The impacts of drift and selection on genomic evolution in insects

K. Jun Tong[1,*], Sebastián Duchêne[1,2,*], Nathan Lo[1] and Simon Y.W. Ho[1]

[1] School of Life and Environmental Sciences, University of Sydney, Sydney, New South Wales, Australia
[2] Centre for Systems Genomics, University of Melbourne, Melbourne, Victoria, Australia
[*] These authors contributed equally to this work.

ABSTRACT

Genomes evolve through a combination of mutation, drift, and selection, all of which act heterogeneously across genes and lineages. This leads to differences in branch-length patterns among gene trees. Genes that yield trees with the same branch-length patterns can be grouped together into clusters. Here, we propose a novel phylogenetic approach to explain the factors that influence the number and distribution of these gene-tree clusters. We apply our method to a genomic dataset from insects, an ancient and diverse group of organisms. We find some evidence that when drift is the dominant evolutionary process, each cluster tends to contain a large number of fast-evolving genes. In contrast, strong negative selection leads to many distinct clusters, each of which contains only a few slow-evolving genes. Our work, although preliminary in nature, illustrates the use of phylogenetic methods to shed light on the factors driving rate variation in genomic evolution.

Subjects Computational Biology, Evolutionary Studies, Genomics
Keywords Mutation, Genomic pacemakers, Molecular evolution, Neutral theory, Insect phylogenomics

INTRODUCTION

Molecular evolution proceeds by the fixation of mutations, a process that balances stochastic drift against natural selection. The relative importance of these two forces depends on population size (*Ohta, 1992*) and on the distribution of fitness effects of new mutations (*Eyre-Walker & Keightley, 2007*). When mutations have neither a beneficial nor detrimental impact on fitness, their fate is determined entirely by the stochastic process of genetic drift (*Kimura, 1968*). In contrast, purifying selection removes deleterious mutations over time. Selection is more efficient in large populations, where even small differences in selection coefficients can substantially change the relative probability of any particular mutation becoming fixed (*Ohta, 1992*). In small populations, mutations with small fitness effects behave similarly to neutral mutations, so drift tends to be more important.

Drift and selection tend to have different impacts on evolutionary rates, leading to patterns of rate variation that can be detected using phylogenetic methods (Fig. 1). Furthermore, different genes are subject to varying degrees of selective constraint, leading to measurable disparities in evolutionary rates. For example, functionally important genes tend to evolve slowly because many of the encoded amino acids are under strong selective constraint (*Dickerson, 1971*). A simple way to detect these "gene effects" is to examine

Submitted 5 December 2016
Accepted 28 March 2017
Published 27 April 2017

Corresponding author
K. Jun Tong, jun.tong@sydney.edu.au

Academic editor
William Amos

Additional Information and Declarations can be found on page 12

DOI **10.7717/peerj.3241**

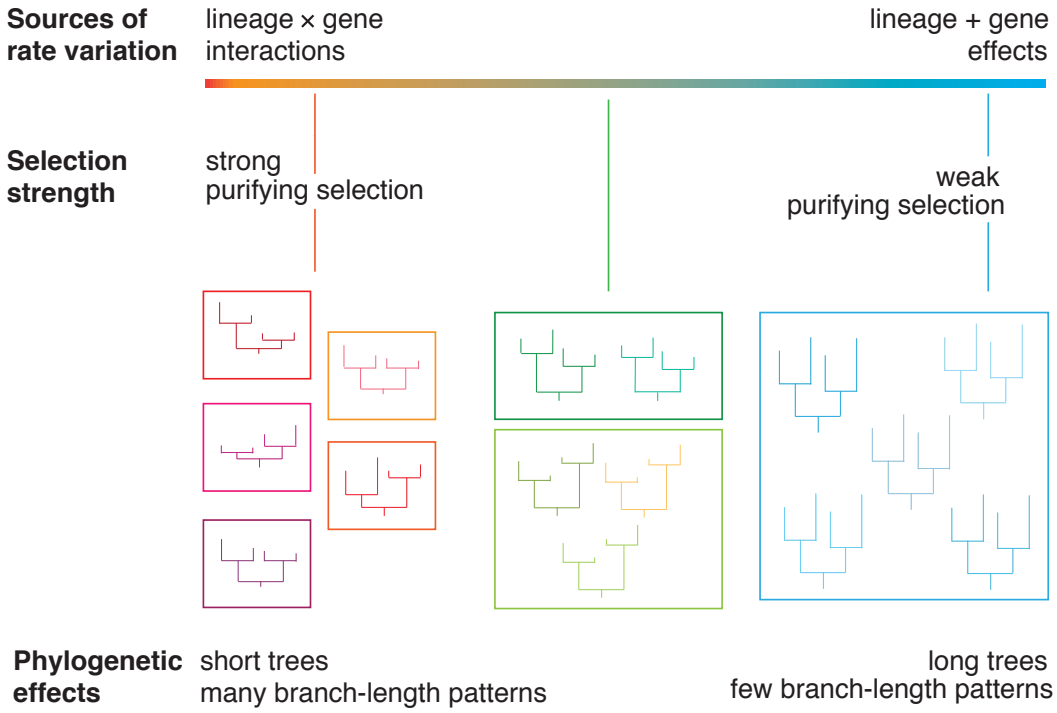

Figure 1 A diagram illustrating the relationship between evolutionary rate and phylogenetic branch-length clusters. Genes that are under strong purifying selection have low rates of evolution, producing short phylogenetic trees. Genes whose evolution is dominated by drift have long phylogenetic trees. We posit that these genes will group into a small number of clusters of branch-length patterns. These genes are primarily subject to lineage effects, which act on a whole-genome scale. In contrast, genes under strong purifying selection experience gene-by-lineage interactions, which lead to distinctive patterns of among-lineage rate variation across genes. These genes will be dispersed into many separate clusters.

the branch lengths of the gene trees. Genes that are subject to weak selective constraints are expected to yield trees with longer branches, representing a larger total amount of genetic change. In contrast, when genetic change is retarded by purifying selection, genes are expected to yield trees with shorter branches.

