Peer Community In Evolutionary Biology

A new statistical tool to identify the determinant of parallel evolution

Stephanie Bedhomme *based on peer reviews by* ***Bastien Boussau*** *and 1 anonymous reviewer*

Susan F Bailey, Qianyun Guo, Thomas Bataillon (2018) Identifying drivers of parallel evolution: A regression model approach. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Evolutionary Biology. 10.1101/118695

Submitted: 22 March 2017, Recommended: 31 January 2018

Cite this recommendation as:
Bedhomme, S. (2018) A new statistical tool to identify the determinant of parallel evolution. *Peer Community in Evolutionary Biology*, 100045. 10.24072/pci.evolbiol.100045

Published: 31 January 2018

In experimental evolution followed by whole genome resequencing, parallel evolution, defined as the increase in frequency of identical changes in independent populations adapting to the same environment, is often considered as the product of similar selection pressures and the parallel changes are interpreted as adaptive. However, theory predicts that heterogeneity both in mutation rate and selection intensity across the genome can trigger patterns of parallel evolution. It is thus important to evaluate and quantify the contribution of both mutation and selection in determining parallel evolution to interpret more accurately experimental evolution genomic data and also potentially improve our capacity to predict the genes that will respond to selection. In their manuscript, Bailey, Guo and Bataillon [1] derive a framework of statistical models to partition the role of mutation and selection in determining patterns of parallel evolution at the gene level. The rationale is to use the synonymous mutations dataset as a baseline to characterize the mutation rate heterogeneity, assuming a negligible impact of selection on synonymous mutations and then analyse the non-synonymous dataset to identify additional source(s) of heterogeneity, by examining the proportion of the variation explained by a number of genomic variables. This framework is applied to a published data set of resequencing of 40 Saccharomyces cerevisiae populations adapting to a laboratory environment [2]. The model explaining at best the synonymous mutations dataset is one of homogeneous mutation rate along the genome with a significant positive effect of gene length, likely reflecting variation in the size of the mutational target. For the non-synonymous mutations dataset, introducing heterogeneity between sites for the probability of a change to increase in frequency is improving the model fit and this heterogeneity can be partially explained by differences in gene length, recombination rate and number of functional protein domains. The application of the framework to an experimental data set illustrates its capacity to disentangle the role of mutation and

selection and to identify genomic variables explaining heterogeneity in parallel evolution probability but also points to potential limits, cautiously discussed by the authors: first, the number of mutations in the dataset analysed needs to be sufficient, in particular to establish the baseline on the synonymous dataset. Here, despite a high replication (40 populations evolved in the exact same conditions), the total number of synonymous mutations that could be analysed was not very high and there was only one case of a gene with synonymous mutation in two independent populations. Second, although the models are able to identify factors affecting the mutation counts, the proportion of the variation explained is quite low. The consequence is that the models correctly predicts the mutation count distribution but the objective of predicting on which genes the response to selection will occur still seems quite far away. The framework developed in this manuscript [1] clearly represents a very useful tool for the analysis of large "evolve and resequence" data sets and to gain a better understanding of the determinants of parallel evolution in general. The extension of its application to mutations others than SNPs would provide the possibility to get a more complete picture of the differences in contributions of mutation and selection intensity heterogeneities depending on the mutation types.

References:

[1] Bailey SF, Guo Q and Bataillon T (2018) Identifying drivers of parallel evolution: A regression model approach. bioRxiv 118695, ver. 4 peer-reviewed by Peer Community In Evolutionary Biology. doi: [10.1101/118695](https://doi.org/10.1101/118695)

[2] Lang GI, Rice DP, Hickman, MJ, Sodergren E, Weinstock GM, Botstein D, and Desai MM (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500: 571–574. doi: [10.1038/nature12344](https://doi.org/10.1038/nature12344)

Reviews

Evaluation round #2

DOI or URL of the preprint: **10.1101/118695**
Version of the preprint: 2

Authors' reply, 23 January 2018

Download author's reply

Decision by Stephanie Bedhomme, posted 23 January 2018

needs some changes before it can be recommended

The two reviewers and myself have now read the revised version of your manuscript. Most of the comments on the previous versions have been addressed and the clarity of the manuscript is improved.

