PLoS one

The Trouble with Sliding Windows and the Selective Pressure in BRCA1

Karl Schmid[1], Ziheng Yang[2]*

1 Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences (SLU), Uppsala, Sweden, **2** Department of Biology, University College London, London, United States of America

Abstract

Sliding-window analysis has widely been used to uncover synonymous (silent, d_S) and nonsynonymous (replacement, d_N) rate variation along the protein sequence and to detect regions of a protein under selective constraint (indicated by $d_N < d_S$) or positive selection (indicated by $d_N > d_S$). The approach compares two or more protein-coding genes and plots estimates $\hat{d}_S$ and $\hat{d}_N$ from each sliding window along the sequence. Here we demonstrate that the approach produces artifactual trends of synonymous and nonsynonymous rate variation, with greater variation in $\hat{d}_S$ than in $\hat{d}_N$. Such trends are generated even if the true d_S and d_N are constant along the whole protein and different codons are evolving independently. Many published tests of negative and positive selection using sliding windows that we have examined appear to be invalid because they fail to correct for multiple testing. Instead, likelihood ratio tests provide a more rigorous framework for detecting signals of natural selection affecting protein evolution. We demonstrate that a previous finding that a particular region of the BRCA1 gene experienced a synonymous rate reduction driven by purifying selection is likely an artifact of the sliding window analysis. We evaluate various sliding-window analyses in molecular evolution, population genetics, and comparative genomics, and argue that the approach is not generally valid if it is not known *a priori* that a trend exists and if no correction for multiple testing is applied.

Introduction

Sliding-window analysis is a popular graphical method for visually revealing trends in synonymous and nonsynonymous rate variation along a protein sequence, and for identifying protein regions that are under functional constraint or positive selection [e.g., 1,2–5]. It is implemented in several computer programs and web servers [e.g., 6,7,8]. Because of its simplicity and intuitive appeal, its legitimacy in such analyses was most often taken for granted.

When applying the approach to compare various gene sequences, we noted two features of the analysis: (i) the estimated number of synonymous substitutions per synonymous site ($\hat{d}_S$) and the number of nonsynonymous substitutions per nonsynonymous site ($\hat{d}_N$) always showed clear trends along the protein sequence, and (ii) $\hat{d}_S$ was more variable than $\hat{d}_N$ along the gene sequence. The greater variation of $\hat{d}_S$ than of $\hat{d}_N$ is particularly surprising. Because processes operating at the DNA level, such as local mutation rate variation [9], should affect both d_S and d_N [10: p. 65] while natural selection on the protein should affects d_N but not d_S, and because protein-level selection is expected to vary across amino acid sites or protein domains, we expect d_N to be more variable than d_S [see also 3]. For d_N to be less variable than d_S, variation in selective constraint on the protein will have to counterbalance variation in mutation rate. Such a scenario appears to be too contrived to apply to many genes. Further examinations, however, suggest that the apparent trends in $\hat{d}_S$ and $\hat{d}_N$ revealed by sliding-window

analysis do not reflect variations in the true d_S and d_N, and are an artifact of the procedure. The effect is inherent in the method and affects many applications of sliding-window analysis.

Here we demonstrate the artifactual effect of sliding-window analysis through a re-analysis of the breast-cancer gene BRCA1 from mammalian species. We also discusses similar problems when sliding-window analysis is used in several other applications in molecular evolutionary studies.

Results

Sliding-window analysis of mammalian BRCA1 genes

The breast-cancer gene BRCA1 is a well-known empirical case of synonymous rate variation, since Hurst and Pál [3,9] conducted a sliding-window analysis to compare the human with the dog and the mouse with the rat genes. Here we reanalyze the data to show that the apparent synonymous rate variation and the purifying selection acting on silent sites in a particular region inferred by those authors is likely an artifact. We follow the common practice of conducting sliding-window analysis in pairwise sequence comparisons but note that our conclusions apply also to simultaneous comparison of multiple sequences. Besides the mouse-rat and human-dog pairs, we also use the orangutan-cow and orangutan-macaque pairs.
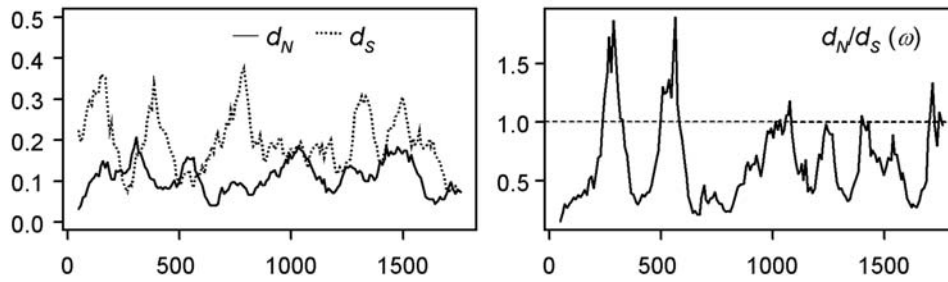
The results are presented in Figure 1. The window size is set to 100 codons, with an offset of one codon between successive windows. In each window, the ω ratio ($= d_N/d_S$) as well as d_S and