The relative impacts of drift and selection also vary across lineages, depending on population size (*Ohta, 1992*). For example, species with small populations are expected to evolve rapidly because of the dominance of genetic drift (*Ohta, 1987*). In addition, differences in life-history traits, such as generation time, can produce rate heterogeneity among branches in the tree (*Bromham, 2009*). Genes that are subject to the same "lineage effects" share the same pattern of relative branch lengths across the tree (*Ho & Duchêne, 2014*).

Gene and lineage effects can interact to produce "residual effects" (*Gillespie, 1991*; *Muse & Gaut, 1997*). Consider two genes, *A* and *B*, sampled from two taxa, *x* and *y*. Both genes are responsible for important biological functions, such that their evolution is constrained. However, gene *A* is under stronger purifying selection in taxon *x* than in taxon *y*. Gene *B* is subject to the reverse conditions, with weaker purifying selection in taxon *x* and stronger selection in taxon *y*. As a consequence, the tree for gene *A* has a longer branch leading to

taxon *y* and a shorter branch leading to taxon *x*, whereas the converse is true in the tree for gene *B*. Thus, the trees for these two genes display disparate branch-length patterns.

On a genomic scale, there might be many different patterns of among-lineage rate variation (*Ho, 2014*; *Snir, 2014*). These can be identified by statistically clustering gene trees according to their branch-length patterns (*Duchêne & Ho, 2015*; *Duchêne, Foster & Ho, 2016*). Each distinct cluster of gene trees identified by this method represents a group of genes that have been subject to a particular combination of gene effects and lineage effects. In the terminology used in previous studies, these genes can be regarded as being governed by the same "pacemaker" (*Snir, Wolf & Koonin, 2012*; *Wolf, Snir & Koonin, 2013*). However, this term implies that groups of genes are subject to an underlying evolutionary driving force. Here we simply refer to these groups as clusters of genes that share the same branch-length patterns.

We expect interactions between gene effects and lineage effects to be more common under conditions of selection, because the strength and direction of selection is unlikely to be uniform across species. Therefore, we predict that genes under strong selection will group into many clusters and yield trees with short branches (Fig. 1). In contrast, we predict that genes that are under much weaker selection will group into few branch-rate clusters and yield trees with long branches. Under these conditions, most rate variation is due to lineage effects, such as those caused by differences in generation time. These lineage effects act on a genome-wide scale (*Gillespie, 1991*), such that different genes share the same pattern of branch-length variation.

In light of the relationships described above, we predict that the clustering of genes according to their branch-length variation is associated with evolutionary rates (Fig. 1). We hypothesize an observable link between evolutionary rate and the dispersion of phylogenetic patterns. This prediction can be tested by analysing genomic data using a phylogenetic approach, because drift and selection leave different signatures in the gene trees. Here we analyse 955 genes from 15 species: two hemimetabolous insects and 13 holometabolous insects. The latter group of insects undergo complete metamorphosis as part of their development. Holometabola arose more than 350 million years ago (*Tong et al., 2015*) and is extraordinarily diverse: its members include those that are eusocial (*Wilson & Hölldobler, 2005*), parasites (*Libersat, Delago & Gal, 2009*), long-distance migrators (*Chapman, Reynolds & Wilson, 2015*), and ecological engineers (*Losey & Vaughan, 2006*). They represent a large proportion of the global biomass and are responsible for the bulk of ecological functions on land. We find that as evolutionary rate increases, genes are assigned to fewer branch-length clusters. The results of our analyses point to a general trend that can be tested using genomic data from other groups of organisms.

METHODS

We used maximum likelihood to infer the phylogeny of 15 species of insects (Table S1A). Our data set is based on that analysed by *Peters et al. (2014)*, which originally comprised 1,343 amino acid sequences from 88 species. We filtered this data set in order to remove missing data, producing a subset of 955 amino acid sequences from 15 species

(Table S1B). The insects in our analysis are: two bees (*Apis mellifera* and *Bombus terrestris*), two ants (*Linepithema humile* and *Pogonomyrmex barbatus*), a wasp (*Nasonia vitripennis*), three mosquitos (*Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*), three flies (*Drosophila melanogaster*, *Drosophila persimilis*, and *Drosophila sechellia*), a beetle (*Tribolium castaneum*), the silkworm (*Bombyx mori*), a louse (*Pediculus humanus*), and an aphid (*Acyrthosiphon pisum*).

Because we were interested in the relationship between tree length and branch-length patterns, our analyses required the topologies of the gene trees to be mutually congruent. We checked for any substantial differences in topologies between gene trees by clustering them using the *k*-means Partitioning Around Medoids (PAM) algorithm (*Kaufman & Rousseeuw, 2009*). This method looks for dissimilarity in the data and characterizes variation using representative medoids. In our approach, we considered the pairwise distances between topologies based on the PH85 metric (*Penny & Hendy, 1985*). We represented these distances in two-dimensional space using multidimensional scaling. As such, each data point corresponds to a gene tree. The algorithm consists of randomly selecting k of n data points as the medoids. This procedure is repeated until the assignment of the data points to medoids does not change. Because this method does not automatically determine the optimal number of clusters, we calculated a measure of goodness of fit, the Gap statistic (described by *Tibshirani, Walther & Hastie, 2001*), for a range of values of k (1–50), and we selected the k with the highest Gap value.

