1 **KaKs_Calculator 3.0: calculating selective pressure on coding and** 2 **non-coding sequences**

3

4 Zhang Zhang^{1,2,3,4,*}

5

6 ¹ China National Center for Bioinformation, Beijing 100101, China

7 ² National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences,

8 Beijing 100101, China

9 ³ CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese

10 Academy of Sciences, Beijing 100101, China

11 ⁴ University of Chinese Academy of Sciences, Beijing 100049, China

12

13 *Corresponding author (zhangzhang@big.ac.cn)

14

15 **Running title:** Zhang Z / Calculating selective pressure on coding and non-coding sequences

16

17 Manuscript type: Application Note

18 Total word counts: 1204

19 Total references: 41

20 Total tables: 1

21 Total figures: 1

22

23 Abstract

24 KaKs_Calculator 3.0 is an updated toolkit that is capable for calculating selective pressure on
 25 both coding and non-coding sequences. Similar to the nonsynonymous/synonymous
 26 substitution rate ratio for coding sequences, selection on non-coding sequences can be
 27 quantified as non-coding nucleotide substitution rate normalized by synonymous substitution
 28 rate of adjacent coding sequences. As testified on empirical data, it shows effectiveness to
 29 detect the strength and mode of selection operated on molecular sequences, accordingly
 30 demonstrating its great potential to achieve genome-wide scan of natural selection on diverse
 31 sequences and identification of potentially functional elements at whole genome scale. The
 32 package of KaKs_Calculator 3.0 is freely available for academic use only at
 33 <https://ngdc.cncb.ac.cn/biocode/tools/BT000001>.

34
 35 **KEYWORDS:** KaKs_Calculator; selective pressure; substitution; coding; non-coding;
 36

37 Introduction

38 Detecting natural selection on molecular sequences is of fundamental significance in
 39 molecular evolution, comparative genomics and phylogenetic reconstruction, which can
 40 provide profound insights for revealing evolutionary processes of molecular sequences and
 41 unveiling complex molecular mechanisms of genome evolution [1]. In principle, estimating
 42 selection on DNA sequences requires a reference set of substitutions that is free from
 43 selection. As synonymous substitutions do not provoke amino acid changes due to the
 44 degeneracy of the genetic code, they are expected to be invisible to selection and thus widely
 45 used as a reference that reflects the neutral rate of evolution [2]. Consequently, the ratio of
 46 nonsynonymous substitution rate (K_a or d_N) to synonymous substitution rate (K_s or d_s),
 47 namely, $\omega = K_a/K_s$ (or d_N/d_s), is widely adopted to differentiate neutral mutation ($\omega \approx 1$) from
 48 negative (purifying) selection ($\omega < 1$) and positive (adaptive) selection ($\omega > 1$), accordingly
 49 providing a powerful tool for illuminating molecular evolution of coding sequences (see a
 50 popular package in [3]).

51

52 Nowadays, a growing body of evidence have shown that non-coding sequences, historically
 53 thought as “junk” due to few knowledge on their function relative to coding sequences, are
 54 recognized as functional elements to play important regulation roles in multiple biological
 55 processes [4] and associate closely with various human diseases [5-7]. Albeit less conserved
 56 by comparison with coding sequences, a larger number of non-coding sequences have been
 57 identified highly conserved across mammalian genomes [8-10]. Importantly, more non-coding
 58 sequences are subject to positive selection and negative selection than previously believed,
 59 and particularly, long non-coding RNA (lncRNA) sequences do experience natural selection
 60 [11]. As a result, several methods have been proposed for computational detection of selection
 61 acting on non-coding sequences [12], which primarily differ in how to choose a reference of
 62 unconstrained evolution, such as, synonymous substitutions of neighboring coding gene [13],
 63 intron sequences [14, 15] and ancestral repeats [16]. However, there lacks of an implemented
 64 algorithm to detect the strength and mode of selective pressure on non-coding sequences,
 65 particularly considering an increasing number of non-coding studies conducted worldwide.
 66 More importantly, an integrated toolkit that is capable to detect selection on both coding and
 67 non-coding sequences is highly desirable, which would help users achieve genome-wide scan
 68 of natural selection on diverse sequences.

69

70 Towards this end, here we present KaKs_Calculator 3.0, an updated toolkit for calculating
 71 selective pressure on both coding and non-coding sequences. Compared with previous
 72 versions [17, 18], we implement an algorithm in KaKs_Calculator 3.0 that employs
 73 synonymous sites of adjacent coding sequences as a reference to estimate selective pressure
 74 acting on non-coding sequences. We test it on empirical data and demonstrate its utility in
 75 diagnosing the strength and form of molecular evolution.