d_N were estimated using maximum likelihood (ML) under model M0 (one-ratio), which assumes that the same ω ratio applies to all codons in the gene [11]. While the method for estimating d_S and d_N may be important, the effects we demonstrate do not depend on the estimation method; use of the approximate methods such as YN00 [12] produced qualitatively identical results (not shown). From Figure 1, the following patterns are apparent: (i) both $\hat{d}_S$ and $\hat{d}_N$ show smooth trends of fluctuation along the sequence; (ii) $\hat{d}_S$
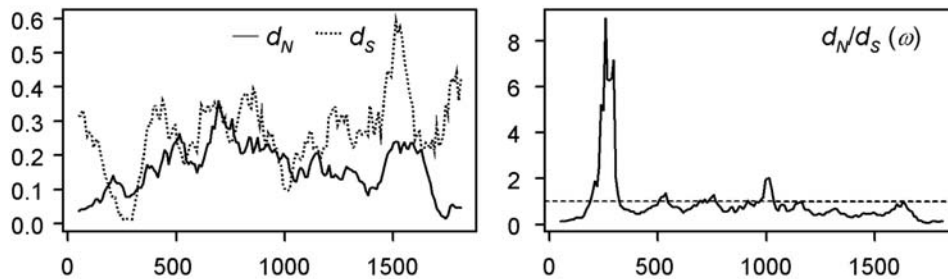
fluctuates more wildly along the sequence than $\hat{d}_N$; and (iii) in some regions, the estimated rate ratio $\hat{\omega} > 1$, which could naïvely be interpreted as indicating positive selection.

As discussed by Hurst and Pál [3], there is a striking plummet in $\hat{d}_S$ around codon 250 in the comparisons between the mouse and the rat and between the human and the dog (Figure 1A&B). Hurst and Pál referred to this region as the 'critical region' and their test suggested that the ω ratio was significantly greater than 1 in the
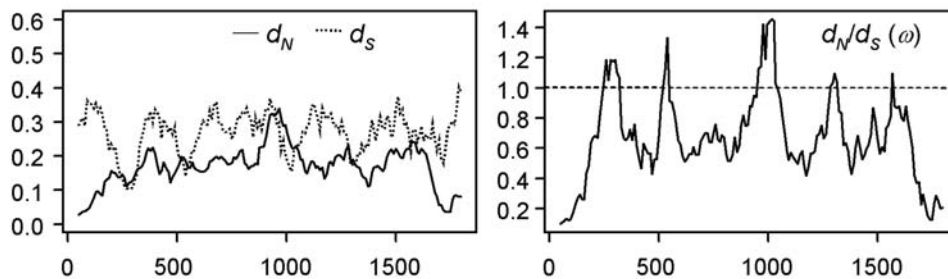
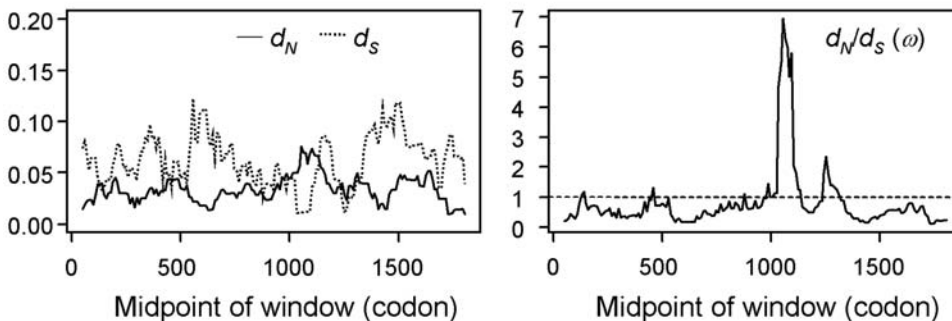Figure 1. Sliding-window plots of $\hat{d}_S$, $\hat{d}_N$ and $\hat{\omega} = \hat{d}_N/\hat{d}_S$ in pairwise comparisons of the BRCA1 genes from mammalian species. The window size is 100 codons, and the offset between windows is one codon.
doi:10.1371/journal.pone.0003746.g001

