant in the statistical analysis. In the following we discuss two examples briefly.

Δ-d_N/d_S test in the relative rate analysis. Consider the related rate (fast or slow) between two species 1 and 2, given the third species as outgroup. When a relative rate test shows a faster molecular evolution in lineage-1 than lineage-2, the challenge is to determine whether fast-evolution in lineage-1 is driven by adaptive evolution. Let $[d_N]_k$, $[d_S]_k$ and $[d_N/d_S]_k$ be the nonsynonymous distance, synonymous distance and the ratio in the k-th lineage, respectively, k=1, 2 and 3, respectively. Under the Δ-d_N/d_S test, the null hypothesis is that lineage-specific fast-evolution follows a strictly neutral evolution whereas some sites are under strong functional constraints, which predicts $[d_N/d_S]_1=1-H$, or $[d_N]_1=(1-H)[d_S]_1$. Meanwhile, in lineage-2 the d_N/d_S ratio represents the ancestral ratio of the protein sequence conservation, that is, $[d_N/d_S]_2=[d_N/d_S]$ in general. Therefore, one can establish the connection between Δ-d_N/d_S and the fast-evolving ratio $R_{12}=[d_N]_1/[d_N]_2$ under the assumption of $[d_S]_1=[d_S]_2$ (synonymous rates are the same between lineages), namely,

$$R_{12}=(1-H)/(d_N/d_S) \qquad (12)$$

One can develop a statistical procedure to test the null hypothesis of Eq.(12). Rejection of the null hypothesis may indicate a positive selection if $[d_N]_1/[d_N]_2>R_{12}$, or indicate either functional relaxation or reduced effective population size if $[d_N]_1/[d_N]_2<R_{12}$. In the case when the two species are not closely related so that the estimation of d_S could be subject to a large sampling variance, one may use the estimate of d_N/d_S from some reliable reference species.

The radical/conservative replacement rate ratio. A major variety of d_N/d_S-related analysis is the D_r/D_c ratio test, the ratio of radical replacement rate (D_r) to conservative replacement rate (D_c) (Dagan et al. 2002; Smith 2003; Hanada et al. 2007; Weber et al. 2014; Figuet et al. 2016; Braun 2018; Chen et al. 2019b). The D_r/D_c ratio analysis postulates that, under positive selection more radical amino acid replacements may occur than conservative replacements, and *vice versa* under negative selection. If we view the D_r/D_c ratio as a proxy to the rate-mutation ration, our Δ-d_N/d_S analysis can be applied to the D_r/D_c with some technical modifications.

Table 1. A summary of Δ-d_N/d_S analysis

Criteria	Interpretations
$d_N/d_S<1-H$	Nearly-neutral evolution, plus some sites under strong selective constraints
$d_N/d_S=1-H$	Neutral evolution, plus some sites under strong selective constraints
$1>d_N/d_S>1-H$	Positive evolution, plus some sites under strong selective constraints
$d_N/d_S=1$	Non-functionalization (being a psedogene without any site under selective constraints)
$d_N/d_S>1$	Dominant positive evolution

Table 2. Constrting patterns between vertebrate protein sequence evolution and the evolution of cancer somatic mutations.

Vertebrate protein sequence alignments			
	$d_N/d_S<1-H$	$1-H<d_N/d_S<1$	$d_N/d_S>1$
	4335	1	0
Cancer genomic data			
	$C_N/C_S<1-H$	$1-H<C_N/C_S<1$	$C_N/C_S>1$
Cancer-related	1 (0.3%)	87 (29.6%)	206 (70.1%)
Genome-wide	30 (0.7%)	2245 (55.5%)	1772 (43.8%)

References

M. H. Bailey, Collin Tokheim, Eduard Porta-Pardo, ..., Gordon B. Mills, Rachel Karchin, Li D. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018;173:371–385.e18. doi:10.1016/j.cell.2018.02.060.

Braun EL. 2018. An evolutionary model motivated by physicochemical properties of amino acids reveals variation among proteins. Bioinformatics 34(13):i350–i356.

Bustamante CD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437(7062):1153–1157.