76 **Algorithm**

77 The major update of KaKs_Calculator 3.0 is to incorporate an algorithm that is capable for
 78 estimating selective pressure on non-coding sequences. Specifically, it uses synonymous
 79 substitutions as a reference baseline (similar to [13]), which, albeit thought to be under weak

selection [19-21], has been widely adopted for determining the strength and type of selection operated on coding sequences [22-27]. Similarly, selective pressure on non-coding sequences (ξ) can be quantified as non-coding nucleotide substitution rate (Kn) normalized by neutral substitution rate (assumed as Ks), viz., $\xi = \text{Kn}/\text{Ks}$, where Ks is inferred from adjacent coding sequences. As the number of observed substitutions is less than the number of real substitutions, we adopt a nucleotide substitution model (e.g., JC/K2P/HKY) to correct multiple substitutions of non-coding sequences. Taking the HKY model [28] as an example, therefore, Kn can be deduced from the observed transitional and transversional substitutions (S and V, respectively) as well as four nucleotide frequencies (π_A , π_T , π_G and π_C), according to Equation 1 (see Equations 1.27 and 1.28 in [29]).

$$\text{Kn} = 2\left(\frac{\pi_T\pi_C}{\pi_Y} + \frac{\pi_A\pi_G}{\pi_R}\right)a - 2\left(\frac{\pi_T\pi_C\pi_R}{\pi_Y} + \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \pi_Y\pi_R\right)b, \quad (1)$$

where $a = -\log\left[1 - \frac{S}{2(\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R)} - \frac{(\pi_T\pi_C\pi_R/\pi_Y + \pi_A\pi_G\pi_Y/\pi_R)V}{2(\pi_T\pi_C\pi_R + \pi_A\pi_G\pi_Y)}\right]$, $b = -\log\left(1 - \frac{V}{2\pi_Y\pi_R}\right)$, $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_T + \pi_C$. To detect and quantify selection on non-coding sequences, KaKs_Calculator 3.0 provides users with two ways to obtain the value of neutral mutation rate or Ks, which is either calculated from adjacent coding sequences uploaded by users or just specified in a straightforward manner by users (Figure 1). As a consequence, KaKs_Calculator 3.0 is capable to detect selection on both coding and non-coding sequences.

KaKs_Calculator 3.0 is implemented in standard C++ language, enabling higher efficiency and easy compilation on different operation systems (Linux/Windows/Mac). In addition to the new functionality for estimating selection on non-coding sequences as mentioned above, it is also updated by fixing bugs and errors. The package of KaKs_Calculator 3.0, including compiled executables, a GUI (graphical user interface) application on Windows, source codes and example data, accompanying with detailed instructions and documentation, is freely available for academic use only at BioCode (<https://ngdc.cncb.ac.cn/biocode/tools/BT000001>), an open-source platform for archiving bioinformatics tools in the National Genomics Data Center [30], China National Center for Bioinformatics.

107 Application on empirical data

108 To test KaKs_Calculator 3.0, we choose three empirical lncRNA genes that are extensively
 109 studied according to LncRNAWiki [7] and collect their human-mouse orthologs as well as
 110 their adjacent coding orthologs from NGDC LncBook [31] and NCBI RefSeq [32].
 111 Specifically, these non-coding and coding gene symbols with accession numbers are: (1) *H19*
 112 (NR_002196.2 vs. NR_130973.1) and *MRPL23* (NM_021134.4 vs. NM_011288.2); (2)
 113 *MALAT1* (NR_002819.4 vs. NR_002847.3) and *SCYL1* (NM_020680.4 vs. NM_001361921.1);
 114 and (3) *HOTAIR* (NR_003716.3 vs. NR_047528.1) and *HOXC12* (NM_173860.3 vs.
 115 NM_010463.2). Based on these orthologous genes, we obtain their corresponding aligned
 116 sequences by MAFFT [33] (using parameters: --maxiterate 1000 --localpair). It is noted that
 117 one non-coding sequence may have multiple adjacent coding genes, which are completely
 118 specified by users and thus can lead to different estimates of Ks and ξ .