There are still some points that should be addressed before I can recommend this manuscript, in particular the two raised by anonymous. The first on MNM is likely to have very few impact on the results of the analysis but I agree that the contradiction between your answer to his previous comment and what you wrote in the manuscript is puzzling and should be clarified. Probably, also, having access to the list of mutations considered will help readers to follow and understand the subset of mutations used for this manuscript.

One additional comment: I find that "mutations per gene" is a confounding wording which should be changed. Indeed, it can both mean "the number of populations in which a particular gene is mutated" or "the number of mutations within this particular gene in a population" or "the number of different mutations found in a particular gene".

Reviewed by anonymous reviewer 1, 28 November 2017

Summary: Two key issues remain for me: 1) The statements that no MNMs were observed, which seem inconsistent with Lang et al. 2013 and the authors' written response and 2) the lack of support for statements assigning selection as the cause of observations rather than mutational heterogeneity.

Major concern:

The authors' statement that "we do not observe any examples of mutations in close physical proximity" is difficult to reconcile with both the source of their data and with the authors' response to reviewers. In the response to reviewers, they write "there are cases of multiple mutations occurring in the same gene within the same population in the data we analyze and so mutations are all >~1000 bps away from each other". This makes it sound to me as if they have filtered the data set to include only mutations that are 1kb apart. The authors retain 414 mutations from the initial set of 995 mutations observed in Lang et al, or 41.6%, which is also close to the fraction of genes retained. Lang et al., the source of the data for this study, observed some 37 SNPs in 20 separate multi-nucleotide mutations (MNM) events (supplemental table S1), 11 of which contain only no indels, accounting for so ~4% of SNPs occurring in the study, which is close to the frequency observed in Schrider et al. 2011. (The MNMs occur at ChrII:21386, 25614, 676465, 713157; ChrIII:207849; ChrIV:276244, 1201939; ChrVI:238808; ChrVIII:275480; ChrIX:370046; ChrX:152679, 152543, 225902; ChrXII:405998, 820866; ChrXIII:542235; ChrXIV:282587; ChrXV:742929; ChrXVI:869233). In the absence of any biases, then MNMs should be observed in the final data in roughly the same proportion they are observed in the unfiltered data set, in which case we would expect to observe roughly 8.3 MNMs. The failure to observe a single MNM is thus somewhat surprising. One possibility is that the authors have selected only mutations coded as "SNPs" or "InDels" in Lang et al table S1, omitting mutations coded as "compound". Mutations are coded as "compound" on the basis of occurring in close physical proximity to another mutation. If compound mutations were omitted, then obviously no mutations in close physical proximity could be observed. Thus, I would like to understand why MNMs were not observed in their data set. If the authors have chosen to exclude or omit MNMs they could provide a justification for doing so, and if they have not, they could explain what biases might have resulted in the surprising absence of MNMs. Additionally, they could provide a list of the retained SNPs and InDels to permit some verification of their claims. However, overall MNMs are a small portion of the data, and their inclusion/exclusion is unlikely to have substantial effects on the authors' conclusions. The most important consideration wrt the authors' paper is a general concern that the methods utilized by the authors do not support the claims made in the abstract and discussion. This may be my own misunderstanding (obviously), but it appears to me that the authors are making a fundamental statistical error. In the first paragraph of the discussion the authors provide a succinct description of what I understand the authors to have done: We are also able to classify genomic variables into those that have affected mutation counts 1) through their effect on the mutation rate (variables that significantly predict synonymous mutations), and/ or 2) through their effect on the probability of a mutation being either observed/ lost due to selection (variables that significantly predict nonsynonymous mutations).