Chen B, Shi Z, et al. . 2019a. Tumorigenesis as the paradigm of quasi-neutral molecular evolution. Mol Biol Evol. 36(7):1430–1441.

Chen Q, He Z, et al. . 2019b. Molecular evolution in large steps—codon substitutions under positive selection. Mol Biol Evol. 36(9):1862–1873.

Comeron JM. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol. 41:1152–1159.

Dagan T, Talmor Y, Graur D. 2002. Ratios of radical to conservative amino acid replacement are affected by mutational and compositional factors and may not be indicative of positive Darwinian selection. Mol Biol Evol. 19(7):1022–1025.

Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Res 2012;22:1589–98. doi:10.1101/gr.134635.111.

Fay, J. C. 2011. Weighing the evidence for adaptation at the molecular level. Trends Genet. 27(9):343–9.

Figuet E, et al. . 2016. Life history traits, protein evolution, and the nearly neutral theory in amniotes. Mol Biol Evol. 33(6):1517–1527.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83. doi:10.1038/nrc1299.

Hahn, M. W. 2008. Toward a selection theory of molecular evolution. Evolution 62:255–65.

Hanada K, Shiu SH, Li WH. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. Mol Biol Evol. 24(10):2235–2241.

Ina Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol. 40(2):190–226.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502:333–9. doi:10.1038/nature12634.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624–626.
Kern, A. D. and M. W. Hahn. 2018. The neutral theory in light of natural selection. Mol. Biol. Evol. 35:1366–71

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press. p. 103–103.

King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. Science 164: 788–98.

Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8. doi:10.1038/nature12213.

Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501. doi:10.1038/nature12912.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol. 36(1):96–99.

Li, W-H Molecular evolution. (Sinauer associates incorporated, 1997).

Li WH, Wu CI, Luo CC. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2(2):150–174.

Lynch, M. 2007. The origins of genome architecture. Sinauer, Sunderland, MA.

Lynch, M., M. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K. Thomas, and P. L. Foster. 2016. Genetic drift, selection, and the evolution of the mutation rate. Nat. Rev. Genet. 17:704–714.

Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science 2015;349:961–8. doi:10.1126/science.aab4082.

Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biol 2016;17:1–13. doi:10.1186/s13059-016-0994-0.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5):418–426.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–8.

Ohta T (1992) The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics 23(1):263-286.

Porta-Pardo E, Godzik A. e-Driver: A novel method to identify protein regions driving cancer. Bioinformatics 2014;30:3109–14. doi:10.1093/bioinformatics/btu499.

Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol 2013;9. doi:10.1038/msb.2012.68.

Smith NG. 2003. Are radical and conservative substitution rates useful statistics in molecular evolution? J Mol Evol. 57(4):467–478.

Weber CC, Nabholz B, Romiguier J, Ellegren H. 2014. Kr/Kc but not dN/dS correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biol. 15(12):542.

Weghorn D, Sunyaev S. Bayesian inference of negative and positive selection in human cancers. Nat Genet 2017;49:1785–8. doi:10.1038/ng.3987.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D. and Lander, E.S. (2006) Positive Natural Selection in the Human Lineage, *Science*, **312**, 1614-1620.

Schroeder MP, Rubio-Perez C, Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. Bioinformatics 2014;30. doi:10.1093/bioinformatics/btu467.

Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719–24. doi:10.1038/nature07943.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 2019;47:D941–7. doi:10.1093/nar/gky1015.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science 2013;340:1546–58. doi:10.1126/science.1235122

Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13, 555–556 (1997).

Yang Z. 2006. Computational Molecular Evolution. Oxford: Oxford University Press. p. 48–49.

Zhou Z, Zou Y, Liu G, Zhou J, Wu J, Zhao S, et al. Mutation-profile-based methods for understanding selection forces in cancer somatic mutations: a comparative analysis. Oncotarget 2017;8:58835–46. doi:10.18632/oncotarget.19371

Fig.1. The d_N/d_S ~ (1-H) scatter plotting of 433 vertebrate genes, showing that almost all of them are below the line of d_N/d_S=1-H, or d_N/d_S<1-H.

