

## Revision round #2

2018-01-23

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”.

Preprint DOI: <https://doi.org/10.1101/118695>

Reviewed by [anonymous](#), .....2017-11-28 15:59

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  $> \sim 1000$  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  $\sim 4\%$  of SNPs occurring in the study, which is close to the frequency observed in Schridder 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](#), 2017-11-28 15:59

## 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.

- 155: "Parallel evolution is an identical change in independently evolving lineages, and the similar processes, convergent evolution": process
- 1122: Is  $\pi$  really a probability? Given that  $\lambda_N = \lambda_S \times \pi$ , and that  $\lambda_S$  already contains a  $\pi$  that describes a probability of fixation, I was under the impression that  $\pi_i$  was a scalar that could be  $>1$ , if selection is such that it favours fixation for gene  $i$ .
- 1276: "and the per nucleotide mutations does not vary significantly across the genome": mutation
- 1259: "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.
- 1310: "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.
- 1360: "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.
- 1367: "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.
- 1432: "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.

**Author's reply: See next page**

We thank the recommender and two reviewers for their time and input. We appreciate the constructive comments and suggestions and feel our manuscript has been greatly improved through this process. We hope that the current manuscript and associated responses to comments below adequately address concerns raised. Recommender/ reviewer comments are shown in **blue**, followed by our specific responses in **black**.

### **Response to comments from Stephanie Bedhomme**

...

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.

We address these concerns in direct response to anonymous's comments below.

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".

We now try to be more clear with this term, specifying "mutations per gene (totaled over all 40 populations in the data set)" on lines 220-221, line 312, and lines 488-489.

### **Response to comments from anonymous**

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.

...

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.

We apologize for being unclear about this in previous comments. MNMs are indeed absent from the data set used in this analysis. Their absence is unintentional, but is due to our pre-analysis filtering. Some MNMs were excluded because they arose in intergenic regions (we only use the 718 genic

mutations in our analysis), and some were excluded due to the removal of genes for which complementary genomic data was unavailable. Our filtering process led to a final mutation count of 393 and happened to remove all the instances of MNMs. We now describe this process more clearly in the manuscript on lines 206-209. A list of the retained genes and their corresponding genomic variable estimates will be available on Dryad following the publication of this manuscript.

- 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.

We understand the point being made here, but respectfully disagree that it applies to our analysis. The comparison suggested by the reviewer certainly makes sense if one is interested in comparing multiple treatments within the same dataset. However, we feel that, conceptually, the non-synonymous and synonymous mutations should be considered more as two different datasets, instead of two treatments within the same dataset. We aim in our analysis to merely use the synonymous dataset as a baseline, to characterize mutation rate across the genome, but not as a direct comparison with the non-synonymous mutations.

However, to satisfy those who may not fully agree with our afore-described conceptual framework, we made two additional statistical comparisons:

- 1) We tested for statistically significant differences between parameter estimates for the synonymous versus nonsynonymous data, by looking for significant differences in fit between a model that estimates the same parameters for both synonymous and nonsynonymous mutations, and a model that estimates different parameters for synonymous and nonsynonymous mutations.
- 2) We addressed the potential power issue raised by the reviewer by randomly down-sampling the non-synonymous mutation data, so that it has the same sample size as the synonymous mutation data. We then re-fit the models to see if we still detect significant effects and repeated this process 1000 times to obtain a distribution of effects.

We attach the results of these analyses in a separate file (results of additional analyses.pdf), but to summarize, we **do** find support for real differences between the synonymous and non-synonymous data. Using approach 1, we found statistically significant differences in the parameter estimates from the principal components model, however we do not find a significant different between parameter estimates in the untransformed genomic variables model. Using approach 2, we found that models fit a down-sampled non-synonymous data set tended to be more significant than the mode fit to the synonymous data. This was significant or marginally significant for all the parameters except *recombination* ( $r$ ).

However, despite the support from these additional analyses, we would prefer to leave these additional results out of the manuscript altogether. We feel they will detract too much from the main message of the study, which is already quite cautious in its claims.

### Response to comments from Bastien Boussau

...

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.

We have added two sentences on lines 54-58 to better qualify what we mean by parallel evolution in the context of this study – specifically we are focusing on parallel evolution at the level of the gene.

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.

We now try better link these ideas at the end of the intro on lines 93-95.

### 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.

- I155: "Parallel evolution is an identical change in independently evolving lineages, and the similar processes, convergent evolution": process

This is now fixed.