I read this as a claim that the authors have shown that some genomic variables have significantly different effects at synonymous and nonsynonymous sites (e.g. I read the authors as claiming that an increased number of protein domains decrease mutation counts at nonsynonymous but does not decrease mutation counts at synonymous sites). The support that they have for this assertion is the fact that the variable has significant predictive power at one type of site, but does not have significant predictive power at a different type of site. However, this approach is incorrect. If a treatment has a significant effect in group A but does not have a significant effect in group B it does not follow that the treatment has a significantly different effect in groups A and B. In other words, an effect acting identically on synonymous and nonsynonymous mutation may not be significant for the former and still be significant for the later. This important because there is more power to detect significance for nonsynonymous mutations counts. Imagine that the authors' analysis was done on a subsample of the data rather than the entire dataset. As the data size is reduced, at some point gene length would no longer have a significant effect on synonymous mutation counts, but, because there are many

more nonsynonymous mutations, gene length might still have a significant effect on nonsynonymous mutation counts. It would be fallacious to conclude from this difference in the detection of significant effects that gene length influences the probability a mutation is either observed/ lost due to selection but does not have an effect on the rate at which mutations occur within a gene. This criticism applies generally to statements applying to nonsynonymous which tend to explain these genomic variables by variation in the strength of selection, rather than heterogeneity of mutation rates. To support these statements the authors would need to show that the effect of these genomic variables was significantly different at synonymous and non-synonymous sites, rather than simply non-significant at synonymous sites.

In particular, I was confused by the statement that "We found that gene length predicts nonsynonymous mutation count via selection, over and above its effects on per gene mutation rate – as estimated from models aimed at explaining the synonymous mutation count only." As I do not see what evidence the authors provided for this statement. If the authors could make reference to the table or data show this it would be appreciated.

To address these criticisms the authors could:

1) provide a clear statement by that they have excluded MNMs (if they have) or provide an explanation of why MNMs are absent from their data in addition to providing a supplemental list of retained genes.

2) Fit an identical model to both synonymous and nonsynonymous sites, then the parameter estimates for this models could be compared to see if the estimated values are significantly different between synonymous and nonsynonymous sites. Obviously, some parameters of the model may have no significant predictive power for one class of sites, but this step would establish that a difference in significance is due to a smaller effect size on one class of sites rather than decrease statistical power for that class of site.

I hope I have not misunderstood the authors due to my own inattention, and I apologize in advance if this is the case.

Reviewed by Bastien Boussau, 28 November 2017

Major comments

Bailey et al. provide an updated version of their manuscript where reviewer's comments have been taken into account.

I think the new version has improved compared to the first one, and at this stage only have some suggestions for improvements that I don't think are mandatory. However, I would argue that they would further improve the manuscript.

Notably, I think the link between the authors'analyses and convergent or parallel evolution is still not sufficiently clear. In particular, the authors now include a reference to Zhang and Kumar about parallel vs convergent evolution: while this strict definition may seem useful at first glance, I think in this manuscript it misleads more than it helps because the authors never actually look at changes at the very same sites in genes. Instead, they use "parallel evolution" to describe the case of the gene IRA1, that "saw mutations in over 50% of the populations sequenced in this experimental data set". I think in that context the use of the term does not fit their early definition. Instead I would suggest that they spend some time discussing different levels of convergence/parallelism, at the nucleotide/gene/pathway level, so that they can state clearly what level they are going to focus on.

Further I would plead for an additional paragraph at the end of the introduction stating the author's reasoning, which seems to be that to understand patterns of convergent or parallel evolution, first one needs to identify the parameters that enable the prediction of synonymous and non-synonymous mutation rates at the gene level. Second, once those parameters have been identified, they can be used to test whether they allow recovering similar patterns of gene-wise parallelism/convergence. More specific comments

In particular, the last comment suggests to add a panel to figure 4, to show how simulated data cannot fit genes like IRA1.