119
 120 By contrasting non-coding nucleotide substitution rate to adjacent synonymous substitution
 121 rate, we reveal that, although the coding genes undergo strong purifying selection ($\omega < 1$),
 122 these three non-coding genes present diverse selective pressure. Strikingly, *HOTAIR* (*Hox*
 123 transcript antisense intergenic RNA) exhibits positive selection ($\xi > 1$), whereas the rest two
 124 genes experience negative selection ($\xi < 1$). *HOTAIR* is a ~2.3 kb intergenic RNA transcribed
 125 from the antisense strand of the *HOXC* gene cluster [34]. The result of positive selection
 126 detected on *HOTAIR* relative to *HOXC12* is consistent well with previous findings that
 127 *HOTAIR* evolves faster than the neighbouring genes [35]. On the contrary, *MALAT1*
 128 (metastasis-associated lung adenocarcinoma transcript 1), a ~8.7 kb non-coding RNA flanked
 129 by the highly conserved kinase-like gene *SCYL1*, is ubiquitously expressed in almost all
 130 human tissues, evolutionarily conserved across mammalian species [36] and associated with
 131 various cancers [37]. Thus, $\xi = 0.464$ indicates strong selective constraint on *MALAT1*, in
 132 accordance with its physiologic and pathophysiological function [38] and conserved RNA
 133 structure [39] as documented by previous studies. Likewise, *H19*, a ~2.3 kb imprinted
 134 maternally expressed transcript located near *MRPL23*, is known for close association with
 135 Beckwith-Wiedemann Syndrome and also involved in tumorigenesis [40]. Our result shows

136 that *H19* presents stronger selection constraint as indicated by $\xi=0.296$, conforming well with
 137 its conserved sequence and structure [41]. Taken together, KaKs_Calculator 3.0 is effective in
 138 estimating natural selection for non-coding sequences, which has the potential to reveal
 139 evolutionarily selective pressures operated on diverse molecular sequences.

140 **Conclusion**

141 KaKs_Calculator 3.0 is significantly updated by achieving the detection of natural selection
 142 on non-coding sequences as well as coding sequences. As testified on empirical data, it is of
 143 great utility in calculating natural selection on molecular sequences and thus identifying
 144 potentially functional elements at genome-wide scale.

145 **Code availability**

146 KaKs_Calculator 3.0 is freely available for academic use only at
 147 <https://ngdc.cncb.ac.cn/biocode/tools/BT000001>.

148 **CRedit author statement**

149 **Zhang Zhang:** Conceptualization, Methodology, Software Development, Supervision,
 150 Writing , Reviewing and Editing

151 **Competing interests**

152 The author has declared no competing interests.

153 **Acknowledgements**

154 I would like to extend special thanks to Lina Ma for constructive suggestions and discussions
 155 on this work and Zhao Li for valuable help on data collection. I also thank Zhuojing Fan for
 156 designing the logo as well as Qing Guo and Lin Dai for fixing a bug on Windows GUI. I am
 157 extremely grateful to a number of users for reporting bugs and sending comments since the
 158 first release of KaKs_Calculator in 2006. This work was supported by the Strategic Priority
 159 Research Program of the Chinese Academy of Sciences [XDA19050302], the National

160 Natural Science Foundation of China [32030021; 31871328], National Key R&D Program of
161 China [2017YFC0907502], and International Partnership Program of the Chinese Academy of
162 Sciences [153F11KYSB20160008].

163

164 Authors' ORCID ID

165 0000-0001-6603-5060 (Zhang Zhang)

166

167 References

- 168 [1] Li W-H. Molecular Evolution. Sunderland, Massachusetts: Sinauer Associates, 1997.
169 [2] Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 2002;18:486.
170 [3] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586-91.
171 [4] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia
172 of DNA elements in the human genome. Nature 2012;489:57-74.
173 [5] Luo H, Bu D, Shao L, Li Y, Sun L, Wang C, et al. Single-cell Long Non-coding RNA Landscape of
174 T Cells in Human Cancer Immunity. Genomics Proteomics Bioinformatics 2021.
175 [6] Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer
176 2018;18:5-18.
177 [7] Liu L, Li Z, Liu C, Zou D, Li Q, Feng C, et al. LncRNAWiki 2.0: a knowledgebase of human long
178 non-coding RNAs with enhanced curation model and database system. Nucleic Acids Res 2022;doi:
179 10.1093/nar/gkab998.
180 [8] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved
181 elements in the human genome. Science 2004;304:1321-5.
182 [9] Habc A, Mattick JS, Calin GA, Krese R, Konc J, Kunej T. Genetic Variations of Ultraconserved
183 Elements in the Human Genome. OMICS 2019;23:549-59.
184 [10] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals
185 over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223-7.
186 [11] Diederichs S. The four dimensions of noncoding RNA conservation. Trends Genet 2014;30:121-3.
187 [12] Zhen Y, Andolfatto P. Methods to detect selection on noncoding DNA. Methods Mol Biol
188 2012;856:141-59.
189 [13] Wong WS, Nielsen R. Detecting selection in noncoding regions of nucleotide sequences. Genetics
190 2004;167:949-58.
191 [14] Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the utility of short
192 intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol
193 Biol Evol 2010;27:1226-34.
194 [15] Hoffman MM, Birney E. Estimating the neutral rate of nucleotide substitution using introns. Mol
195 Biol Evol 2007;24:522-31.
196 [16] Bush EC, Lahn BT. A genome-wide screen for noncoding elements important in primate evolution.