human-dog pair and significantly higher than the average for the whole gene in the mouse-rat pair. The authors suggested purifying selection at silent sites as the most likely mechanism for the reduced $\hat{d}_S$ and for the elevated $\hat{\omega}$ for the region. Nevertheless, the authors' tests do not appear to be valid, because the 'critical region' was identified by analyzing the data and not specified *a priori*, and because no correction for multiple testing was applied (see below). The orangutan-macaque comparison (Figure 1D) is largely independent phylogenetically of the mouse-rat and human-dog comparisons, and does not show a dip in $\hat{d}_S$ in the critical region. The orangutan-cow comparison (Figure 1C) overlaps somewhat with the human-dog comparison, and shows a small dip in $\hat{d}_S$ in the critical region, but is by no means out of the ordinary. It is noteworthy that even between the mouse-rat and human-dog comparisons, the peaks and valleys in $\hat{d}_S$ and $\hat{d}_N$ do not occur at similar locations except for the dip in $\hat{d}_S$ in the critical region.

Sliding-window analysis of simulated data

To examine whether the patterns of Figure 1 are statistically significant and may thus reflect real biological processes, we apply the sliding-window analysis to data sets simulated under model M0 (one-ratio), which assumes the same d_S, d_N, and ω across the whole sequence and independent evolution among codons. The ML estimates of parameters under M0 from the original pair of real sequences [11] were used to simulate replicate data sets using program EVOLVER in the PAML package [13]. The results obtained from simulations based on the four pairs of sequences are qualitatively similar, so we present in Figure 2 only those for the first two replicate data sets based on the mouse-rat comparison. The original parameter estimates for this pair are $\hat{t} = 0.391$, $\hat{\kappa} = 3.304$, and $\hat{\omega} = 0.504$, with $\hat{d}_S = 0.204$ and $\hat{d}_N = 0.103$.

Simply from visual inspection, we were unable to distinguish the plots in Figure 1A for the real data from those in Figure 2A&B for the simulated data. The peaks and valleys in $\hat{d}_S$ and $\hat{d}_N$ in Figure 2 are random and differ between simulated replicates. However, like the real data, the simulated data show considerable and smooth fluctuations in $\hat{d}_S$ and $\hat{d}_N$, greater fluctuations in $\hat{d}_S$ than in $\hat{d}_N$, and also windows with $\hat{\omega} > 1$. All those features are artifactual.

We suggest that the following reasons may explain the features. First, $\hat{d}_S$ and $\hat{d}_N$ calculated from the sliding windows will fluctuate due to chance effects in a small window. Because two neighboring windows share many codons, $\hat{d}_S$ and $\hat{d}_N$ will change smoothly when plotted against the sequence. Of course the amount of smoothing depends on the window size and the offset between consecutive windows. Second, the fluctuations in $\hat{d}_S$ and $\hat{d}_N$ are due to fluctuations in the estimated numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions and in the numbers of synonymous (S) and nonsynonymous (N) sites. Consider the numbers of sites S and N in a window. Their sum is $3w$, where w is the number of codons in each window. Random fluctuations in amino acid composition or codon usage will generate fluctuations in S and N. Because N is about three times as large as S, the same amount of change will proportionally affect S much more than it affects N. As a result, $\hat{d}_S$ tends to fluctuate more than $\hat{d}_N$. Because the data of Figure 2 are generated under model M0 (one-ratio), with constant d_S and d_N along the sequence and with independent evolution among codons, the apparent variation and trends in $\hat{d}_S$ and $\hat{d}_N$ are artifacts.

Multiple testing in sliding-window analysis and likelihood ratio test of positive selection

We examined the validity of previous uses of sliding-window analysis to test for regions of a protein under selective constraint or

A Replicate 1

B Replicate 2

Midpoint of window (codon)