- l55: "Parallel evolution is an identical change in independently evolving lineages, and the similar processes,

convergent evolution": process

- l122: Is \pi*i really a probability? Given that* *lambda*iN=\lambda*iS* *times* *pi*i, and that \lambda*iS already contains a* *pi*O that describes a probability of fixation, I was under the impression that \pi_i was a scaler that could be >1, if selection is such that it favours fixation for gene i.
- l276: "and the per nucleotide mutations does not vary significantly across the genome": mutation
- l259: "and an example script for implementing our model framework and hypothesis testing is available on Dryad (doi will be inserted here).": I think it is a very useful idea.
- l310: "only a single principal component, PC10, was significant in the model (see model M N .NB PC in Table 3)": How much variation did this component explain? I assume it must be very low, being the 10th component.
- l360: "However, rates of HGT tend to be higher in bacteria, and in particular E. coli, as compared to yeast and other eukaryotes (e.g. Boto 2010).": I could not find this Boto 2010 reference.
- l367: "dS and dN/dS are noisy to estimate at the gene level and that tends to downplay their predictive power in our analysis of counts in evolve and re-sequence experiment.": to further investigate this noise hypothesis it could be interesting to look at the predicted numbers of substitutions in the gene alignments (e.g. sum of branch lengths * alignment lengths), because I expect more noise if the alignments are very conserved, or on the contrary extremely divergent.
- l432: "For example, one gene (IRA1) saw...": In Fig. 4, the authors show the distribution of Jaccard indices between pairs of genes over 40 simulated replicate populations. While this shows that the model cannot quite fit the amount of convergent evolution observed in the real data, it does not show cases like IRA1 that appear in 50% of the replicates. I think it would have been nice to show in addition to the distribution of Jaccard indices the true and simulated distributions of numbers of replicates where each gene was hit with a mutation.

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/118695
Version of the preprint: 1

Authors' reply, 17 August 2017

Download author's reply

Decision by Stephanie Bedhomme, posted 17 August 2017

Revise

The preprint "Identifying drivers of parallel evolution: A regression model approach" has now been read by two reviewers and myself. We all agree that the central topic of the paper and the methods derived are of interest but that the paper in its actual form cannot be recommended. Various problems have been pointed by the reviewers and I synthesize and complete them below:

- There seems to be a disconnection between the title and the introduction which focus on parallel evolution and the results and discussion which focus on the factors affecting the probability of a gene to carry a mutation by the end of the experimental evolution. The introduction makes the reader expect that the methods developed is going to be able to determine to what extent parallel evolution is due

to the probability of the mutation to happen and to selection. In other words, from the introduction, I expected the method to be able to discriminate cases where parallel evolution can be truly taken as a strong signal for adaptive mutations and cases where parallel evolution is due to neutral processes. The methods developed is not reaching this goal, at least not explicitly, and the added value of the methods does not appear clearly to the reader.

- More details should be given on the experimental design, in particular on the ploidy of the yeast and the reproduction mode they had during experimental evolution (see point 4 of MA).

- The authors rely on the hypothesis that synonymous mutations are neutral to selection, which is a classical one, but they write in their discussion (l 403): "Relying on the assumption that synonymous mutations are selectively neutral (which does appear to be the case for these data)" and I do not see where the neutrality is tested in the study. More importantly, for non-synonymous mutations, they fit a number of models to try and detect the effect of different genomic variables on the heterogeneity in the probability that a mutation rises. Among this genomic variables, some are likely to affect non-synonymous as well as synonymous mutations and their link to selection is not obvious and straightforward. The comment on GC content by MA is going in this direction and similar argument could be developed for CAI and recombination rate. As far as I understand the effect of these variables on the synonymous mutations has not been tested, so it cannot be claimed that they have an effect on NS mutation that they have not on S mutations.

- All the manuscript is focussed on SNP when high levels of parallelism have been found for IS and large duplications and deletions (see for example Tenaillon et al. 2012). I recognized that it is more difficult to derive a modelling framework for them and that the present one cannot be easily adapted but I think that these mutations have a strong impact on adaptation and would like to see some comments on them, at least, in the discussion.