We found strong support for a single cluster of tree topologies, whereby every gene supported the same set of evolutionary relationships among the 15 species of insects. Our topology matches that of previously published insect phylogenies (e.g., *Peters et al., 2014*; *Misof et al., 2014*). Accordingly, we inferred the maximum-likelihood tree from a data set comprising the 955 genes in concatenation, using RAxML v8.1 (*Stamatakis, 2014*). Based on this estimate of the tree topology, we optimized the branch lengths for each gene. Thus, the resulting gene trees shared the same topology but had their own sets of maximum-likelihood branch lengths. The same substitution model, GTR+G with four categories of site rates, was used to estimate the branch lengths for each gene tree. We ran ten replicates of each search and chose the tree with the highest likelihood score.

Using these data, we first tested the assumption that evolutionary rates are associated with the strength of purifying selection. To do this, we determined the relative average rate in each gene by taking the sum of the expected number of substitutions along all of the branches in the corresponding gene tree (i.e., the tree length). We then plotted gene-specific ratios of radical and conservative amino acid substitutions, referred to as the Kr/Kc ratio, against the lengths of the corresponding gene trees. The Kr/Kc ratio is a commonly used and is a robust indicator of selection pressure (*Hughes & Nei, 1988*; *Hanada, Shiu & Li, 2007*). It is calculated from protein data, so it is more resistant to the impacts of sequence saturation than Dn/Ds, the equivalent ratio for nucleotide data (*Smith & Smith, 1996*). Our method of identifying radical and conserved substitutions is similar to that of *Zhang (2000)*. We used a model-free, non-parametric approach to estimate this ratio. This statistic has a similar interpretation to the Kr/Kc ratio (*Zhang, 2000*), but the absolute values are expected to be different because Kr/Kc is estimated using an explicit substitution model and phylogenetic

tree. Although our method is biased towards radical substitutions, with a consequent skew in our results, it provides a fast estimate of the degree of selection. Regardless of the absolute values of Kr/Kc, we consider that our comparison is valid because our data set contains the same set of genes for all taxa, such that they have evolved over the same timescale and are expected to have similar levels of saturation at sites under weak selective constraints. The code for this method is available at www.github.com/kjuntong/tree-length.

We then tested for a relationship between evolutionary rate and the clustering of genes by their branch-length patterns. We assigned genes to clusters by grouping them according to their branch-length patterns using a Gaussian mixture model (GMM) clustering algorithm from the Python machine learning toolkit, Scikit-learn (*Pedregosa et al., 2011*). GMM algorithms assign data to multivariate normal components and appear to work well when used to identify clusters of branch-length patterns (*Duchêne, Foster & Ho, 2016*). Importantly, this method clusters the gene trees by their pattern of branch lengths (lineage effects), but not their overall relative evolutionary rate (gene effects). In this study, we did not aim to find the optimal number of clusters for the data; instead, we wished to test our hypothesis using different numbers of clusters. Therefore, we compared the results obtained using different numbers of clusters, from five to 100. We expect that a scheme with few clusters would not provide sufficient resolution to allow us to detect the hypothesized relationship, whereas a scheme with many clusters relative to the total number of genes carries the risk of overfitting. The gene trees were ranked according to evolutionary rate, as denoted by tree length, and divided into deciles. For each tree-length decile, we identified the number of clusters of branch-length patterns that were represented, and plotted these results in a histogram. We then performed a Kendall rank correlation test to measure the association between evolutionary rate and cluster size.

We tested whether our clustering method carried the risk of producing a spurious relationship between tree length and number of clusters of branch-length patterns. To do this, we simulated three sets of 300 gene trees based on the median tree length, the 10th percentile of tree lengths, and the 90th percentile of tree lengths from the insect data. Each tree was generated by simulation according to one of 40 branch-length patterns. Sequence evolution was simulated on each tree to produce an alignment of 353 amino acids, the mean sequence length of the loci in our insect data set. If our method is not biased, we expect that each of the three tree-length categories (short, medium, and long) will contain the same number of branch-length patterns.

In addition to testing the role of evolutionary rate, we investigated whether the clustering of genes by their branch-length patterns could be explained by gene function. Our data set is poorly annotated, which is typical of large data sets generated by high-throughput sequencing. This limited the scope of our investigation to enzymes because enzyme commission (EC) numbers were available for only a subset of our data. EC numbers refer to particular catalytic processes that are enabled by the enzymes. These classifications were available for 297 genes in our data set, but other genes either had incomplete annotations or did not encode enzymes. We looked at the number of clusters of branch-length patterns represented for each of six EC numbers. To correct for an imbalance in the number of genes

