Identification of Differentially Evolved Genes: An Alternative Approach to Detection of Accelerated Molecular Evolution from Genome-Wide Comparative Data

Evolutionary Bioinformatics

METHODOLOGY

Identification of Differentially Evolved Genes: An Alternative Approach to Detection of Accelerated Molecular Evolution from Genome-Wide Comparative Data

Kyu-Won Kim[1], David W. Burt[2], Heebal Kim[3] and Seoae Cho[4]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea. [2]The Division of Genetics and Genomics, The Roslin Institute and Royal (Dick) School of Veterinary Studies, Edinburgh University, EH25 9RG, UK. [3]Department of Agricultural Biotechnology, Seoul National University, Seoul 151-742, Korea. [4]C&K genomics, 514 Main Building, Seoul National University Research Park, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.
Corresponding author email: seoae@cnkgenomics.com

Abstract: One of the most important measures for detecting molecular adaptations between species/lineages at the gene level is the comparison of relative fixation rates of synonymous (dS) and non-synonymous (dN) mutations. This study shows that the branch model is sensitive to tree topology and proposes an alternative approach, devogs, which does not require phylogenetic topology for analysis. We compared devogs with a branch model method using virtual data and a varying ω ratio, in which parameters were obtained from real data. The positive predictive value, sensitivity, and specificity of the branch model were affected by the phylogenic tree topology. Devogs showed greater positive predictive value, whereas the branch model method had greater sensitivity. In a working example using devogs, a group of human RNA polymerase II-related genes, which are important in mediating alternative splicing, were significantly accelerated compared to four other mammals.

Keywords: devogs, dN/dS, branch model

Evolutionary Bioinformatics 2013:9 285–299

doi: 10.4137/EBO.S12166

This article is available from http://www.la-press.com.

Introduction

One of the most important measures for detecting molecular adaptation is to compare the relative substitution rates of synonymous (dS) and non-synonymous (dN) mutations.[1] The non-synonymous/synonymous rate ratio (ω = dN/dS) measures selective pressure, with $\omega = 1$ indicative of neutral mutation, $\omega < 1$ indicative of purifying or negative selection, and $\omega > 1$ indicative of positive diversifying selection. Several methods have been developed to apply this criterion to particular lineages on a phylogeny (branch methods)[2,3] or to subsets of gene sites (site methods).[4–7] Based on the site or branch-site model, a series of likelihood ratio tests (LRTs) have been used in a comprehensive examination of positively selected genes in six eutherian mammals.[8] Current methods for detecting molecular adaptation use phylogenetic trees for analysis. However, phylogenetic relationships require further examination before they can be confirmed, although the phylogenetic relationships of orders have been addressed in several recent molecular studies. Previous studies have yielded inconsistent results with respect to some ordinal relationships. For example, the phylogenetic positions of rodents, primates, and carnivores remain unclear. Traditional morphology supports a primate–rodent clade,[9] but molecular studies support either a primate–rodent clade[10] or a primate–carnivore clade.[11] Even a primate–cetartiodactyla clade is supported by mitochondrial DNA analysis.[12] The phylogenetic tree of Brucella, a genus of host-specific bacteria, is not consistent with host species; for example, Brucella canis (dog host) is closer to Brucella suis (pig host) than to another dog host in the phylogenetic tree.[13]

Here, we propose an alternative approach for detecting accelerated molecular evolution between species at the individual gene level using relative criteria and not phylogenetic trees. This method uses similar data structures and analysis approaches as gene expression microarray data, which involves the identification of "differentially evolved genes" (devogs). This method can be used with data from one gene up to large comparative genomic data sets with no prior biological assumptions, which may be responsible for the evolutionary differentiation between two clades of interest. We show that the present branch method is significantly affected by phylogenetic tree topology using various phylogenetic trees with real data from five mammalian species. We also compare devogs with the present method in terms of positive predictive value (PPV), which is defined as the proportion of predicted positives that are actually positive, sensitivity, and specificity using virtually evolved data from five real mammals. Additionally, we compare human-specific accelerated genes with four other mammals using devogs. Devogs has already been applied to an evolutionary study in avian lineages at the turkey genome project.[14]

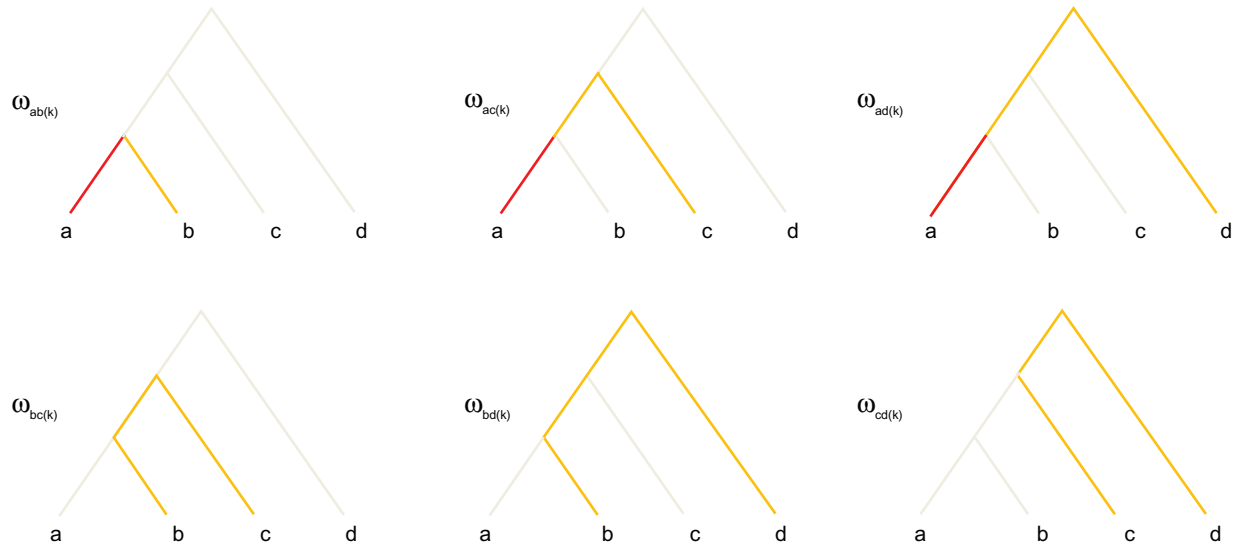
Methods

Concept

Let $\omega_{ij(k)}$ be the pairwise ω ratio between i and j species of an orthologous gene $k \in O$, where i and $j \in S$ are in a species set S and their orthologous set O. For example, six pairs ($\omega_{ab(k)}$, $\omega_{ac(k)}$, $\omega_{ad(k)}$, $\omega_{bc(k)}$, $\omega_{bd(k)}$, and $\omega_{cd(k)}$) were generated from four species: a, b, c, and d. Focusing on species a, the pair ω_{ij} can be divided into two groups: $\omega_{*a(k)} = \{\omega_{ij(k)} \mid i \text{ or } j \text{ is } a,\ a, i, j \in S,\ k \in O\}$ and $\omega_{\wedge a(k)} = \{\omega_{ij(k)} \mid i \text{ and } j \text{ are not } a,\ a, i, j \in S,\ k \in O\}$ (Fig. 1). Orthologous gene k under accelerated evolution within species a would increase the mean ω ratio in $\omega_{*a(k)}$ compared to $\omega_{\wedge a(k)}$. Verification of whether the mean ω ratio of $\omega_{*a(k)}$ is higher than $\omega_{\wedge a(k)}$ can be used to detect accelerated evolution within species a.