Reviewed by anonymous reviewer 1, 28 November 2017

While this article has many merits, unfortunately I cannot recommended it at this time. My primary concern is that it is unclear what novel conclusions should be drawn. The authors provide clear evidence that large genes experience more mutations than small genes, and that selective constraints vary between genes; however, these observations are trivial. I suspect that more important questions can be addressed with their approach, but they have failed to articulate these questions. These criticisms can be addressed by clarifying how the models tested change our interpretation of previous experimental results. Additionally, I suggest some methodological changes. If the authors feel that I have misunderstood key points of this paper, I suggest they attempt to divine the source of my misunderstanding and make appropriate clarifications so that future reviewers do not make similar mistakes. Despite these criticisms, I feel that the core of the paper is potentially interesting, and I look forward to seeing a future versions of this manuscript.

General summary.

The authors use several models to show that mutation rates are, to a first order approximation, constant across the genomes of sets of yeast growing under controlled conditions, and that non-synonomous mutations are subject to varying amounts of selection. They identify several features of genes that correlate with the frequency at which mutations arise, such as GC content, and also features that correlate with the strength of purifying selection, such as the number of functional domains in a protein. The abstract, discussion, and title of the paper focus our attention on the importance of parallel evolution, which in this context means identical mutations arising to detectable frequencies independently in multiple lines. I presume that the paper intends to contrast the possibility that a non-synonymous mutation observed in many replicate lines was selectively advantageous with the possibility that the site in question was hyper-mutable. This is a reasonable and interesting question. However, I did not find these hypotheses stated clearly. The discussion contrasts

these hypotheses when noting the failure of any tested model to predict the high number of mutations seen in some genes, but this observation should be central to the manuscript.

Major comments:

1) GC content is included as a variable in non-synonymous mutation rates, but not as a variable in synonymous mutation rates. One could argue that the failure to detect substantial mutational heterogeneity between genes (i.e. the Poisson had a lower AIC than the negative binomial) implies that GC would not be a significant predictor of mutation rates. This may be correct; however, the significance of GC content in the non-synonymous models is most probably explained by the effect of GC content on mutation rate and not as a predictor of the strength of selection. If GC content does not significantly correlate with synonymous mutation counts, then this points to a difference in the power to detect mutational heterogeneity at non-synonymous and synonymous sites. This difference in power has implications for the interpretation of the results and should be addressed.

2) It has been consistently found that some substantial fraction of mutations occur in complex events that alter many nearby nucleotides (multinucleotide mutations or MNM; Schrider 2011). This is problematic if the authors' method would tabulate a single MNM event as two or more parallel mutations. Additionally, because MNM events happen on very small scales, typically affecting adjacent nucleotides, they disproportionally cause adjacent non-synonymous changes rather than adjacent synonymous changes. This can be addressed by counting MNM events as single events.

3) A persistent challenge in experimental evolution is separating relaxed selection on a gene from adaptation. While relaxed selection is arguably a form of parallel evolution, the methods adopted by the authors could provide insight into separating these two forms of evolution. It would be an interesting addition to discuss this in some detail.

4) Because this paper analyzes data from a single experiment, more details on the conditions in that experiment should be included, particularly information on general growth conditions (batch size, frequency of transfer, volume transferred, etc), whether the yeast were grown as haploid or diploid, and whether they were given the opportunity to have sex. This information is crucial to determining the meaning of these results, and should be at least broadly summarized in this paper.

Minor comments:

Line 174 "essential genes" reads awkwardly in the list modifying "each gene." Perhaps "essentiality of the gene."

Line 199 reports a result in the methods section… omitted word "whether."

Line 360: This observation would be much more interesting and informative if the authors had tested for an effect of r on synonymous mutation counts.

Line 366: Are the yeast growing as haploids or diploids? If they are growing as haploids w/o sex, then then there should be no opportunity for BGC to occur.