197 BMC Evol Biol 2008;8:17.

198 [17] Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks
199 through model selection and model averaging. Genomics Proteomics Bioinformatics 2006;4:259-63.

200 [18] Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating
201 gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics
202 2010;8:77-80.

203 [19] Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to
204 biological regulation and complexity. Nucleic Acids Res 2013;41:2073-94.

205 [20] Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet 2008;42:287-99.

206 [21] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias.
207 Nature Reviews Genetics 2011;12:32-42.

208 [22] Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of
209 nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol.
210 Evol. 1985;2:150-74.

211 [23] Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and
212 nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 1986;3:418-26.

213 [24] Li WH. Unbiased estimation of the Rates of synonymous and nonsynonymous substitution. J. Mol.
214 Evol. 1993;36:96-9.

215 [25] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA
216 sequences. Mol. Biol. Evol. 1994;11:725-36.

217 [26] Yang Z, Nielsen R. Estimating Synonymous and Nonsynonymous Substitution Rates Under
218 Realistic Evolutionary Models. Mol. Biol. Evol. 2000;17:32-43.

219 [27] Zhang Z, Li J, Yu J. Computing Ka and Ks with a consideration of unequal transitional
220 substitutions. BMC Evol Biol 2006;6:44.

221 [28] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of
222 mitochondrial DNA. J Mol Evol 1985;22:160-74.

223 [29] Yang Z. Computational Molecular Evolution. London: Oxford University Press, 2006.

224 [30] CNCB-NGDC Members & Partners. Database Resources of the National Genomics Data Center,
225 China National Center for Bioinformation in 2021. Nucleic Acids Res 2021;49:D18-D28.

226 [31] Ma L, Cao J, Liu L, Du Q, Li Z, Zou D, et al. LncBook: a curated knowledgebase of human long
227 non-coding RNAs. Nucleic Acids Res 2019;47:D128-D34.

228 [32] Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretin A, et al. RefSeq: expanding the
229 Prokaryotic Genome Annotation Pipeline reach with protein family model curation. Nucleic Acids Res
230 2021;49:D1020-D8.

231 [33] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements
232 in performance and usability. Mol Biol Evol 2013;30:772-80.

233 [34] Tang Q, Hann SS. HOTAIR: An Oncogenic Long Non-Coding RNA in Human Cancer. Cell
234 Physiol Biochem 2018;47:893-913.

235 [35] He S, Liu S, Zhu H. The sequence, structure and evolutionary features of HOTAIR in mammals.
236 BMC Evol Biol 2011;11:102.

237 [36] Gutschner T, Hammerle M, Diederichs S. MALAT1 -- a paradigm for long noncoding RNA
238 function in cancer. J Mol Med (Berl) 2013;91:791-801.

239 [37] Meseure D, Vacher S, Lallemand F, Alsibai KD, Hatem R, Chemlali W, et al. Prognostic value of a
240 newly identified MALAT1 alternatively spliced transcript in breast cancer. Br J Cancer

241 2016;114:1395-404.
 242 [38] Zhang X, Hamblin MH, Yin KJ. The long noncoding RNA Malat1: Its physiological and
 243 pathophysiological functions. RNA Biol 2017;14:1705-14.
 244 [39] Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in
 245 mammals. Nucleic Acids Res 2013;41:8220-36.
 246 [40] Hurst LD, Smith NG. Molecular evolutionary evidence that H19 mRNA is functional. Trends
 247 Genet 1999;15:134-5.
 248 [41] Juan V, Crain C, Wilson C. Evidence for evolutionarily conserved secondary structure in the H19
 249 tumor suppressor RNA. Nucleic Acids Res 2000;28:1221-7.
 250

251 Tables

252 Table 1 Estimates of selective pressure as well as substitution rates in human-mouse orthologs

Non-coding			Coding			
Gene Symbol	Kn	$\xi=Kn/Ks$	Gene Symbol	Ka	Ks	$\omega=Ka/Ks$
<i>H19</i>	0.340	0.296	<i>MRPL23</i>	0.088	1.150	0.077
<i>MALAT1</i>	0.324	0.464	<i>SCYL1</i>	0.040	0.697	0.058
<i>HOTAIR</i>	0.544	1.114	<i>HOXC12</i>	0.020	0.488	0.041

253

254 Figure legends

255 Figure 1 Graphical user interface of KaKs_Calculator 3.0.

256

257