Implementation

t-test comparison between $\omega_{*a(k)}$ and $\omega_{\wedge a(k)}$ was used to identify differences between the two means. We normalized the ω ratios because one of the assumptions of the t-test is data normality. Base-2-logarithm transformation of the ω ratio was performed since the ω ratio was similar to the red/green (R/G) intensity of two-channel expression microarrays. Quantile normalization Q: $x \rightarrow$ quantile normalized x on $\{\log_2 \Omega_{ij} \mid i, j \in S\}$ was then performed, where Ω_{ij} was $\{\omega_{ij(k)} \mid i, j \in S,\ k \in O\}$ and $\log_2 \Omega_{ij}$ was $\{\log_2 \omega_{ij(k)} \mid i, j \in S,\ k \in O\}$. The assumption that bias exists between Ω_{ab} and Ω_{cd} is more reasonable than supposing that evolution between species a and b is faster than that in species c and d when the mean ω ratio of Ω_{ab} is higher than that of Ω_{cd} ($a \neq b \neq c \neq d$, $a, b, c, d \in S$). The bias among $\{\Omega_{ij} \mid i, j \in S\}$ affects the t-test. For example, the mean of $\omega_{*a(k)}$ is higher than $\omega_{\wedge a(k)}$ in many orthologous genes when Ω_{*a} is higher than $\Omega_{\wedge a}$, where Ω_{*a} is $\{\omega_{*a(k)} \mid a \in S,\ k \in O\}$ and $\Omega_{\wedge a}$ is $\{\omega_{\wedge a(k)} \mid a \in S,\ k \in O\}$.



Gene	Paring with a			Paring without a		
K	$\omega_{ab(k)}$	$\omega_{ac(k)}$	$\omega_{ad(k)}$	$\omega_{bc(k)}$	$\omega_{bd(k)}$	$\omega_{cd(k)}$

Figure 1. Schematic view of two groups of the ω ratio. a, b, c, and d represent species. $\omega_{ij(k)}$ represents the pairwise ω ratio between i and j species of k orthologous genes.

Orthologous gene k was verified as an accelerated gene based on two conditions: the mean of $\{Q(\log_2\omega_{*a(k)})|a\in S,\ k\in O\}$ was higher than the mean of $\{Q(\log_2\omega_{\wedge a(k)})|a\in S,\ k\in O\}$ and the P-value of $\omega_{*a(k)}$ against $\omega_{\wedge a(k)}$ was below 0.05, which considered to be significant. We referred to this method as devogs, or identification of differentially evolved genes.

In addition to correction for multiple hypothesis testing, Benjamini and Hochberg false discovery rates can be applied.

Validation

Branch model under various phylogenic tree topologies

To examine the consistency of the branch method with tree structures, we performed the branch method under various tree topologies. The tree topologies were obtained from real data. We prepared 10,891 orthologs of human, mouse, rat, dog, and opossum from ENSEMBL.[15] Trees for each ortholog were generated using Mrbayes v3.1.2 (n-generation: 2×10^5, burn-in: 2×10^3).[16] Distinct tree topologies from all orthologs were identified using TOPD-fMtS v3.3 (split method).[17] Branch model using codeml in PAML4.2 (F3X4) was applied to all orthologs with changing tree topologies along distinct trees.[18]

Creation of simulated test data and evaluation of devogs and the branch model

To compare the performance of devogs and the branch model in codeml in PAML4.2,[18] we generated simulated data with evolved genes. We then used the two methods to evaluate the simulated data and measured their performance.

Generation of test data

To compare the two methods under real conditions, we constructed virtual data using real parameters, including tree topology (excluding ω, which is the variable being manipulated). A total of 30 orthologous gene sets, which reduced the computational burden, were selected randomly from the genomes of human, mouse, rat, dog, and opossum obtained from ENSEMBL[15] to choose parameter values, including codon usage, substitutions per codon, and sequence length. Test DNA sequences from the five species were generated using the evolver program in PAML4.2 (kappa = 2.0, tree length = 0) based on the

actual parameters.[18] The phylogenetic tree and branch lengths converted to species divergence times, as required by the evolver program, were obtained from a previous report (Fig. 2).[19] In addition, the branch lengths were rescaled based on the average number of substitutions per codon between mice and humans. We included an accelerated branch with a higher ω ratio to compare human, mouse, rat, dog, and opossum branches when generating data. Differences in the ω ratio between the accelerated branch and other branches ranged from 0 to 1.2 (steps of 0.1).

Testing devogs and the branch model with virtual data

Orthologous sequences were aligned using ClustalW2[20] and alignments were converted to codon alignments using pal2nal.[21] Both devogs and the branch model were applied to the data. Each species was set as the foreground once to test acceleration. In this case, we estimated ω values for devogs using codeml in PAML4.2 (runmode = –2 option),[18] which adopts the maximum-likelihood method. Compared to the branch model, the only difference was that tree topology information was not used when estimating pairwise ω between two species; therefore, the comparison is valid and not affected by bias in maximum-likelihood methods.

Devogs analysis of real data

We downloaded 1:1 orthologous protein and reference mRNA sequences of human, mouse, rat, dog, and opossum from ENSEMBL.[15] The phylogenetic trees were obtained from a previous report.[19] A total of 10,891 1:1 orthologous genes for the five species were collected, and the orthologous gene sets were aligned using ClustalW2.[20] The devogs method was applied for identification of accelerated genes in humans. Orthologs with dS $>$ 3 or $\omega > 5$ were filtered.[22,23] A total of 8,407 orthologs were examined.

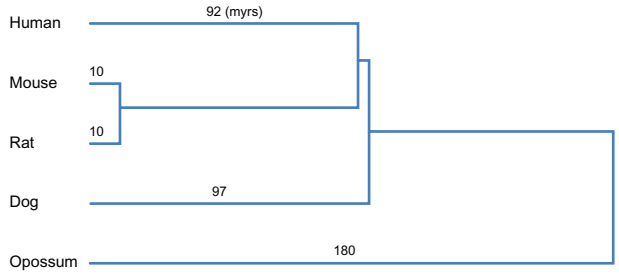


Figure 2. Branch length for generating virtual data.

Results

Branch model under various phylogenic tree topologies

A total of 10,889 gene trees for 10,891 orthologs from human, mouse, rat, dog, and opossum from ENSEMBL[15] were generated using Mrbayes.[16] The 10,889 trees were grouped into 14 distinct topologies. Trees were arranged in order of the number of orthologous genes mapping to each tree, which were named A, B, C, … N. The branch model was used to detect genes showing accelerated evolution in each of five lineages within each of these 14 trees.

Some genes showed evidence of accelerated evolution. However, the number of predicted accelerated genes varied from 9.7%–15.2% between tree topologies. Figure 3 shows the statistical results of the branch model for all orthologous gene sets in the 14 distinct gene trees. The accelerated genes in the lineage with their own trees were considered to be true accelerated genes for the calculation of PPV, sensitivity, and specificity. The positive predictive value (PPV), which is defined as the proportion of predicted positives that are positive, varied from 0.17–0.98 among gene trees. The degree of variation of the PPV in the tree topologies was high for mouse and rat. The sensitivities also showed wide variation, with values ranging from 0.29–0.98. The degree of decrease in sensitivity on the minor trees was high for humans. The sensitivities of mice and rats were typically higher than the others, in contrast to PPV. Specificities were high for all trees, ranging from 0.77–0.98. However, the specificities of mice and rats in the G, H, I, M, and N trees were relatively low, which reduced the PPV of mice and rats in the trees, despite their high sensitivity.