Expression levels are sensitive to growth conditions. If available, the expression data from growth under experimental conditions should be used for all analyses.

I favor the definition of parallel evolution being used here, but quite a lot of confusion exists between the use of the terms parallel evolution and convergent evolution. Since both of these terms are used in this paper, it would be useful to clearly define the terms. I would recommend citing an authoritative usage of the term, such as Zhang and Kumar 1997.

Reviewed by **Bastien Boussau**, 28 November 2017

This manuscript aims at understanding the variables that affect parallel evolution in an experiment conducted in yeast. It compares statistical models that include different variables and conclude that gene length or recombination rate affect the rate of mutation. I found this paper interesting and I think the model comparison approach is sound, but in the end I was a bit confused about what had really been achieved. The introduction focuses on parallel evolution, but looking at the methods, it seems like all mutations have been analyzed in

the manuscript (lines 145-151, page 7), not only the mutations that occur in genes that have been hit multiple times. So in the end it is unclear to me why the results apply to parallel mutations and not to mutations in general. The authors analyze 414 substitutions in total, assuming that non-synonymous substitutions are under selection, and synonymous mutations are evolving neutrally. However it may be that not all non-synonymous substitutions are under selection. In the Lang paper where the sequencing was conducted, it is noted that some genes have been hit multiple times in the populations, and it is concluded that these genes are likely targets of selection. I think it would be interesting to analyze separately the subset of mutations occurring in those genes only (if there are enough), because several non-synonymous substitutions that the authors chose to analyze may in fact be neutral or nearly neutral. The other experiment I would be curious to see conducted is an analysis of the mutations with respect to the GC content of the arrival state. As I suggest below, GC-biased gene conversion may partly explain why there is a correlation between the number of mutations and the local recombination rate.

More specific comments follow. Many of those are just typos, but in the mix there are also genuine scientific questions.

p4 l66: ”genes that exhibit a higher that expected number”: than ”We used a codon table model with a fixed tree topology

165 (a comparison of AICs among alternative codon based models indicated this was the most appropriate

166 model for the data set).”: this is not clear to me, I’s prefer to see the name of the model according to PAML (e.g. M0, M1...).

p10 l215: ”permutation tests instead of relying on asymptotic distribution of the LRTs”: it is not clear to me how the permutations were done. What variables were permuted, and how were they permuted?

p12 l262 ”and MS1: λS = constant*(Li)α”: I think in other parts of the manuscript alpha was alpha*1.*

p13 l277: ”evenly loaded with a number genomic”: number of

p13 l291: ”that can significantly predict the distribution of mutations”: I’m not sure what significantly predicting means.

p14 l307: ”mutation counts from Lenksi’s long term evolution experiment”: Lenski’s

*p15 l343:” Further evidence that gene length acts as a summary variable comes from the M3 results (summarized in Table 3), where we see that gene length is no longer significant when other summary variables – the principal components – are included in the model.”: I’m confused. M3 is a new notation, not found in table 3. If M3 is in fact MN.NB*PC, then gene length is included in PC10 already, so I don’t understand the argument.

p16 l355: what about the other correlations? Could the number of domains be another ”summary variable”?

p16 l362 ”double strand breaks in substantially increases the frequency of nearby point mutations in nearby intervals”: remove the first in, and too many ”nearby”s.

p16 l365 ”Another non exclusive possibility might be the fact that biased gene conversion might vary from gene to gene and also – like selection - affect the probability of detecting variants in evolve and re-sequence experiments”: (a point is missing at the end of the sentence) Indeed, biased gene conversion behaves as selection in terms of its impact on the probability of fixation. In that case, wouldn’t we expect the variants to be GC biased (cf https://www.ncbi.nlm.nih.gov/pubmed/23505044)? Would it be possible to check the GC content of those variants?

p18 l408 ”and move closer the goal of predicting which genes”: closer to

p25 table 1 ”based growth assays of deletion strains.”: based on