Midpoint of window (codon)
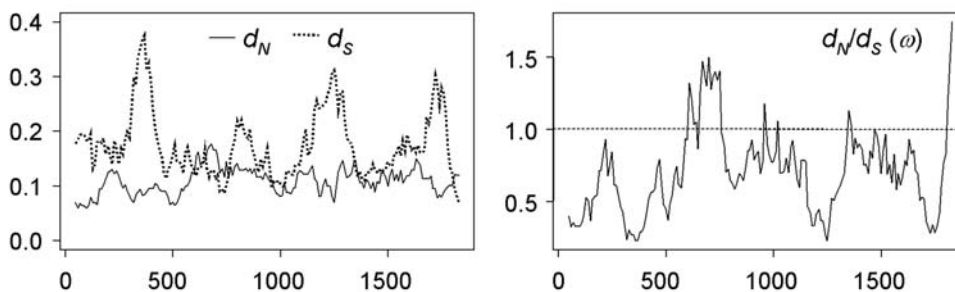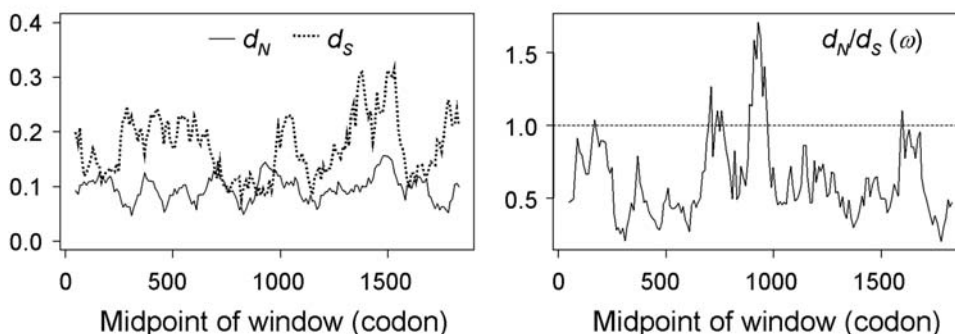
Figure 2. Sliding window plots of $\hat{d}_S$, $\hat{d}_N$ and $\hat{\omega} = \hat{d}_N/\hat{d}_S$ from two simulated data sets, generated under model M0 (one-ratio) using parameter estimates obtained from the comparison of the mouse and rat BRCA1 genes. The window size is 100 codons, and the offset between windows is one codon.
doi:10.1371/journal.pone.0003746.g002

positive selection [e.g., 1,4,7,8,14]. Most such studies used simplistic methods to estimate d_S and d_N, ignoring major features of DNA sequence evolution such as unequal codon frequencies or different transition and transversion rates. Here we claim that most such tests we have examined appear to be invalid, partly because they did not correct for multiple testing. If one conducts 100 independent tests at the 5% significance level, one is expected on average to reject falsely the null hypothesis by chance in 5 tests. Here the tests are not independent because the windows overlap, but the problem of multiple testing exists. The overall false-positive rate or the family-wise error rate refers to rejection of at least one true null hypothesis when multiple null hypotheses are tested. This error rate can be much higher than the significance level if no correction for multiple testing is applied.

Figure 3A shows the relationship between the overall false-positive rate and the size of the sliding window, when a pair of sequences is simulated under a model of no positive selection and then analyzed using sliding windows to test for positive selection. We used two null models to simulate the data. The first is model M0 (one-ratio) with the single ω ratio fixed at 1. The second is the site model M1a (neutral), which assumes two site classes with $\omega_0 = 0$ and $\omega_1 = 1$, in proportions $p_0 = p_1 = \frac{1}{2}$. Each simulated data set is analyzed using a sliding window, using an LRT to test whether $\hat{\omega}$ for that window is significantly greater than 1. A false positive is recorded if the test is significant in at least one window. The error rate rises quickly with the increase of the window size, peaks at an intermediate window size of between 5 and 10 codons, and then drops with the further increase of the window size. The false positive rates are unacceptably high at low and intermediate window sizes. Note that in datasets simulated under M1a, the overall error rate is nearly zero in large windows, because the test based on M0 (one-ratio), which requires the average ω ratio for the whole sequence to be >1, is very stringent. The effect of the offset is examined in Figure 3B, which shows that for a fixed window size (20 codons), the error rate drops when the offset increases.

A simulation approach may be used to correct for multiple testing. One may use the number of windows in which $\hat{\omega} > 1$ as the test statistic; let this be W. The null distribution can be generated by simulating under a null model of no positive selection. An appropriate null model is the site model M1a (neutral), which assumes two site classes with $\omega_0 < 1$ and $\omega_1 = 1$ [15]. We applied this test to the four pairs of BRCA1 genes. For each pair, we calculated the test statistic from the original data, W. The original data were then used to estimate parameters under M1a (neutral), and the estimates were used to simulate 1000 datasets under M1a. Each dataset i was then analyzed using a sliding window to calculate the number of windows in which $\hat{\omega}$ under M0 is >1, $W^{(i)}$. The p value is the proportion of simulated datasets in which $W^{(i)} \geq W$. We used the window size of 100 codons, with an offset of 10 codons to analyze the BRCA1 genes. The results are shown in table 1. The test is significant in the human-dog ($p < 1\%$) and orangutan-cow ($p < 5\%$) pairs, but not in the mouse-rat and orangutan-macaque pairs.