Comparison of the performance of devogs and the branch model for simulated data

We compared the devogs approach with a branch model method using virtual data from human, mouse, rat, dog, and opossum, which evolved under varying ω ratios. To compare the two methods under realistic conditions, the parameters, including tree topology (excluding ω), were obtained from real data in ENSEMBL.[15] A total of 10,965 orthologous sets were generated after filtering, and devogs and the branch

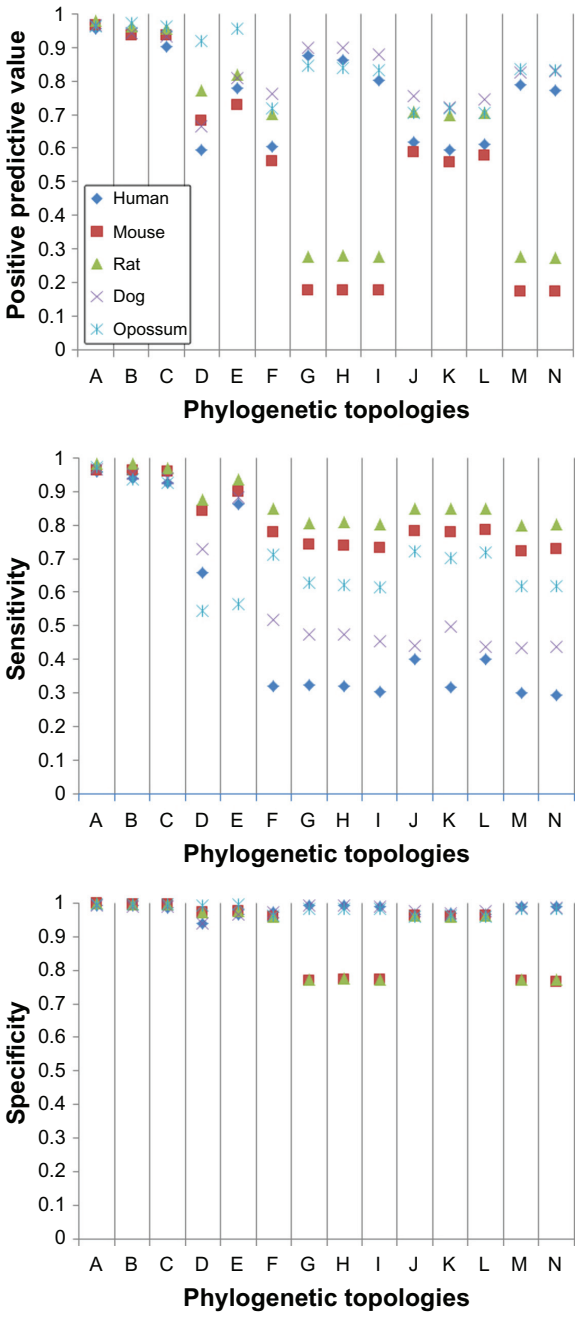


Figure 3. Statistical representations of performance for detecting accelerated genes from each foreground lineage using the branch model with real orthologous gene sets on 14 trees. Positive predictive value = TP/(TP + FP). Sensitivity = TP/(TP + FN). Specificity = TN/(TN + FP). TP: number of true-positives; FP: number of false-positives; TN: number of true-negatives; FN: number of false-negatives.

model were used to detect accelerated evolution of one species at a time among the five species (five times in total). A total of 54,825 examinations were conducted using both models. Additionally, we included an altered devogs approach where normalization and *t*-tests were replaced with the non-parametric Mann Whitney U-test.

Of the 54,825 tests, the devogs method predicted that a total of 5,514 genes (10.0%) showed evidence of accelerated evolution. Additionally, 4,630 (84.0%) of these genes were true-positives. Devogs with the Mann Whitney U-test predicted that a total of 5,690 genes (10.4%) showed evidence of accelerated evolution. Of these, 4,909 (86.3%) genes were true-positives. The branch model predicted that a total of 7,781 genes (14.2%) showed evidence of accelerated evolution. Of these, 6,419 (82.5%) genes were true-positives.

Figure 4 shows the Venn diagram of positives and true-positives from two types of devogs and the branch model. Many overlapped genes were identified among the models. There are more overlapped genes per model than non-overlapped genes in all models, whereas the branch model contained the largest number of non-overlapped genes (2,609 genes) and devogs incorporating the Mann Whitney U-test showed the lowest number of non-overlapping genes (234 genes). The ratio of true-positives was higher for overlapped genes than non-overlapped genes, and the ratio of true-positives was highest (95.4%) at the intersection of the three models. As shown in Table S1, total PPV, sensitivity, and specificity of both methods were similar. PPV and specificity of devogs with the *t*-test were slightly higher than the branch model, and sensitivity of the branch model was higher than that of devogs. Sensitivity of devogs with the Mann Whitney U-test was higher than that of devogs with the *t*-test.

Next, we measured the performance of both methods under varying ω ratio differences between the

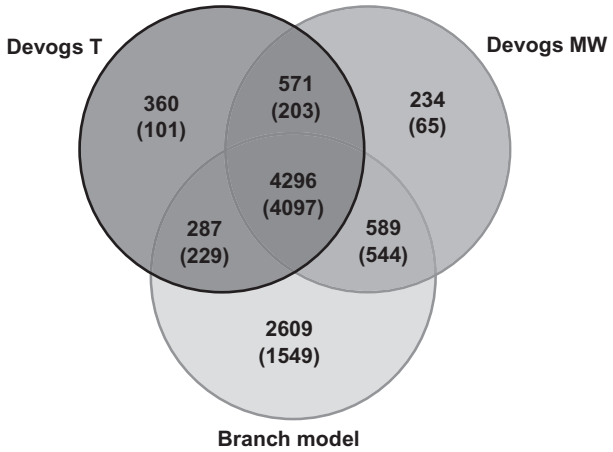


Figure 4. Venn diagram of positives and true positives from two types of devogs and the branch model. Devogs T: devogs with *t*-test. Devogs MW: devogs with Mann Whitney U-test.

accelerated branch and other branches from 0–1.2 (steps of 0.1). Figure 5 shows the statistical measures of performance with differences between foreground and background ω ratios. The "longs" is the group consisting of humans, dogs, and opossums with long-branch lengths between the species, as shown in Figure 2. The "shorts" is the group consisting of the mouse and rat, with a short-branch length between species. We divided the species into short- and long-branch length groups based on the results of foreground ω having a tendency to be grouped into mainly two patterns according to branch length (Fig. S1).