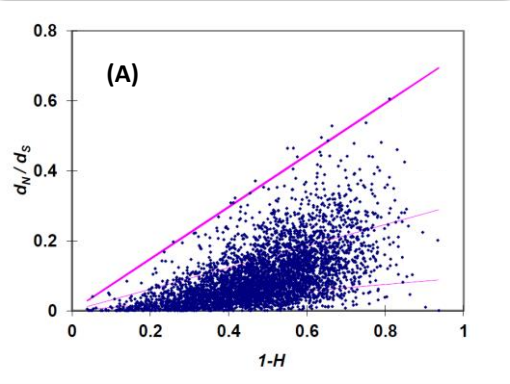
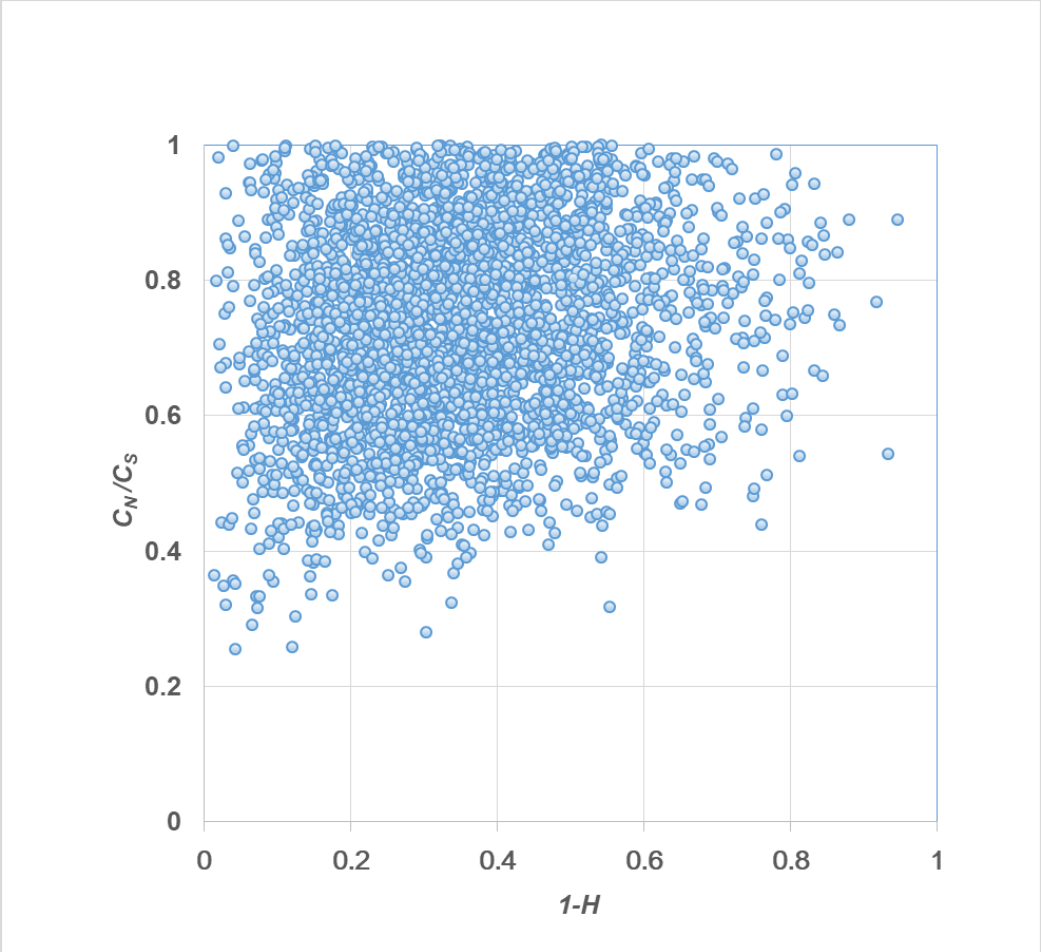
Fig.2. The C_N/C_S ~ (1-H) scatter plotting of 2275 human genes with C_N/C_S<1, showing that most of them are above the line of C_N/C_S=1-H, or C_N/C_S>1-H.