For comparison, we applied two likelihood ratio tests of positive selection to the same data, comparing the site models M1a (neutral) against M2a (selection) and M7 (beta) against M8 (beta&ω). Both LRTs are significant in the human-dog comparison and not significant in all other pairs (table 1). We note that the test based on sliding windows is a goodness of fit test, although the test statistic is designed such that rejection of the null indicates positive selection.

Previous simulation studies suggest that the LRTs based on site models may be more sensitive when multiple sequences are compared jointly on a phylogenetic tree [16], so we applied the LRTs to the dataset of nine mammalian species. The phylogeny is shown in Figure 4. The results are summarized in table 2. M0 (one-ratio) has much lower log likelihood than the site models which allow ω to vary among sites, indicating highly variable selective pressure along the protein. The M1a-M2a test gave $2\Delta\ell = 1.3$, and the difference was not significant. While the parameter estimates under M2a suggested a small proportion of sites with $\omega > 1$, the BEB calculation [17] detected no sites with high posterior probability of being under positive selection ($P < 0.6$ for all sites). The M7-M8 comparison is significant, with
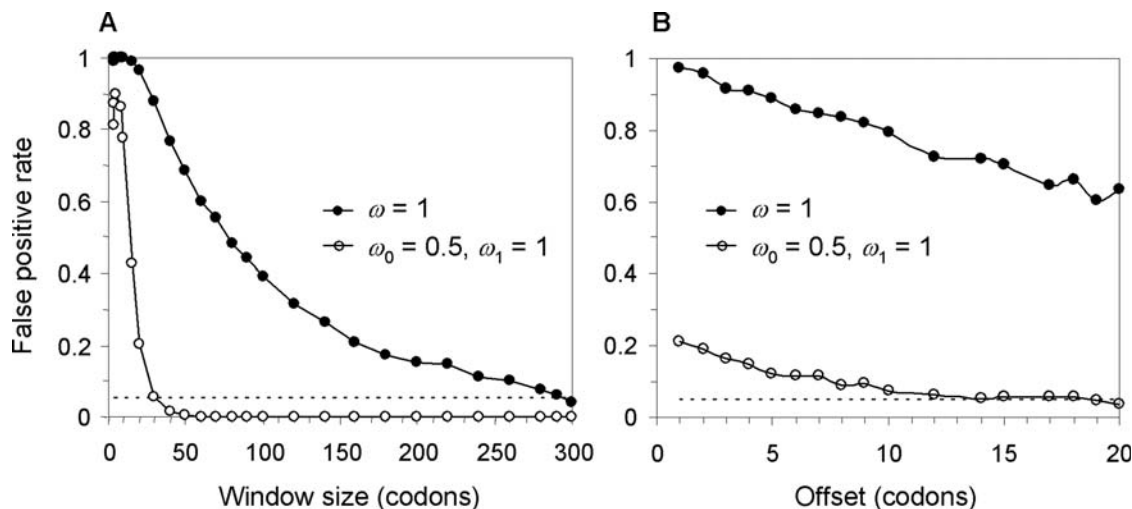
Figure 3. The overall false-positive rate of the sliding-window test of positive selection plotted against the window size. Data of a pair of sequences are simulated under either model M0 (one-ratio) with $\omega = 1$ (•) or model M1a (neutral) with two site classes in proportions $p_0 = p_1 = \frac{1}{2}$ with $\omega_0 = 0$ and $\omega_1 = 1$ (○). An LRT is used to test for positive selection in each window, by fitting model M0 (one-ratio) to the data, either with $\omega \geq 1$ estimated or with $\omega = 1$ fixed, and by comparing twice the log likelihood difference between the two hypotheses with 2.71, at the 5% level. If the test is significant in any window, positive selection is claimed to be detected for the replicate data set. The false-positive rate is calculated as the proportion of replicate datasets in which the test is significant in at least one window. The sequence length is 300 codons. The impact of the window size is examined in **A**, with the offset fixed at one codon, while the impact of the offset is examined in **B**, with the window size fixed at 20.
doi:10.1371/journal.pone.0003746.g003

Table 1. Test of sites under positive selection by the sliding-windows analysis and by the LRT.