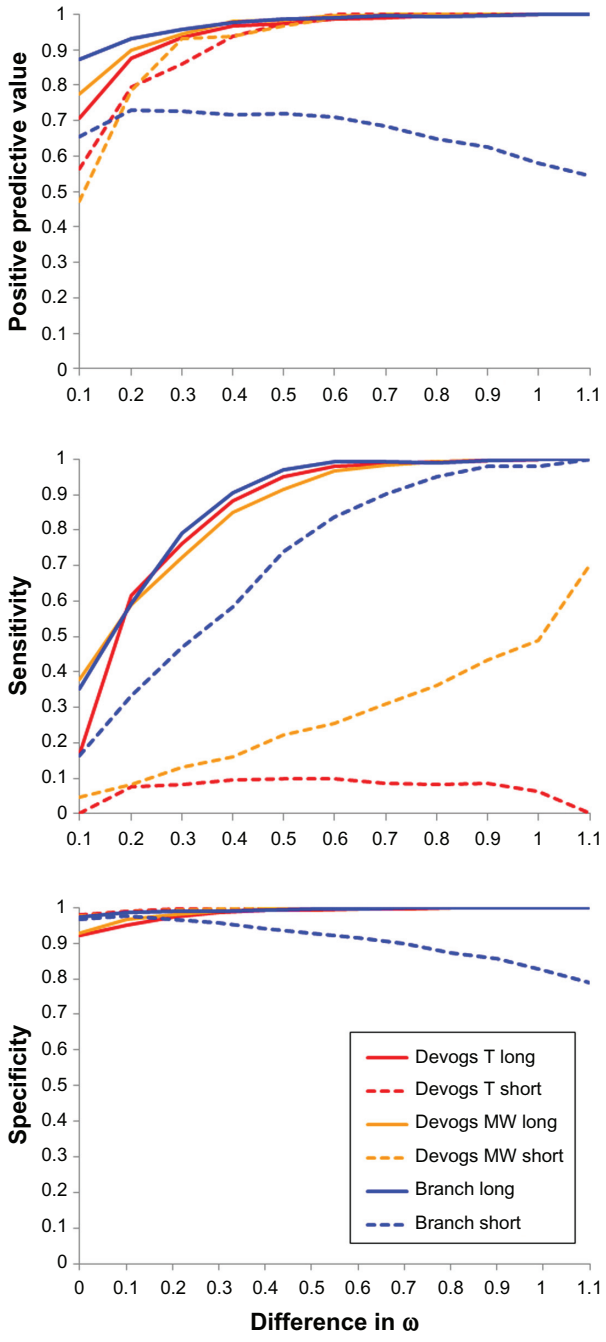


Figure 5. Statistical assessments of the performance of two types of *devogs* and the branch model based on differences in the ω ratio between foreground and background lineages from virtually accelerated genes through simulation. The longs are the human, dog, and opossum groups, while the shorts are the mouse and rat groups. Devogs T: *devogs* with *t*-test. Devogs MW: *devogs* with Mann-Whitney U-test.

PPVs of all methods (excluding the short group in the branch model) are high and increased rapidly with increasing differences in ω ratios between the accelerated branch and other branches (Fig. 5). The PPV of the branch model in the short group peaked at 0.73 at a difference in ω ratios of 0.2, and decreased to 0.54. The overall sensitivity increased gradually with increasing differences in the ω ratio (excluding the short group in devogs). The curve slope of the short group was lower than that of the long group. The sensitivity in the short group of devogs incorporating the *t*-test remained very low (<0.10) across all differences in ω ratio, whereas the sensitivity in the short group of devogs with the Mann Whitney U-test increased gradually to 0.70. Specificities were very high from a difference in ω ratio of 0, and increased linearly to 1.0, excluding the short group in the branch model. The specificity of the short group in the branch model decreased from 0.97 to 0.79 with increasing differences in the ω ratio.

Devogs analysis of real data

We applied devogs analysis to identify genes with accelerated evolution in humans compared to other mammals with 1:1 orthologous sequences of human, mouse, rat, dog, and opossum from ENSEMBL.[15] A total of 50 genes showed accelerated evolution in humans compared with other species with FDR < 0.05 from 8,407 orthologous genes with dS ≤ 3 (Table S2). Differences in the ω ratio between these two groups ranged from 0.005–0.603, although the ω ratios of the group pairing with humans were <1. The difference in the ω ratio of the *NAA30* gene was 0.443. Differences in the ω ratio of *ETV6*, *LEPROTL1*, *TM4SF19*, *SMEK1*, and *HTRA4* were all >0.20. Gene enrichment analysis of GO terms was performed using the DAVID functional annotation tool[24] with thresholds (count 2, EASE 0.1). Table 1 shows enriched GO terms of biological process, cellular

Table 1. Functional annotation of enriched GO terms for human accelerated genes detected using devogs.

Category	Term	*P*-value*
Bioprocess	ER-associated protein catabolic process	0.062
Cellular component	DNA-directed RNA polymerase II, core complex	0.043
	Nuclear DNA-directed RNA polymerase complex	0.075
	DNA-directed RNA polymerase complex	0.075
	RNA polymerase complex	0.078
Molecular function	Protein kinase activity	0.003
	N-acetyltransferase activity	0.012
	Acetyltransferase activity	0.016
	N-acyltransferase activity	0.017
	Insulin-like growth factor binding	0.058
	Protein heterodimerization activity	0.088
	RNA polymerase activity	0.098
	DNA-directed RNA polymerase activity	0.098

Note: *EASE score.

component, and molecular function. Terms related to RNA polymerase II were also significantly enriched.

We also applied the branch model to the same actual data. A total of 553 genes were predicted to show accelerated evolution. Additionally, 39 of 50 genes identified during devogs analysis agreed with genes from the branch model (Table S3). Agreement of genes between devogs and the branch model increased as the *P*-values of the branch model results decreased (Fig. S3).

Discussion

Performance of the branch model and devogs approaches under various phylogenic tree topologies

The accuracy of results obtained using the branch model was affected by the phylogenic tree topology. In addition, use of the incorrect tree topology reduced overall accuracy. A PPV of 0.17, as shown in Figure 3, indicated that 83% of the results from the branch model were affected by tree topology, although very low PPVs were determined for specific orthologous genes in our experiment. However, devogs does not require phylogenetic tree topology during the analysis process. Therefore, devogs was not affected by phylogenic topology.

Comparison of the performance of devogs and branch model approaches

Both devogs and the branch model approaches showed acceptable overall performance, as shown in Figures 4 and 5. However, both devogs and the branch model showed low performance for detecting accelerated genes in mice and rats. This may be because the numbers of synonymous and non-synonymous substitutions between the mouse and rat genes were very small and did not cause statistically significant differences in ω ratios, as the branch lengths of the mouse and rat are relatively shorter than the other lineages, as shown in Figure 2. However, the responses of devogs and the branch model under their low performance conditions differed significantly. Sensitivity was markedly reduced in devogs, while PPV was reduced in the branch model. This difference in responses may explain the advantages and disadvantages of using tree topology information, particularly under low performance conditions since the only difference in input information for the branch model compared to devogs is tree topology information. The topology information input into the branch model may increase sensitivity, but may also increase the false-positive rate, thus decreasing specificity and PPV. In actual analysis, the true-positive and true-negative conditions as determined by the gold standard remain unknown, while the positive and negative values from the test outcome are available. False-positives are problematic for identifying truly accelerated genes, although the sensitivity of the model has been shown to be high in many studies. However, PPV can be used since it is defined as the sum of true positives divided by the test outcome positives. An advantage of devogs is that the number of false-positives is low, despite the reduction in sensitivity under low-performance conditions.

The *t*-test used in the devogs analysis has several underlying assumptions. The first is that the data follow a standard normal distribution under the null hypothesis. The distribution of ω is positively skewed as ω itself is a ratio value of the non-synonymous rate divided by the synonymous rate, and genes are typically under purifying selection. We adjusted the

skewness by $\log_2$ transformation to approximate the normal distribution (Fig. S2) to fulfill this assumption, although the *t*-test is sufficiently robust to moderate violations of the normality assumption.[25] However, another assumption is that the data used to perform the *t*-test should be sampled independently from the two populations being compared. The pairwise ω ratio is calculated on the branch across the two species. However, $\omega_{*a(k)}$ and $\omega_{\wedge a(k)}$ which are compared by *t*-tests, share some branches between the two groups, as shown in Figure 1. This violates the assumption of independency in the *t*-test and may reduce devogs sensitivity. If rapid evolution occurred in the branch, both means of the two ω-ratio groups would increase; conversely, if evolution were suppressed, both means of the two ω-ratio groups would decrease. This would reduce the differences in means between the two groups, which may reduce the discrimination of *t*-tests and lower the sensitivity of devogs. Alternatively, the Mann Whitney U-test (a non-parametric test) can be applied. We generated receiver operating characteristic and PPV curves using the *P*-value (Figs. S4 and S5).