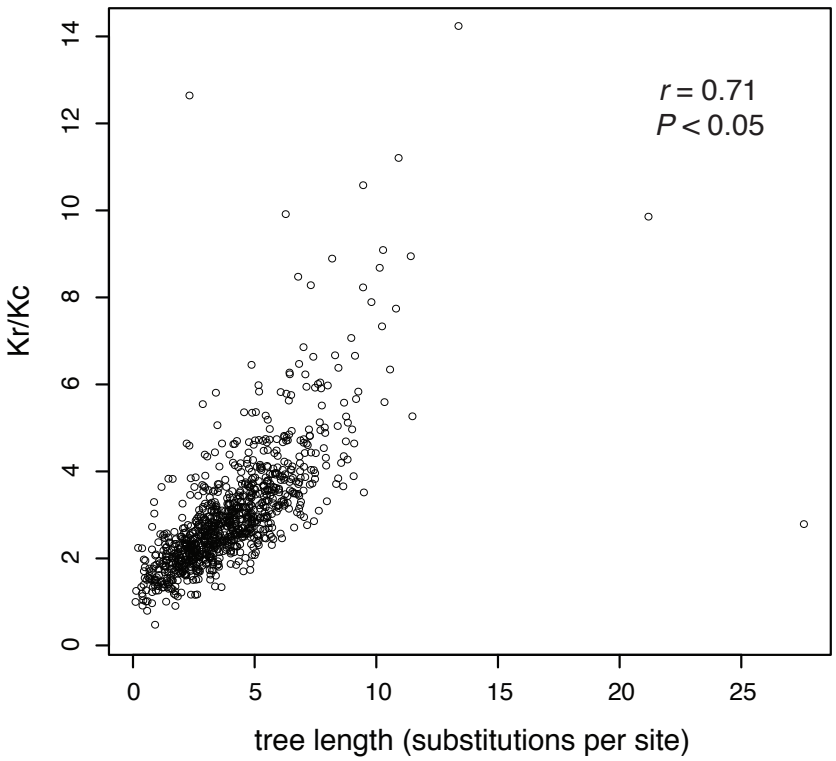

Figure 2 The ratio of estimated radical to non-radical amino acid substitutions (Kr/Kc) shows a positive relationship with evolutionary rate, as measured by gene-tree length. Each point represents one of 955 gene trees. High Kr/Kc values indicate that radical substitutions outnumber non-radical substitutions, reflecting weak selective constraints. Thus, this plot shows that the strength of purifying selection is negatively correlated with evolutionary rate.

within each EC category, we divided the number of represented clusters by the number of genes.

Finally, we fitted a random forest classifier to test whether the tree length, ratio of radical and conserved amino acid substitutions, or EC number could predict the cluster assignments of the genes (*Liaw & Werner, 2002*). Predictive accuracy was quantified using Gini coefficients, a measure of statistical dispersion. A variable with a Gini coefficient of 1 predicts the data perfectly, whereas a coefficient of 0 indicates that the variable is not predictive at all.

RESULTS

In our analysis of 955 amino acid sequences, we first tested the assumption that evolutionary rate is linked to the strength of purifying selection. We found a positive relationship between the ratio of radical and conservative amino acid substitutions and evolutionary rate (as measured by gene-tree length), meaning that more rapidly evolving genes are under weaker purifying selection (Fig. 2).

After confirming the relationship between rate and purifying selection for our data set, we tested our prediction of a relationship between evolutionary rate (tree length) and

the clustering of genes by their branch-length patterns (Fig. 3). Our results confirmed this, showing that slowly evolving genes group in many clusters whereas rapidly evolving genes group in fewer clusters. Kendall rank correlations found significant relationships for five clustering schemes: 20 clusters ($P = 0.0007$, $\tau = -0.86$), 30 clusters ($P = 0.0007$, $\tau = -0.86$), 40 clusters ($P = 0.0029$, $\tau = -0.88$), 50 clusters ($P = 0.0029$, $\tau = -0.75$), and 100 clusters ($P = 0.0003$, $\tau = -0.90$). We also conducted five- and ten-cluster analyses, but these schemes provided a low level of resolution and we were unable to identify a relationship between the evolutionary rate and clusters of branch-length patterns.

Slowly evolving genes yield short phylogenetic trees that are prone to stochastic estimation errors. Such estimation errors have the potential to confound our analysis by artificially producing variegated patterns of branch lengths. However, it is unlikely that the signal we detect is an artefact because the pattern remains apparent even when we exclude the slowest two deciles of genes. When we excluded the slowest two deciles of genes for each clustering scheme, our results are as follows: 20 clusters ($P = 0.0078$, $\tau = -0.79$), 30 clusters ($P = 0.0102$, $\tau = -0.77$), 40 clusters ($P = 0.0237$, $\tau = -0.67$), 50 clusters ($P = 0.044$, $\tau = -0.59$), and 100 clusters ($P = 0.004$, $\tau = -0.84$).

To evaluate the robustness of our clustering method, we repeated our analysis using two additional data sets. We used a 17-taxon data set and a 10-taxon data set that have approximately 250 fewer and more genes than the original data set, respectively. The former consists of 707 genes and the latter 1192 genes, compared with the 955 genes in the original 15-taxon data set. The 17-taxon data set includes the mosquitos *Aedes albopictus* and *Anopheles funestus*, in addition to those represented in the 15-taxon data set (Table S2A). The 10-taxon data set excludes *Drosophila persimilis*, *Drosophila sechellia*, *Bombyx mori*, *Bombus terrestris*, and *Culex quinquefasciatus* from the 15-taxon data set (Table S3A). For both of these data sets, we recovered the same negative relationship between evolutionary rate and the number of clusters of branch-length patterns (Figs. S1A and S1B).

We also investigated potential links between branch-length patterns and gene function. We isolated the genes that coded for enzymes in our 15-taxon data set and found that of this subset, isomerase genes (EC number 2) are more likely to group in the same cluster than the genes assigned to other EC numbers (Fig. 4). In contrast, transferase genes (EC number 5) are represented across many clusters.