Data	Sliding windows		$2\Delta\ell$	$2\Delta\ell$
	Statistic [a]	*p*-value	(M1a-M2a)	(M7-M8)
Mouse-rat	19	0.096	2.24	2.39
Human-dog	18	0.007**	8.18*	9.08*
Orangutan-cow	23	0.016*	0.35	0.40
Orangutan-macaque	16	0.762	0.00	0.00

[a]The test statistic in the sliding window analysis is the number of windows in which $\hat{\omega} > 1$. The window size is 100 codons, and the offset is 10 codons. In the LRTs, the test statistic $2\Delta\ell$ is the log likelihood difference between the null and alternative models.
*: significance with $p < 5\%$.
**: significance with $p < 1\%$.
doi:10.1371/journal.pone.0003746.t001

$2\Delta\ell = 11.32$. The BEB calculation suggested three sites (897N, 914N, 919I) to be potentially under positive selection, with $0.80 < P < 0.86$. Thus both models M2a and M8 provide some evidence for presence of sites under positive selection, but the disagreement between the two tests and the lower posterior probabilities for sites indicate that the evidence is not strong.

A plausible explanation for the conflicts between the joint analysis and the pairwise tests is that the selective pressure on the protein has been variable among lineages, and the various tests used here either average over sites or average across lineages, leading to somewhat inconsistent results. Previously Huttley *et al.* [2] detected positive selection in BRCA1 affecting the human and chimpanzee lineages. Indeed estimates from the free-ratios model, which assigns an ω ratio to every branch on the tree [18], suggested that the human and chimpanzee branches had the highest average ω ratios.

Discussion

A search in the literature reveals that sliding-window analysis is widely used in molecular evolution, population genetics, and comparative genomics. In between-species comparisons, it has been used to detect regions of protein under selective constraint [19] and to assess local variations in certain properties of a protein such as solvent accessibility [20] and amino acid hydrophobicity [21]. In population genetics, it has been used to identify variations in synonymous and nonsynonymous polymorphisms within species [22–28], to detect balanced selection [29], to detect recombination in a gene sequence [30,31], and to detect associations between SNPs and human diseases [32]. We do not claim that all those analyses are invalid. Indeed, Andolfatto *et al.* [33] corrected for multiple tests when they used a sliding window analysis to detect recombination. Tajima [34] discussed determination of the optimal window size. Furthermore, Ardell [35] wrote a program for performing neutrality test in a sliding window analysis by adjusting for multiple testing. Similarly Talbert *et al.* [36] used Comeron's [6] program K-ESTIMATOR to conduct a sliding-window analysis of the gene sequences of the mammalian centromere protein C (CENP-C) to detect regions under purifying and positive selection. Comeron's sliding-window approach does not correct for multiple testing, but Talbert *et al.* used a trial-and-error approach to decide empirically that positive selection was supported only if $\omega > 1.5$ and purifying selection was indicated by $\omega < 0.67$ in sliding windows of 33 codons. The trial-and-error approach was an attempt to guide against the high false positives of the sliding-window analysis.

We suggest that if a certain trend is known to exist along the sequence, it is legitimate to use sliding windows to visually illustrate it. Certain amino acid properties (such as hydrophobicity) may be expected to vary gradually along the protein sequence, because neighboring residues are often in the same secondary structural categories or in the same protein fold. If such a trend is not known to exist, it is in general invalid to use sliding windows to infer the
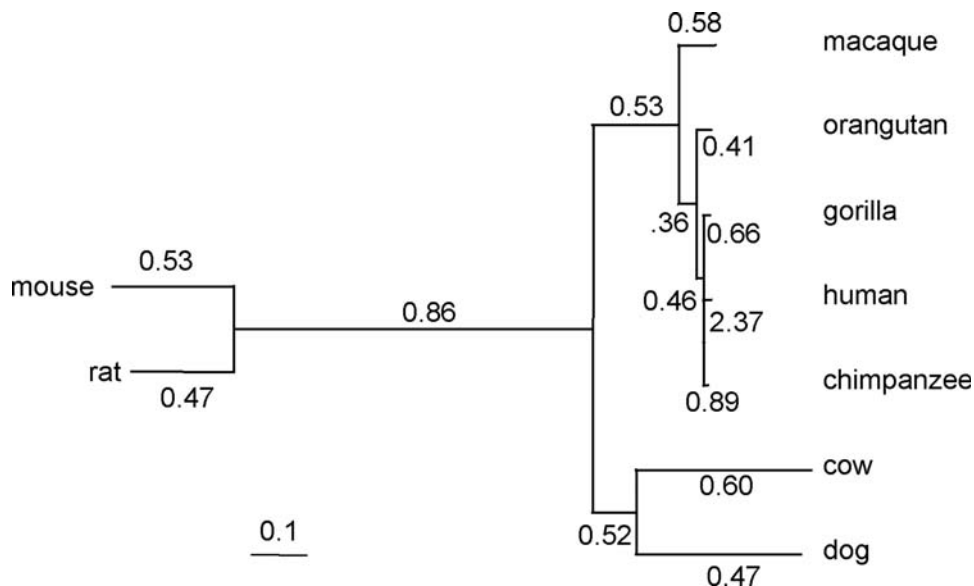
Figure 4. The phylogeny for nine mammalian species. The branch lengths, in the expected number of nucleotide substitutions per codon, are estimated under the free-ratios model [18] from analysis of the BRCA1 genes, while the estimated ω ratios for branches are shown along the branches.
doi:10.1371/journal.pone.0003746.g004