Devogs analysis of real data

A total of 50 genes showed accelerated evolution for humans compared with the other four species when analyzing real data with devogs. Terms related to RNA polymerase II were significantly enriched from gene enrichment analysis with this set of 50 genes. Developmentally complex organisms do not appear to be distinguished by the total number of genes they encode, but rather by the number of ways these genes can be expressed and controlled. The surprisingly small number of genes found in the human genome[26] illustrates the importance of evolutionary advances in the control of gene expression. RNA polymerase II in animals has a very important role in mediating alternative splicing of exon junctions to produce different tissue-specific or developmentally specific products from the same gene. The C-terminal domain of RNA polymerase II binds the mediator that transduces control signals to the polymerase II promoter complex, as well as recruits serine-arginine proteins and other splicing factors to the elongating message.[27] Rapid evolution of genes related to the RNA polymerase promoter or RNA elongation may be related to the observation that approximately 40% of genes in the human genome are subject to such alternative splicing, resulting in a more than a three-fold increase in the complexity of gene products over gene content.[28] Similar trends between devogs and the branch model with actual data are observed when compared with our virtual data analysis (Fig. 4). The branch model predicted many accelerated genes, and non-overlapped genes were primarily identified using the branch model. The ratio of true-positives may be higher in overlapped genes when we refer our virtual data analysis (Fig. 4). Additionally, most genes (39 of 50) overlapped between devogs and the branch model, and the concordance of genes between devogs and branch model increased as the *P*-value of the branch model decreased.

Devogs can be used to complement the branch model method and yields reliable results under marginal conditions but has low sensitivity, such as short evolutionary distances between lineages, and makes no assumptions regarding phylogenetic relationships. The branch model is accurate with well-defined phylogenetic tree structures and longer distances between lineages. However, the devogs method currently corresponds only to the branch model. Therefore, further studies are required to optimize devogs and apply it to detect positive selection in sites level with particular lineages corresponding to the branch-site model.[29,30]

Author Contributions

Analyzed the data: K-WK, HK. Wrote the first draft of the manuscript: K-WK. Contributed to the writing of the manuscript: DWB, SC. Agree with manuscript results and conclusions: K-WK, HK, DWB, SC. Made critical revisions and approved final version: K-WK, SC, DWB. All authors reviewed and approved of the final manuscript.

Funding

This work was supported by a grant (PJ009019, PJ009032) from Next-Generation BioGreen 21 Program, Rural Development Administration, Republic of Korea.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with

ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Miyata T, Yasunaga T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol.* 1980;16(1):23–36.
2. Messier W, Stewart CB. Episodic adaptive evolution of primate lysozymes. *Nature.* 1997;385(6612):151–4.
3. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol.* 1998;15(5):568–73.
4. Fitch WM, Bush RM, Bender CA, Cox NJ. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci U S A.* 1997;94(15):7712–8.
5. Suzuki Y, Gojobori T. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol.* 1999;16(10):1315–28.
6. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics.* 1998;148(3):929–36.
7. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics.* 2000;155(1):431–49.
8. Kosiol C, Vinar T, da Fonseca R, et al. Patterns of positive selection in six mammalian genomes. *PLoS Genetics.* 2008;4(8):e1000144.
9. Shoshani J, McKenna MC. Higher taxonomic telationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Mol Phylogenet Evol.* 1998;9(3):572–84.
10. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. Molecular phylogenetics and the origins of placental mammals. *Nature.* 2001;409(6820):614–8.
11. Graur D. Towards a molecular resolution of the ordinal phylogeny of the eutherian mammals. *FEBS Lett.* 1993;325(1–2):152–9.
12. Arnason U, Gullberg A, Janke A, Xu X. Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *J Mol Evol.* 1996;43(6):650–61.
13. Wattam AR, Williams KP, Snyder EE, et al. Analysis of ten Brucella genomes reveals evidence for horizontal gene transfer despite a preferred intracellular lifestyle. *J Bacteriol.* 2009;191(11):3569.
14. Dalloul RA, Long JA, Zimin AV, et al. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* 2010;8(9):e1000475.
15. Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. *Nucleic Acids Res.* 2002;30(1):38–41.
16. Ronquist F, Huelsenbeck J. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics.* 2003;19(12):1572–4.
17. Puigbò P, Garcia-Vallvé S, McInerney J. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics.* 2007;23(12):1556–8.
18. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 2007;24(8):1586–91.
19. Thomas J. V Comparative Vertebrate Genomics. *Comparative Genomics: Basic and Applied Research.* 2007:105.
20. Larkin M, Blackshields G, Brown N, et al. Clustal W and Clustal X version 2.0. *Bioinformatics.* 2007;23(21):2947–8.
21. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 2006;34(Web Server issue):W609–12.
22. Castillo-Davis C, Kondrashov F, Hartl D, Kulathinal R. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res.* 2004;14(5):802–11.
23. Peacock CS, Seeger K, Harris D, et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. *Nat Genet.* 2007;39(7):839–47.
24. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2008;4(1):44–57.
25. Sawilowsky SS, Blair RC. A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin.* 1992;111(2):352–60.
26. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001;291(5507):1304–51.
27. Stiller JW, Hall BD. Evolution of the RNA polymerase II C-terminal domain. *Proc Natl Acad Sci U S A.* 2002;99(9):6091–6.
28. Brett D, Hanke J, Lehmann G, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* 2000; 474(1):83–6.
29. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* Dec 2005;22(12):2472–9.
30. Yang Z, dos Reis M. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol.* 2011;28(3):1217–28.