Lastly, we fitted a random forest classifier to compare the influence of different factors on the assignment of genes to clusters. We found that the length of the gene tree has the best predictive accuracy, with a Gini coefficient of 0.64, for explaining the cluster assignment of a gene. This was followed by our Kr/Kc ratio and the EC number, with respective coefficients of 0.60 and 0.25. However, the classifier has overall low predictive accuracy, suggesting that more gene features might need to be considered to provide a more complete and satisfying model for cluster assignment. This can be improved in the future with further progress in genome annotation. We hope that our results will form the basis for future investigation into the question of whether evolutionary rate is linked to phylogenetic patterns.

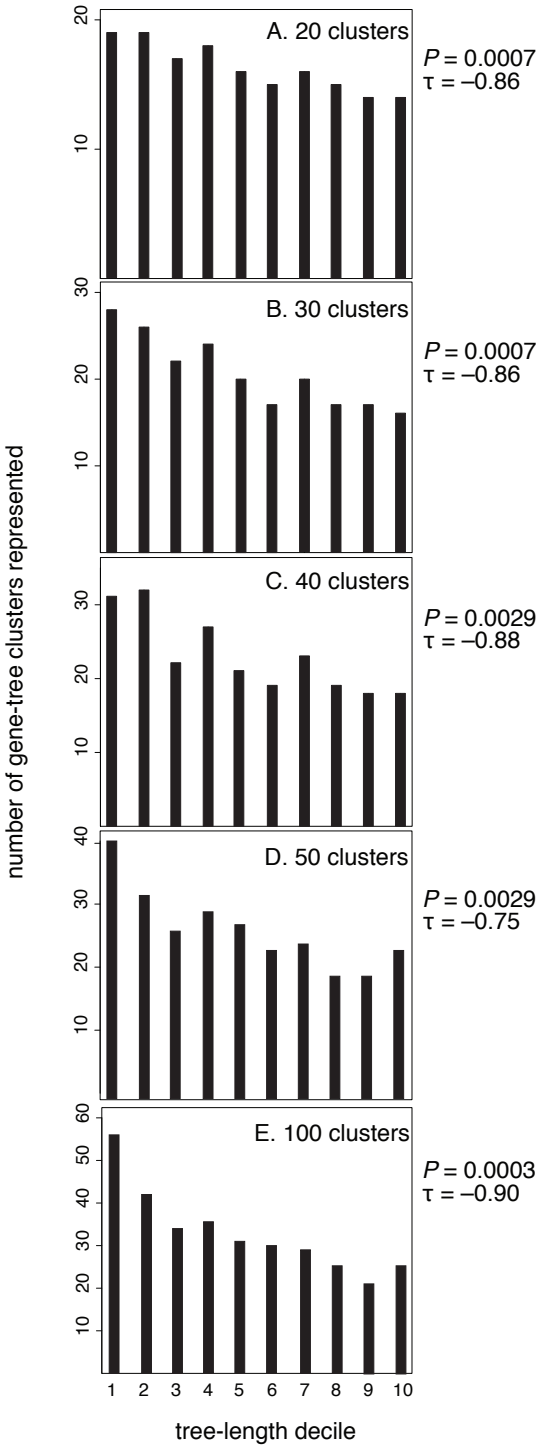

Figure 3 Genes with the longest trees have fewer branch-length patterns than the decile of genes with the shortest trees. Here, each gene has been sorted incrementally into a decile category according to its tree length, where decile 10 contains the longest 10% of trees. Tree length reflects the rate of molecular evolution that has been experienced by a gene, measured in substitutions per site. For each of the five branch-rate clustering schemes (20, 30, 40, 50, and 100 clusters), deciles of genes with higher rates are assigned to fewer clusters.