Table 2. Log-likelihood values and parameter estimates under site models for the nine mammalian BRCA1 genes.

Model	p	ℓ	Estimates of parameters	Positively selected sites
M0: one-ratio	26	−20,210.50	$\hat{\omega} = 0.585$	None
M1a: neutral	27	−20,139.17	$\hat{p}_0 = 0.484$, $\hat{p}_1 = 0.515$	
			$\hat{\omega}_0 = 0.225$, $\omega_1 = 1$	
M2a: selection	29	−20,137.43	$\hat{p}_0 = 0.508$, $\hat{p}_1 = 0.461$, $\hat{p}_2 = 0.031$	644M 897N 914N 919I (BEB, $0.5 < P < 0.6$)
			$\hat{\omega}_0 = 0.250$, $\omega_1 = 1$, $\hat{\omega}_2 = 2.276$	
M7: beta	27	−20,142.53	$\hat{p} = 0.581$, $\hat{q} = 0.365$	Not allowed
M8: beta&ω	29	−20,136.87	$\hat{p}_0 = 0.882$, $\hat{p} = 1.069$, $\hat{q} = 1.003$	897N, 914N, 919I (BEB, $0.80 < P < 0.86$)
			$\hat{p}_1 = 0.118$, $\hat{\omega} = 1.706$	

Note.— p: Number of parameters including 15 branch lengths in the tree (Figure 4). Estimates of κ range from 3.4 to 3.6 among models. Sites potentially under positive selection are listed for different models, with the human sequence used as reference. The posterior probability P that a site is from the positive-selection class is calculated using BEB [17].
doi:10.1371/journal.pone.0003746.t002

trend, because the approach will always generate a trend whether or not one exists. In addition, one has to correct for multiple testing if sliding windows are used to detect significant departures in a certain property of the molecule from null or neutral expectations. Many studies, both early and recent, did not use sliding-window analysis appropriately due to lack of an *a priori* hypothesis to stipulate the existence of the trend and due to lack of correction for multiple testing.

Materials and Methods

Mammalian BRCA1 genes

We retrieved from GenBank sequences for the breast-cancer gene BRCA1 from nine mammalian species: human (NM_007294), chimpanzee (AY365046), gorilla (AY5890), orang-utan (AY589040), macaque (AY58904), cow (NM_178573), dog (U50709), mouse (U35641) and rat (AF036760). The sequences were aligned manually. Codons with alignment gaps were removed from all species, with 1768 codons in every sequence.

Sliding-window analysis

The data in each sliding window were analyzed using the CODEML program in the PAML package [13] to fit codon model M0 (one ratio). This model involves the following parameters: sequence divergence t, measured in the number of nucleotide substitutions per codon, the transition/transversion rate ratio κ and the rate ratio $\omega = d_N/d_S$. Codon frequencies were estimated using the observed frequencies (the Fcodon model), while other parameters were estimated by ML.

Likelihood ratio test under site models

Two likelihood ratio tests of positive selection were implemented using the CODEML program [13,37,38]. The first test compares M1a (neutral) against M2a (selection). M1a assumes two site classes with $0 \le \omega_0 < 1$ (conserved sites) and $\omega_1 = 1$ (neutral sites), while M2a (selection) adds an extra class with $\omega_2 \ge 1$. The second test compares M7 (beta) against M8 (beta&ω). M7 assumes a beta distribution beta(p, q), while M8 adds an extra site class with $\omega_s \ge 1$. In both tests, twice the log likelihood difference was compared against χ_2^2 [13].