Supplementary Data
Supporting information

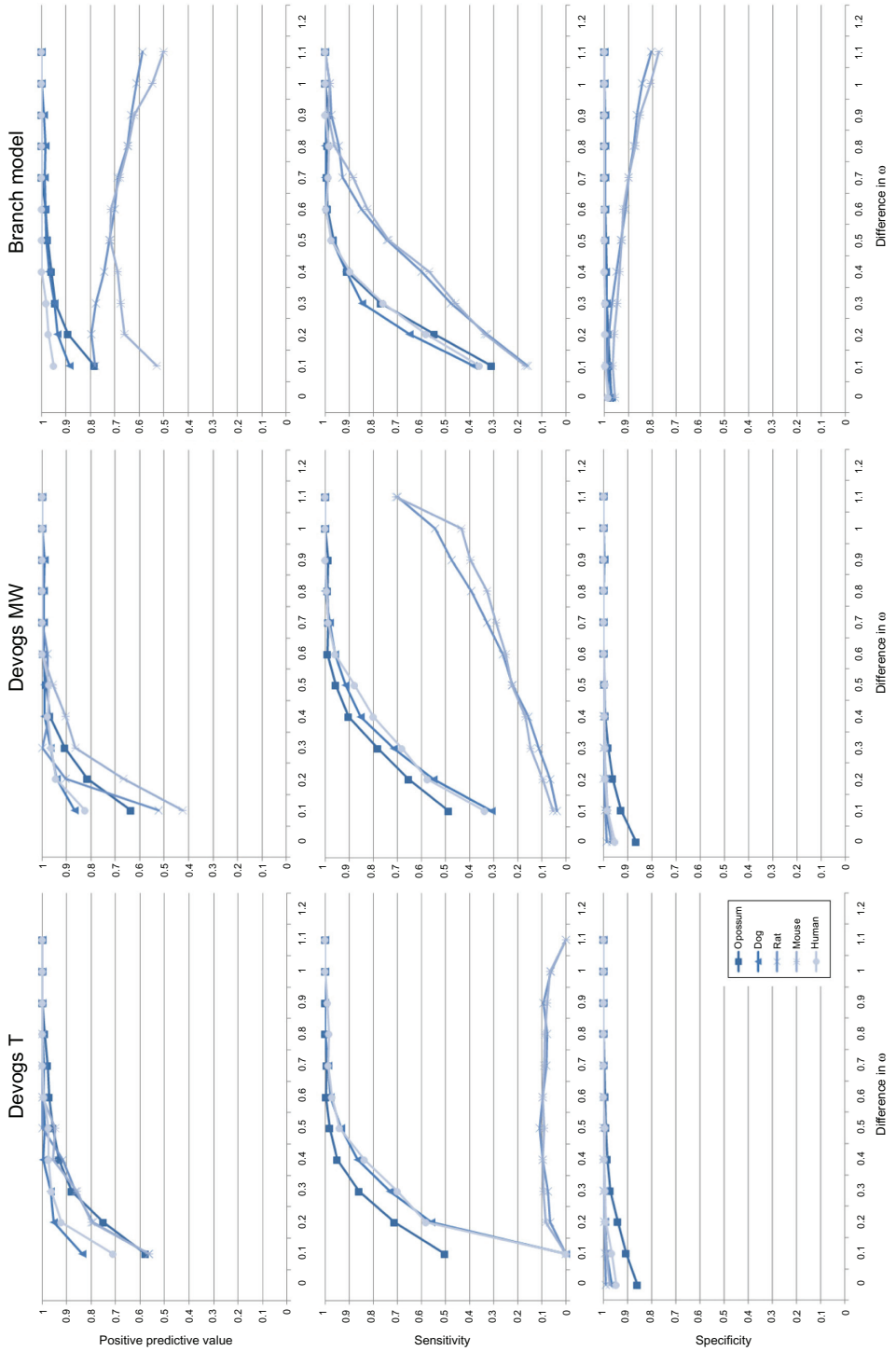


Figure S1. Statistical assessment of the performance of two types of *devogs* and the branch model based on differences in the ω ratio between foreground and background lineages from virtually accelerated genes through simulation. PPV and specificity cannot be defined at a 0 difference. No data at a difference of 1.2 were observed after screening out those with dS $>$ 3. Devogs T: *devogs* with *t*-test. Devogs MW: *devogs* with Mann-Whitney U-test.

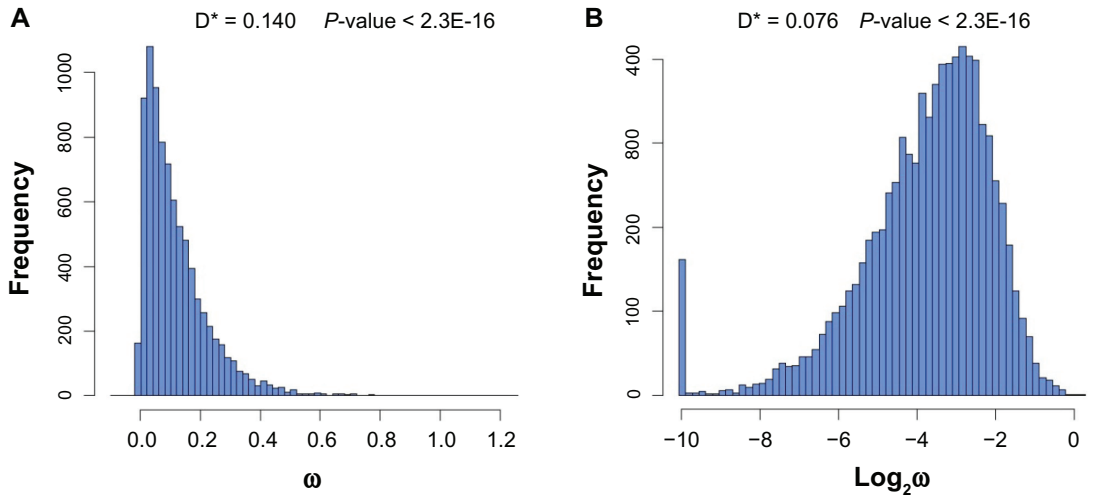


Figure S2. Log_2 transformation of the ω ratio between 8,407 orthologous genes from human and mouse. Orthologous genes with dS $>$ 3 or $\omega > 5$ were filtered. (**A**) is the distribution of the ω ratio before transformation (mainly positively skewed), and (**B**) is that after transformation.
Note: *Lilliefors test statistics for normality.

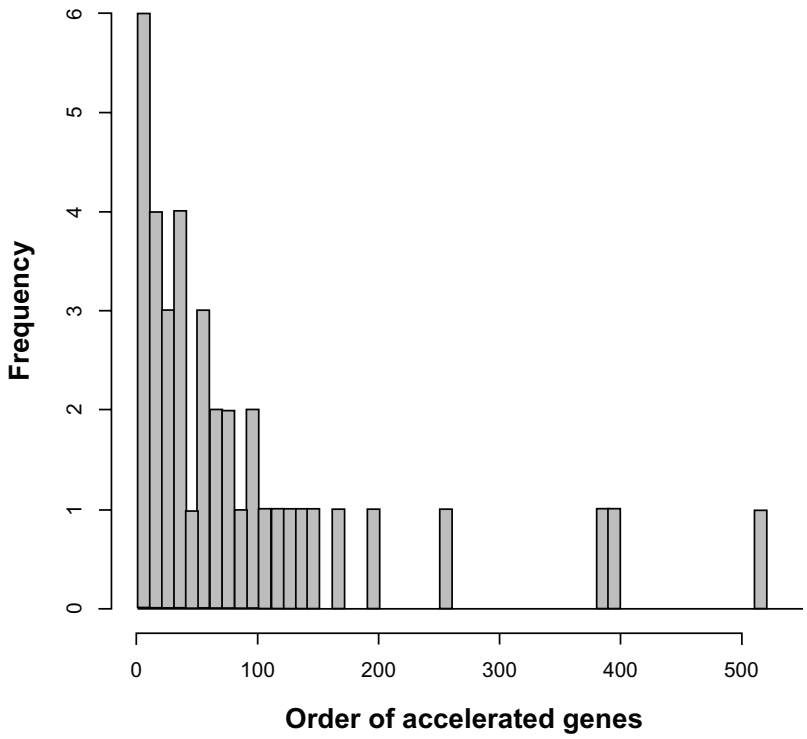


Figure S3. Distribution of co-accelerated genes from both *devogs* and the branch model with real data. X-axis represents the order of accelerated genes from low to high FDR values, while the Y-axis represents the number of accelerated genes from both *devogs* and the branch model.