- I122: Is  $\pi_i$  really a probability? Given that  $\lambda_i = \lambda \times \pi_i$ , and that  $\lambda$  already contains a  $\pi_0$  that describes a probability of fixation, I was under the impression that  $\pi_i$  was a scalar that could be  $>1$ , if selection is such that it favours fixation for gene  $i$ .

Yes, you are correct. Thank you for pointing this out. We have now corrected our language here, see lines 127-128.

- I276: "and the per nucleotide mutations does not vary significantly across the genome": mutation

This is now fixed on line 222.

- I259: "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.

Great, we will be sure to do this.

- I310: "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.

This model explained about 16% of the total variation. We now include this measure on lines 320-321.

- I360: "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.

Reference is now added.

- I367: "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.

This is an interesting idea, but we feel it is outside of the main message of this study.

- I432: "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.

Conceptually we treat the replicates as exchangeable entities so when we summarize the patterns of observed evolution at the gene level, we effectively sum the number of mutations for each gene across replicates. So, in our opinion, when suggesting that we should report "the true and simulated distributions of numbers of replicates where each gene was hit with a mutation" this is actually the same as the distributions illustrated in Figure 2 where we report the observed versus expected distribution of total number of hits (mutations) per gene for both the synonymous and non-synonymous hits. The figure 2 is also meant to illustrate as honestly as possible the fact that even our best fitting model does exhibit some lack of fit to the data.



## Additional analyses of the authors

**1A) Look for significant differences between parameter estimates** for nonsynonymous versus synonymous mutation counts

Parameters of interest:  $constant$ ,  $\alpha 1$ ,  $\alpha 2$ ,  $\alpha 3$ ,  $\theta$

Model 1: Same parameters are fit for both non-synonymous and synonymous mutations

$$NB(\lambda(N) = constant * L_i * L_i^{\alpha 1} * num.dom_i^{\alpha 2} * r_i^{\alpha 3}, \theta) + NB(\lambda(S) = constant * L_i * L_i^{\alpha 1} * num.dom_i^{\alpha 2} * r_i^{\alpha 3}, \theta)$$

Model 2: A different 'constant' is fit for non-synonymous and synonymous mutations. All other parameters are the same for non-synonymous and synonymous mutations.

$$NB(\lambda(N) = \underline{constant}_N * L_i * L_i^{\alpha 1} * num.dom_i^{\alpha 2} * r_i^{\alpha 3}, \theta) + NB(\lambda(S) = \underline{constant}_S * L_i * L_i^{\alpha 1} * num.dom_i^{\alpha 2} * r_i^{\alpha 3}, \theta)$$

Model 3: All parameters are different for non-synonymous and synonymous mutations.

$$NB(\lambda(N) = \underline{constant}_N * L_i * L_i^{\alpha 1N} * num.dom_i^{\alpha 2N} * r_i^{\alpha 3N}, \theta) + NB(\lambda(S) = \underline{constant}_S * L_i * L_i^{\alpha 1S} * num.dom_i^{\alpha 2S} * r_i^{\alpha 3S}, \theta)$$

Results:

	df	LogLik	AIC
Model 1:	5	-1231.2	2472.302
Model 2:	6	-1222.1	2456.285
Model 3:	9	-1219.4	2456.814

Therefore, the synonymous and nonsynonymous counts have significantly different constants (i.e. means). But there is no significant difference between any other parameters.

There is also no significant difference if one tests each parameter in isolation. Doing the same comparison focused on each variable, we obtain the following P-values:

$\alpha 1$ (parameter for <i>length</i> ):	P = 0.1863
$\alpha 3$ (parameter for <i>num.dom</i> ):	P = 0.5933
$\alpha 3$ (parameter for <i>r</i> ):	P = 0.1529

**1B) Test for significant differences between parameter estimates in the principal component model:**

Model 1:  $NB(\lambda(N) = constant * L_i * PC10_i^{\alpha 1}, \theta)$

Model 2:  $NB(\lambda(N) = \underline{constant}_N * L_i * PC10_i^{\alpha 1}, \theta) + NB(\lambda(S) = \underline{constant}_S * L_i * PC10_i^{\alpha 1}, \theta)$

Model 3:  $NB(\lambda(N) = \underline{constant}_N * L_i * PC10_i^{\alpha 1N}, \theta) + NB(\lambda(S) = \underline{constant}_S * L_i * PC10_i^{\alpha 1S}, \theta)$

	df	LogLik	AIC
Model 1:	3	-1066.0	2137.983
Model 2:	4	-1065.5	2138.908
Model 3:	5	-1056.4	2122.755

The model fits significantly better if we calculate different PC10 parameters for the synonymous and non-synonymous mutations.