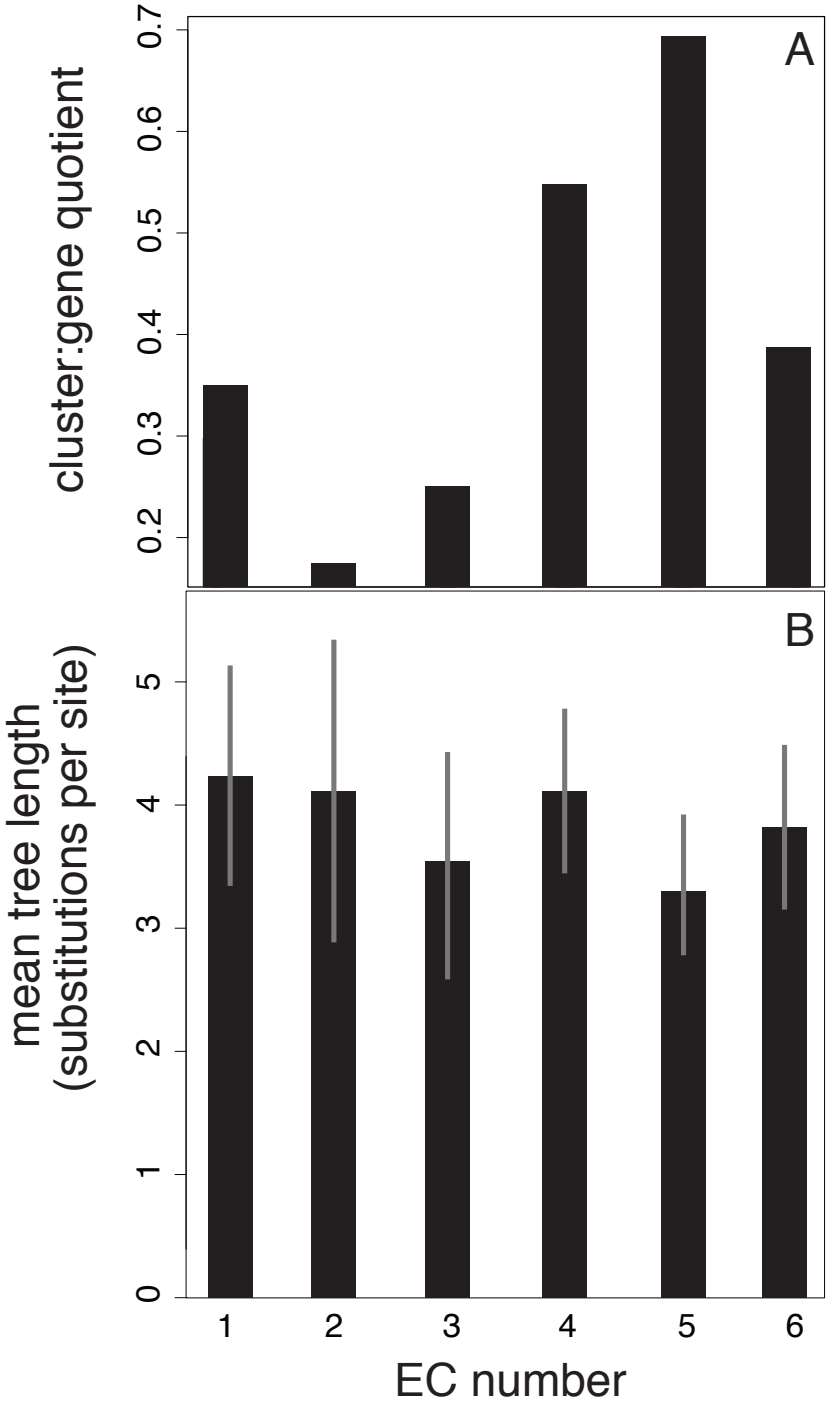

Figure 4 Relationship between EC number and clusters of branch-length patterns for 297 genes in our data set. Each EC number represents a collection of genes that share a common enzyme function. EC number 5, in which genes code for transferases (enzymes that move chemical functional groups from one molecule to another), contains genes with the shortest trees and that are grouped into the largest number of clusters of branch-length patterns. EC number 2 contains genes that are grouped into the smallest number of clusters. EC number 2 codes for isomerases, which are enzymes that convert molecules from one isomer to another. The results are based on a 40-cluster scheme.

DISCUSSION

Evolutionary rate informs structure of branch-length patterns

Our analyses reveal that the clustering of branch-length patterns across genes can be at least partly explained by the competing effects of selection and drift. Genes that are the most weakly selected are subject to the vagaries of drift, and they tend to have the highest evolutionary rates across the genome. The main driver of rate heterogeneity in these genes is lineage effects, which explains our finding that large groups of rapidly evolving genes share the same branch-length patterns. The most well studied cause of lineage effects is that of differences in generation time (e.g., *Thomas et al., 2010*; *Weller & Wu, 2015*). Generation time has a negative relationship with evolutionary rate because genome replication occurs more infrequently in species with long generations than in those with short generations. This is tempered by the fact that long-lived species tend to have small populations, where drift is the dominant driver of molecular evolution and leads to a higher evolutionary rate (*Ohta & Kimura, 1971*). However, theoretical examinations suggest that certain mutagenic conditions allow the fixation of neutral mutations to be independent of population size (*Welch, Eyre-Walker & Waxman, 2008*).

A key problem in our attempt to describe the clustering of branch-length patterns across genes is the effect of fluctuating selection pressures over time. As population sizes change, the effectiveness of selection can increase or decrease. Furthermore, the fitness effects of individual mutations can vary through time, with the potential for selection that is realized under new environmental and ecological conditions (*Dykhuizen & Hartl, 1980*; *Ohta, 1992*). The converse might also be true: as selection dynamics shift, the magnitude of selection acting upon a gene might vary over time. Nevertheless, our phylogenetic approach is able to detect an underlying signal through these various sources of noise.

The clustering of branch-length patterns across genes that we observe here reflects groups of genes that share the same temporal patterns of rate variation. The clusters represented in the most rapidly evolving decile of genes might differ from one another by the shifting balance between selection and drift that has occurred over time. For instance, the sets of genes that display two different branch-length patterns might have experienced the same total amount of evolutionary change due to drift, but differ in the periods of time in which they were subject to selection and drift, thereby generating different branch-length patterns. Our results suggest that in the clusters of the most rapidly evolving genes, these sources of fluctuation are genome-wide factors.

Among the most slowly evolving genes, there is a variety of branch-length patterns because of gene-by-lineage interactions that lead to highly heterogeneous evolutionary rates. Genes that have evolved under these conditions are probably important housekeeping genes, such as those that encode histone or ribosomal proteins. These genes would be subject to strong purifying selection.

Enzyme function and branch-length patterns

The results of our analyses suggest that isomerases are more likely to share the same branch-length pattern compared to other enzymes (Fig. 4). Interestingly, isomerases are more likely to evolve new functions in different EC classes (*Martinez Cuesta et al., 2014*).

Such isomerase sequences might possess latent potential for selection (*Dykhuizen & Hartl, 1980*), whereby long periods of drift produce a stream of raw genetic variation that can be subject to selection under particular conditions (*Ohta, 1987*; *Ohta, 1992*). We speculate that if this is the case, selection is probably occurring at the secondary or tertiary level of protein structure because the trees of the isomerase genes display fewer types of branch-length patterns, indicating that they are subject to little selection pressure at the sequence level. Alternatively, the sharing of branch-length patterns might partly indicate the presence of protein-protein interactions (*Lovell & Robertson, 2010*).