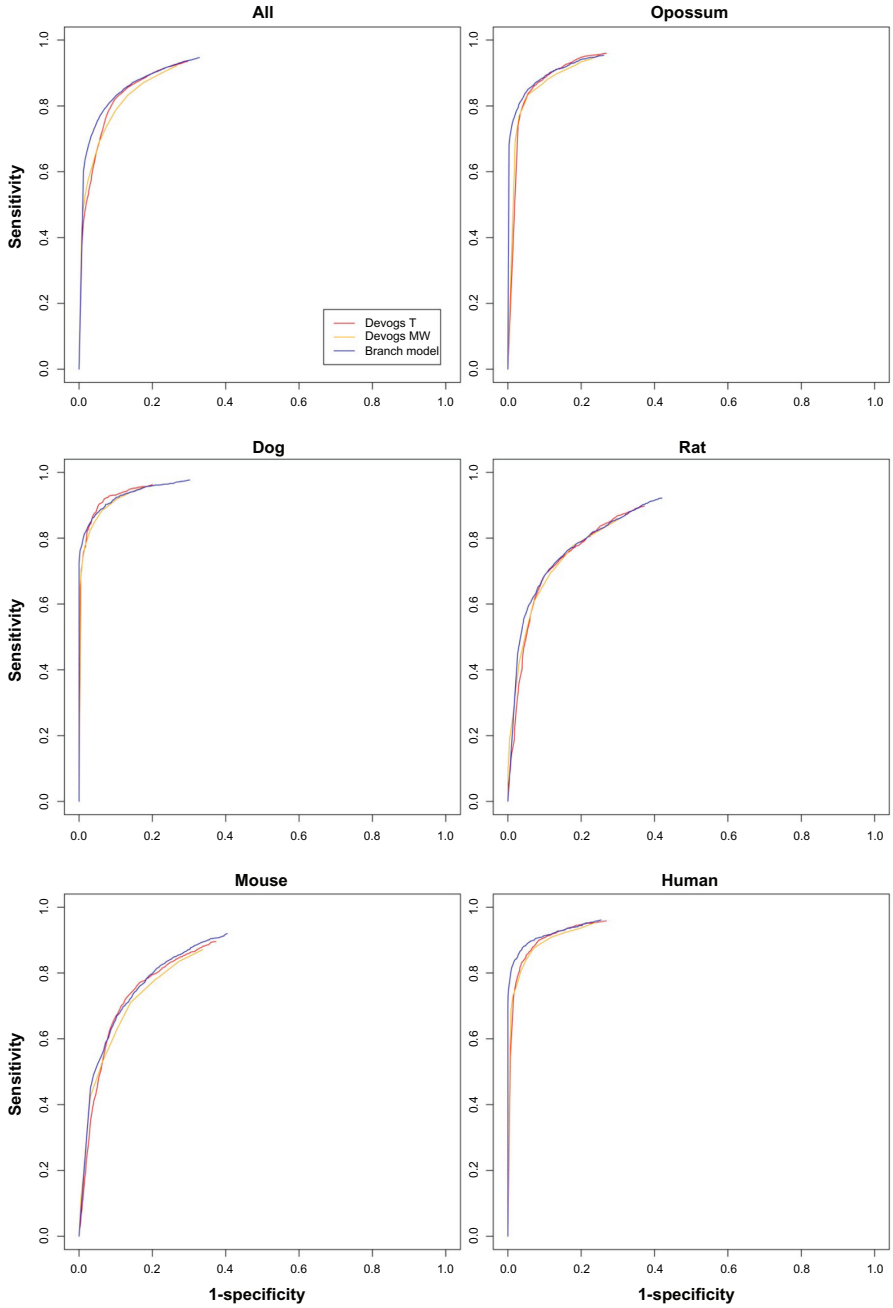


Figure S4. ROC curves by *P*-value of all and each species of *devogs* with *t*-test, *devogs* with Mann-Whitney U test and the branch model. Devogs T: *devogs* with *t*-test. Devogs MW: *devogs* with Mann-Whitney U-test.

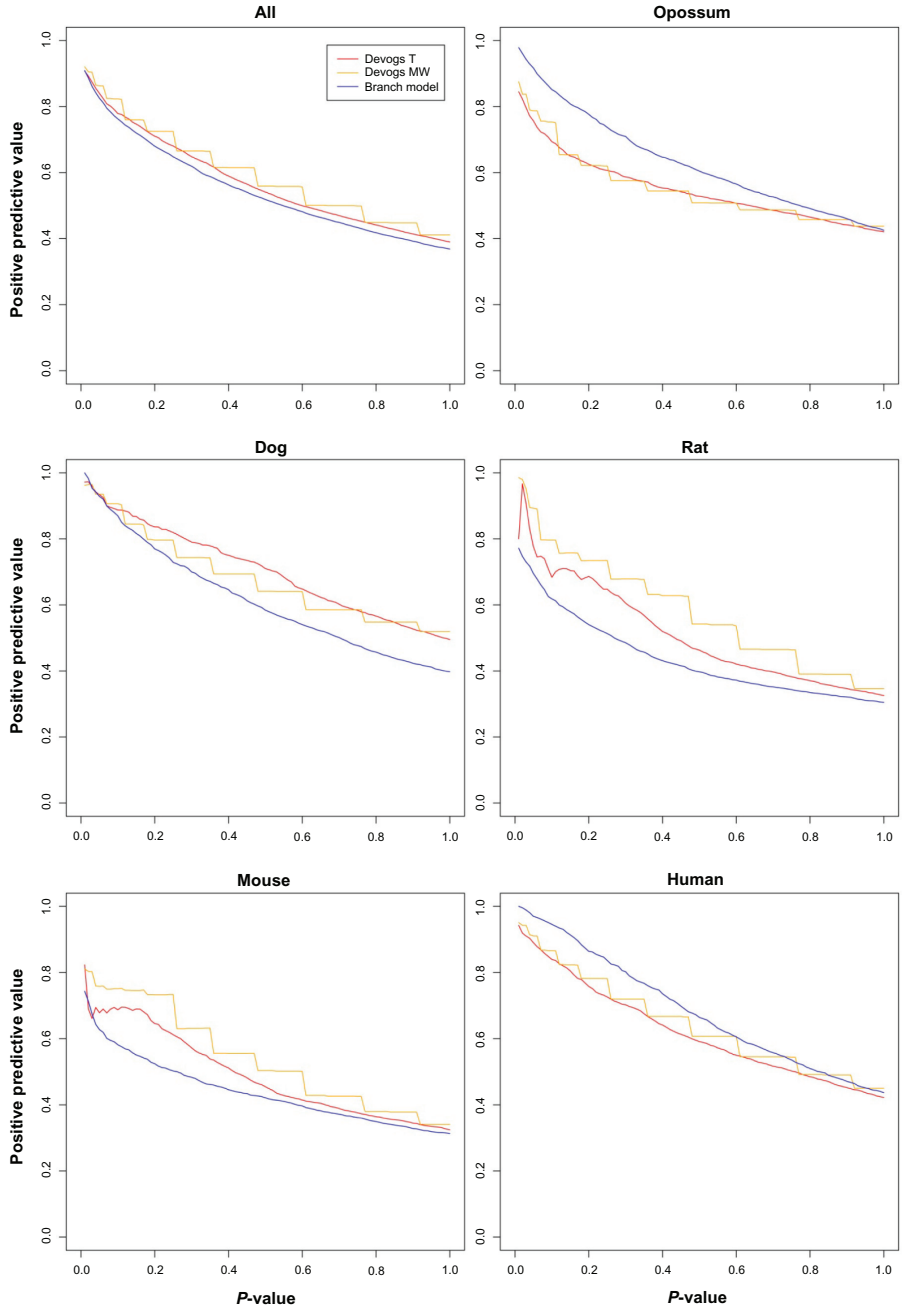


Figure S5. Positive predictive value by *P*-value of all and each species of *devogs* with *t*-test, *devogs* with Mann-Whitney U test and the branch model. Devogs T: *devogs* with *t*-test. Devogs MW: *devogs* with Mann-Whitney U-test.