Simulation to evaluate the false positive rate of sliding-window analysis

Data sets consisting of a pair of sequences were simulated under a codon model of neutral evolution and analyzed using sliding windows to test for positive selection. Two null models were assumed to simulate datasets, with the number of replicates to be 1000. The first model was M0 (one-ratio) with $\omega = 1$. The second was M1a (neutral) with $p_0 = p_1 = \frac{1}{2}$, $\omega_0 = 0$ and $\omega_1 = 1$. In both models, the sequence distance was fixed at $t = 1$ nucleotide substitution per codon, and the transition/transversion rate ratio was fixed at $\kappa = 1$. Codon frequencies were assumed to be equal (1/61). Each simulated data set was analyzed using sliding windows, with an LRT used to test whether the single ω in M0 (one-ratio) is significantly greater than 1. The null distribution is the 50:50 mixture of point mass 0 and χ_1^2 [39], with the critical value to be 2.71 at the 5% level. Positive selection was claimed to be detected for the replicate data set if the LRT was significant in at least one window. The sequence length used was 300 codons.

Author Contributions

Conceived and designed the experiments: KS ZY. Performed the experiments: KS. Analyzed the data: KS ZY. Wrote the paper: KS ZY.

References

1. Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. Mol Biol Evol 13: 685–690.
2. Huttley GA, Easteal S, Southey MC, et al. (2000) Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Nature Genet 25: 410–413.
3. Hurst LD, Pál C (2001) Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet 17: 62–65.
4. Fares MA, Elena SF, Ortiz J, et al. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J Mol Evol 55: 509–521.
5. Sawyer SL, Emerman M, Malik HS (2004) Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol 2: E275.
6. Comeron JM (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics 15: 763–764.
7. Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analyses. Bioinformatics 15: 174–175.
8. Fares MA (2004) SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics 20: 2867–2868.

9. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7: 98–108.

10. Yang Z (2006) Computational Molecular Evolution. Oxford, England: Oxford University Press.

11. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.

12. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17: 32–43.

13. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

14. McClellan DA (2000) The codon-degeneracy model of molecular evolution. J Mol Evol 50: 131–140.

15. Wong WSW, Yang Z, Goldman N, et al. (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041–1051.

16. Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19: 950–958.

17. Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107–1118.

18. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568–573.

19. Simon AL, Stone EA, Sidow A (2002) Inference of functional regions in proteins by quantification of evolutionary constraints. Proc Natl Acad Sci USA 99: 2912–2917.

20. Yuan Z, Burrage K, Mattick JS (2002) Prediction of protein solvent accessibility using support vector machines. Proteins: Structure, Function, and Genetics 48: 566–570.

21. Peek AS, Souza V, Eguiarte LE, et al. (2001) The interaction of protein structure, selection, and recombination on the evolution of the type-1 fimbrial major subunit (*fimA*) from *Escherichia coli*. J Mol Evol 52: 193–204.

22. Llopart A, Aguade M (1999) Synonymous rates at the RpII215 gene of Drosophila: variation among species and across the coding region. Genetics 152: 269–280.

23. Kawabe A, Yamane K, Miyashita NT (2000) DNA polymorphism at the cytosolic phosphoglucose isomerase (PgiC) locus of the wild plant *Arabidopsis thaliana*. Genetics 156: 1339–1347.

24. Makova KD, Ramsay M, Jenkins T, et al. (2001) Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter. Genetics 158: 1253–1268.

25. Malik HS, Henikoff S (2001) Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*. Genetics 157: 1293–1298.

26. Polley SD, Conway DJ (2001) Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. Genetics 158: 1505–1512.

27. Presgraves DC, Balagopalan L, Abmayr SM, et al. (2003) Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature 423: 715–719.

28. Barbash DA, Awadalla P, Tarone AM (2004) Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus. PLoS Biol 2: 839–848.

29. Tian D, Araki H, Stahl E, et al. (2002) Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci USA 99: 11525–11530.

30. Grassly NC, Holmes EC (1997) A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol Biol Evol 14: 239–247.

31. Ladoukakis ED, Zouros E (2001) Recombination in animal mitochondrial DNA: evidence from published sequences. Mol Biol Evol 18: 2127–2131.

32. Mathias R, Gao P, Goldstein J, et al. (2006) A graphical assessment of *p*-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. BMC Genetics 7: 38.

33. Andolfatto P, Wall JD, Kreitman M (1999) Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of Drosophila melanogaster. Genetics 153: 1297–1311.

34. Tajima F (1991) Determination of window size for analyzing DNA sequences. J Mol Evol 33: 470–473.

35. Ardell DH (2004) SCANMS: Adjusting for multiple comparisons in sliding window neutrality tests. Bioinformatics 20: 1986–1988.

36. Talbert PB, Bryson TD, Henikoff S (2004) Adaptive evolution of centromere proteins in plants and animals. J Biol 3: 18.

37. Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929–936.

38. Yang Z, Nielsen R, Goldman N, et al. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.

39. Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82: 605–610.