Our investigation of the relationship between enzyme function and clustering of branch-length patterns is limited in its statistical power. Despite our correction for the imbalance in the number of genes represented across the six EC categories, three of the six categories have 13 or fewer genes; these relatively small groups of genes might have had a large bearing on the results (Fig. 4). Further clouding any signal in the data set is the fact that enzyme function can change without substantial alterations to the nucleotide sequence (*Martinez Cuesta et al., 2014*). Enzymes can also exhibit 'promiscuity', whereby they evolve to catalyse new suites of reactions in addition to their normal functions (*O'Brien & Herschlag, 1999*; *Duarte, Amrein & Kamerlin, 2013*). This uncertain correspondence between amino acid changes and biological function, the ultimate target of selection, is potentially a contributor to the statistical noise in our clustering analyses. Amino acid substitutions are also thought to be more insensitive to generation-time effects than nucleotide substitutions, particularly nucleotide changes occurring in non-coding regions, because proteins are more likely to be targets of selection (*Ohta, 1992*).

Implications for phylogenomic analysis

Identifying the relationship between evolutionary rates and branch-length patterns across genes provide some useful insights into how genome-scale data might be handled in phylogenetic analysis. There is a need for new analytical methods to extract phylogenetic and temporal signals from genome-scale data without creating excessive computational demands (*Kumar & Hedges, 2016*; *Tong, Lo & Ho, 2016*). One promising new approach involves data-clustering to identify subsets of genes that share similar evolutionary characteristics (*Duchêne, Molak & Ho, 2014*; *Mirarab et al., 2014*). These techniques have already been used in phylogenomic analyses of mammals (*Dos Reis et al., 2012*), birds (*Jarvis et al., 2014*), and insects (*Misof et al., 2014*).

Our results show that slowly evolving genes tend to yield trees with different patterns of branch lengths. These genes are especially useful for studying ancient divergences because they have experienced less saturation, but they also display greater variation across genes in terms of their among-lineage rate heterogeneity. Therefore, understanding the variation in branch-length patterns across genes has important practical implications for evolutionary dating using molecular clocks (*Duchêne & Ho, 2014*). Any molecular dating study must be based on a compromise between selecting genes with an appropriate rate of evolution, and selecting genes to minimize the variation in patterns of among-lineage rate heterogeneity.

CONCLUSIONS

In summary, our analysis of insects has revealed that the variation in branch-length patterns across genes can be at least partly explained by the different impacts of drift and selection, which produce predictable patterns of rate variation. This is in spite of the noise created by the complexities of the evolutionary process over hundreds of millions of years. The trends that we report here should be understood as an initial demonstration of phylogenetic tools in studying mutation over a vast timescale. Further detailed annotation of genomes and improved methodologies will open the way for deeper insights into the impacts of gene function on shaping phylogenetic information. We also hope that our results will spur the discovery of other widespread patterns in genome evolution and lead to improvements in phylogenomic analysis.

ACKNOWLEDGEMENTS

The authors acknowledge the University of Sydney High Performance Computing services for providing computational resources.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was suppported by the Australian Research Council (grant number DP160104173). KJT was supported by an Australian Postgraduate Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Australian Research Council: DP160104173.
Australian Postgraduate Award.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- K. Jun Tong conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Sebastián Duchêne conceived and designed the experiments, performed the experiments, analyzed the data, reviewed drafts of the paper.
- Nathan Lo contributed reagents/materials/analysis tools, reviewed drafts of the paper, coordinated and supervised the study.
- Simon Y.W. Ho conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

Gihub: www.github.com/kjuntong/tree-length.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3241#supplemental-information.

REFERENCES

Bromham L. 2009. Why do species vary in their rate of evolution? *Biology Letters* **5**:401–404 DOI 10.1098/rsbl.2009.0136.

Chapman JW, Reynolds DR, Wilson K. 2015. Long-range seasonal migration in insects: mechanisms, evolutionary drivers, and ecological consequences. *Ecology Letters* **18**:287–302 DOI 10.1111/ele.12407.

Dickerson RE. 1971. The structure of cytochrome c and the rate of molecular evolution. *Journal of Molecular Evolution* **1**:26–45 DOI 10.1007/BF01659392.

Dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society London B* **279**:3491–3500 DOI 10.1098/rspb.2012.0683.

Duarte F, Amrein BA, Kamerlin SCL. 2013. Modeling catalytic promiscuity in the alkaline phosphatase superfamily. *Physical Chemistry Chemical Physics* **15**:11160–11177 DOI 10.1039/c3cp51179k.

Duchêne S, Foster CSP, Ho SYW. 2016. Estimating the number and assignment of clock models in analyses of multigene datasets. *Bioinformatics* **32**:1281–1285 DOI 10.1093/bioinformatics/btw005.

Duchêne S, Ho SYW. 2014. Using multiple relaxed-clock models to estimate evolutionary timescales from DNA sequence data. *Molecular Phylogenetics and Evolution* **77**:65–70 DOI 10.1016/j.ympev.2014.04.010.

Duchêne S, Ho SYW. 2015. Mammalian genome evolution is governed by multiple pacemakers. *Bioinformatics* **31**:2061–2065 DOI 10.1093/bioinformatics/btv121.

Duchêne S, Molak M, Ho SYW. 2014. ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics* **30**:1017–1019 DOI 10.1093/bioinformatics/btt665.

Dykhuizen D, Hartl DL. 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* **96**:801–817.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**:610–618 DOI 10.1038/nrg2146.