Table S1. Statistical assessments of the performance of *devogs* and the branch model for simulated data.

Approaches	PPV*	Sensitivity	Specificity
Devogs *t*-test	0.84	0.50	0.98
Devogs Mann-Whitney U	0.86	0.53	0.98
Branch	0.82	0.70	0.97

Note: *Positive predictive value.

Table S2. Accelerated human genes. The lists are ordered by (a) to (b). (a) Average ω of group pairing with humans. (b) Average ω of group pairing without humans.

Gene symbol	(a) average of ω_{*human}	(b) average of $\omega_{\wedge human}$	(a)-(b)	FDR
NAA30	0.603	0.160	0.443	2.0.E-03
ETV6	0.297	0.001	0.296	3.5.E-04
LEPROTL1	0.321	0.039	0.282	4.1.E-03
TM4SF19	0.477	0.240	0.236	3.3.E-03
SMEK1	0.229	0.008	0.221	7.5.E-03
HTRA4	0.343	0.126	0.217	4.3.E-02
CYFIP2	0.196	0.001	0.195	3.5.E-04
C7orf59	0.235	0.042	0.194	4.1.E-03
FAM18A	0.255	0.073	0.182	9.3.E-03
MPDU1	0.313	0.143	0.170	1.6.E-02
FBXL17	0.192	0.029	0.162	8.8.E-03
IFT20	0.165	0.013	0.151	4.6.E-02
C9orf114	0.224	0.080	0.144	9.3.E-03
SCG5	0.201	0.059	0.142	1.7.E-02
IKBKB	0.164	0.026	0.139	4.1.E-03
CHEK1	0.201	0.065	0.136	4.6.E-02
POLR2H	0.136	0.001	0.135	8.5.E-04
NAA60	0.158	0.027	0.131	1.5.E-02
POLR2F	0.129	0.007	0.122	5.6.E-04
P4HA3	0.204	0.084	0.120	9.3.E-03
SYVN1	0.163	0.058	0.105	3.6.E-02
GNB2L1	0.102	0.001	0.101	8.5.E-04
PRRX1	0.104	0.004	0.100	2.0.E-03
PTDSS1	0.137	0.039	0.098	3.6.E-02
ST6GALNAC2	0.230	0.153	0.077	3.3.E-02
XKR6	0.091	0.016	0.076	3.6.E-02
NPLOC4	0.090	0.018	0.072	2.0.E-03
IFT74	0.123	0.052	0.071	3.3.E-02
FYN	0.071	0.002	0.068	1.8.E-02
S1PR1	0.093	0.025	0.068	3.5.E-02
MRS2	0.152	0.084	0.068	3.7.E-02
PPIE	0.098	0.032	0.066	2.1.E-02
CNOT4	0.113	0.049	0.064	2.8.E-03
BRE	0.075	0.013	0.062	9.3.E-03
CDH8	0.081	0.019	0.061	1.8.E-02
CRIM1	0.120	0.062	0.058	3.9.E-03
PPP2R4	0.102	0.047	0.055	1.8.E-02
CHMP3	0.073	0.020	0.053	1.8.E-03
PDGFD	0.076	0.036	0.039	9.3.E-03
KAT8	0.031	0.002	0.029	9.3.E-03
CSNK2A1	0.053	0.024	0.028	2.6.E-02
KLHL5	0.045	0.019	0.027	4.6.E-02
ARHGEF7	0.062	0.039	0.023	4.1.E-03
OSBPL3	0.077	0.057	0.020	2.7.E-02
FAT1	0.078	0.068	0.009	3.3.E-03
GABRR1	0.030	0.021	0.009	4.6.E-02
TECTA	0.034	0.025	0.009	9.3.E-03
ARCN1	0.022	0.015	0.008	2.1.E-02
TNNI1	0.011	0.005	0.006	3.1.E-02
TSPAN12	0.022	0.017	0.005	4.0.E-02

Table S3. Human genes accelerated by both *devogs* and the branch model. (ω_f: foreground ω, ω_b: background ω)

Gene symbol	ω_f	ω_b	ω_f-ω_b	2ΔlnL	FDR
GNB2L1	0.21713	0.0001	0.21703	142.9604	3.21E-29
FYN	0.17872	0.00142	0.1773	117.2497	6.77E-24
NPLOC4	0.21826	0.01957	0.19869	88.33955	5.90E-18
SMEK1	0.54245	0.00902	0.53343	84.75537	2.58E-17
CDH8	0.27516	0.02168	0.25348	81.12733	1.13E-16
FBXL17	0.48817	0.02551	0.46266	79.07555	2.91E-16
CHEK1	0.36934	0.05885	0.31049	76.26699	1.02E-15
HTRA4	1.12494	0.13605	0.98889	67.98521	5.51E-14
SYVN1	0.38628	0.06539	0.32089	66.44629	1.07E-13
S1PR1	0.2133	0.02751	0.18579	58.64361	4.81E-12
PRRX1	0.24999	0.00515	0.24484	57.70917	7.39E-12
KAT8	0.07173	0.00218	0.06955	57.36686	8.41E-12
NAA60	0.323	0.03072	0.29228	51.50981	1.27E-10
POLR2F	0.31863	0.0089	0.30973	49.44172	3.42E-10
P4HA3	0.42905	0.08903	0.34002	46.80893	1.20E-09
CRIM1	0.25088	0.06587	0.18501	42.37855	1.03E-08
LEPROTL1	0.57632	0.04803	0.52829	42.20085	1.10E-08
IKBKB	0.35023	0.02784	0.32239	38.98667	4.66E-08
BRE	0.17412	0.01632	0.1578	35.75642	2.17E-07
KLHL5	0.1485	0.01919	0.12931	35.47817	2.46E-07
SCG5	0.49508	0.06306	0.43202	35.20808	2.73E-07
IFT74	0.31864	0.06139	0.25725	32.71837	8.70E-07
ETV6	0.58307	0.0001	0.58297	30.92364	2.02E-06
FAM18A	0.46072	0.07708	0.38364	29.57495	4.00E-06
XKR6	0.32755	0.01835	0.3092	28.93148	5.35E-06
PPIE	0.2062	0.03426	0.17194	27.10987	1.20E-05
CNOT4	0.25645	0.04921	0.20724	25.9242	2.06E-05
MRS2	0.4065	0.08935	0.31715	25.77011	2.19E-05
PPP2R4	0.24671	0.0505	0.19621	22.94705	8.18E-05
MPDU1	0.48702	0.14541	0.34161	22.23522	0.000112
C7orf59	0.42855	0.05606	0.37249	20.81865	0.00021
CHMP3	0.16945	0.02443	0.14502	20.59236	0.000229
NAA30	0.71871	0.1542	0.56451	19.72839	0.000325
TM4SF19	0.76833	0.238	0.53033	18.07518	0.000673
ARHGEF7	0.13243	0.03913	0.0933	16.36756	0.001446
TECTA	0.06032	0.02751	0.03281	14.23274	0.003388
CSNK2A1	0.11741	0.02413	0.09328	10.26646	0.018676
PTDSS1	0.26073	0.05253	0.2082	9.925722	0.021687
OSBPL3	0.15242	0.06439	0.08803	8.251596	0.041802