Gillespie JH. 1991. *The causes of molecular evolution*. Oxford: Oxford University Press.

Hanada K, Shiu SH, Li WH. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Molecular Biology and Evolution* **24**:2235–2241 DOI 10.1093/molbev/msm152.

Ho SYW. 2014. The changing face of the molecular evolutionary clock. *Trends in Ecology and Evolution* **29**:496–503 DOI 10.1016/j.tree.2014.07.004.

Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology* **23**:5947–5965 DOI 10.1111/mec.12953.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitutions at major histocompatibility complex class I loci reveals over-dominant selection. *Nature* **335**:167–170 DOI 10.1038/335167a0.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, Da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldså J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli K-P, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Wang J, Gilbert MTP, Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**:1320–1331 DOI 10.1126/science.1253451.

Kaufman L, Rousseeuw PJ. 2009. *Finding groups in data: an introduction to cluster analysis*. New York: John Wiley.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**:624–626 DOI 10.1038/217624a0.

Kumar S, Hedges SB. 2016. Advances in time estimation methods for molecular data. *Molecular Biology and Evolution* **33**:863–869 DOI 10.1093/molbev/msw026.

Liaw A, Werner M. 2002. Classification and regression by randomforest. *R News* **2**:18–22.

Libersat F, Delago A, Gal R. 2009. Manipulation of host behaviour by parasitic insects and insect parasites. *Annual Review of Entomology* **54**:189–207 DOI 10.1146/annurev.ento.54.110807.090556.

Losey JE, Vaughan M. 2006. The economic value of ecological services provided by insects. *Bioscience* **56**:311–323 DOI 10.1641/0006-3568(2006)56[311:TEVOES]2.0.CO;2.

Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein-protein interactions. *Molecular Biology and Evolution* **27**:2567–2575 DOI 10.1093/molbev/msq144.

Martinez Cuesta S, Furnham N, Rahman SA, Sillitoe I, Thornton JM. 2014. The evolution of enzyme function in the isomerases. *Current Opinion in Structural Biology* **26**:121–130 DOI 10.1016/j.sbi.2014.06.002.

Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* **346**:1250463 DOI 10.1126/science.1250463.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu S, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, Von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M, Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TKF, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang J, Kjer KM, Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**:763–767 DOI 10.1126/science.1257570.

Muse SV, Gaut BS. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using the relative ratio test. *Genetics* **146**:393–399.

O'Brien PJ, Herschlag D. 1999. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* **6**:R91–R105 DOI 10.1016/S1074-5521(99)80033-7.

Ohta T. 1987. Very slightly deleterious mutations and the molecular clock. *Journal of Molecular Evolution* **26**:1–6 DOI 10.1007/BF02111276.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annual Review of Ecology, Evolution, and Systematics* **23**:263–286 DOI 10.1146/annurev.es.23.110192.001403.

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. *Journal of Molecular Evolution* **1**:18–25 DOI 10.1007/BF01659391.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830.

Penny D, Hendy MD. 1985. The use of tree comparison metrics. *Systematic Zoology* **34**:75–82 DOI 10.2307/2413347.

Peters RS, Meusemann K, Petersen M, Mayer C, Wilbrandt J, Ziesmann T, Donath A, Kjer KM, Aspöck H, Aberer A, Stamatakis A, Friedrich F, Hünefeld F, Niehuis O, Beutel RG, Misof B. 2014. The evolutionary history of holometabolous insects inferred from transcriptome-based phylogeny and comprehensive morphological data. *BMC Evolutionary Biology* **14**:52 DOI 10.1186/1471-2148-14-52.

Smith JM, Smith NH. 1996. Synonymous nucleotide divergence: what is "saturation"? *Genetics* **142**:1033–1036.

Snir S. 2014. On the number of genomic pacemakers: a geometric approach. *Algorithms for Molecular Biology* **9**:26 DOI 10.1186/s13015-014-0026-0.

Snir S, Wolf YI, Koonin EV. 2012. Universal pacemaker of genome evolution. *PLOS Computational Biology* **8**:e1002785 DOI 10.1371/journal.pcbi.1002785.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Thomas JA, Welch JJ, Lanfear R, Bromham L. 2010. A generation time effect on the rate of molecular evolution in invertebrates. *Molecular Biology and Evolution* **27**:1173–1180 DOI 10.1093/molbev/msq009.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B* **63**:411–423 DOI 10.1111/1467-9868.00293.

Tong KJ, Duchêne S, Ho SYW, Lo N. 2015. Comment on "Phylogenomics resolves the timing and pattern of insect evolution". *Science* **349**:487b DOI 10.1126/science.aaa5460.

Tong KJ, Lo N, Ho SYW. 2016. Reconstructing evolutionary timescales using phylogenomics. *Zoological Systematics* **41**:343–351 DOI 10.11865/zs.201640.

Welch JJ, Eyre-Walker A, Waxman D. 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. *Journal of Molecular Evolution* **67**:418–426 DOI 10.1007/s00239-008-9146-9.

Weller C, Wu M. 2015. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution* **69**:643–652 DOI 10.1111/evo.12597.

Wilson EO, Hölldobler B. 2005. Eusociality: origin and consequences. *Proceedings of the National Academy of Sciences of the United States of America* **102**:13367–13371 DOI 10.1073/pnas.0505858102.

Wolf YI, Snir S, Koonin EV. 2013. Stability along with extreme variability in core genome evolution. *Genome Biology and Evolution* **5**:1393–1402 DOI 10.1093/gbe/evt098.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution* **50**:56–68 DOI 10.1007/s002399910007.