**2) Reduce the sample size of the non-synonymous counts** and re-run the analysis to see if significance is lost. This is meant to get at the possibility that we don't see significant effects in the synonymous model simply because there are far fewer mutations and so we don't have the power to detect anything.

Procedure:

1. Count the number of mutations in the synonymous mutation data set =  $no.S = 52$
2. Re-sample the nonsynonymous mutation data set by randomly selecting  $no.S$  genes (using `sample()`, with replacement), setting the probability of re-sampling a particular gene,  $i$ , equal to (NS mutations in gene  $i$ )/(total number of NS mutations)
3. Fit GLMs to the re-sampled data  
M1: with the variable of interest (e.g. Length, Num.dom, r, theta (i.e. Poisson vs NB))  
M2: without the variable of interest
4. Perform a likelihood ratio test for models M1 and M2. A significant difference in fit tells us that the variable of interest significantly improved the model fit.
5. Re-sample again, and re do the analysis. Do it 1000 times, report the distribution of likelihood ratio test P-values for each variable of interest.
6. For each variable of interest, see where the synonymous mutations P-values falls in relation to the distribution of P-values from the re-sampled non-synonymous mutations models.

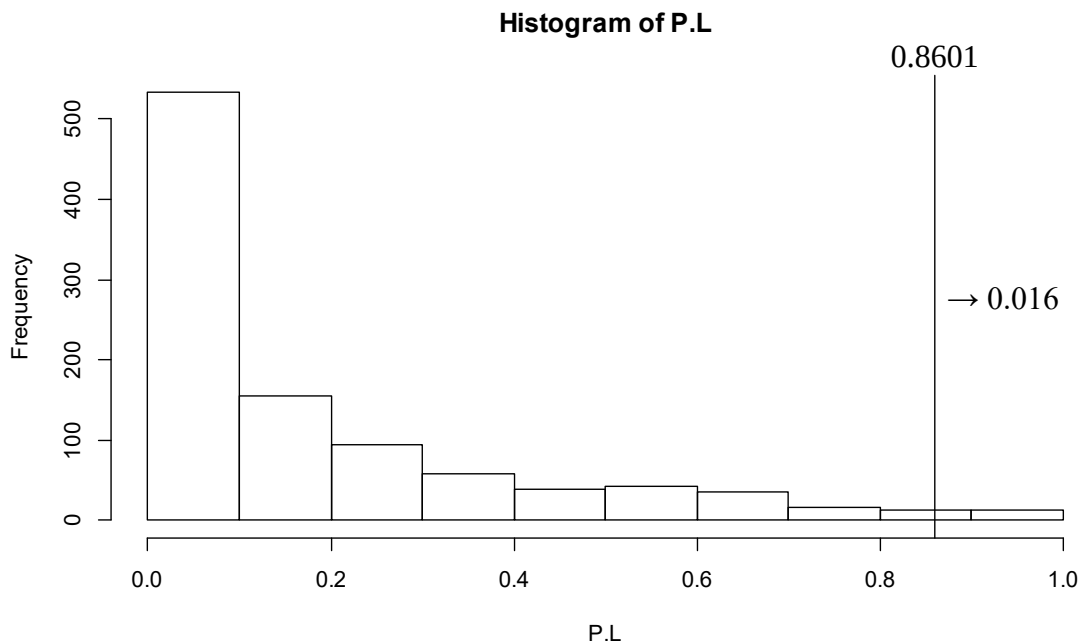
Results:

Figures below show the distribution of P-values for the variable of interest from 1000 re-sampled non-synonymous models. Vertical lines indicate the P-value for the synonymous model. To summarize, this analysis suggests that the parameter for *length* and *theta* (the dispersion parameter) are truly not significant in the synonymous model because the re-sampled nonsynonymous data had a lower P-value more than 95% of the time (i.e. this is not just a power issue; *length*  $P = 0.016$  and *theta*  $P < 0.001$  respectively). There is marginal support that this is also the case for the *num.dom* parameter ( $P = 0.097$ ) and no support for this with the *r* parameter ( $P = 0.29$ ).

**Length**

Synonymous model:  $P = 0.8601$

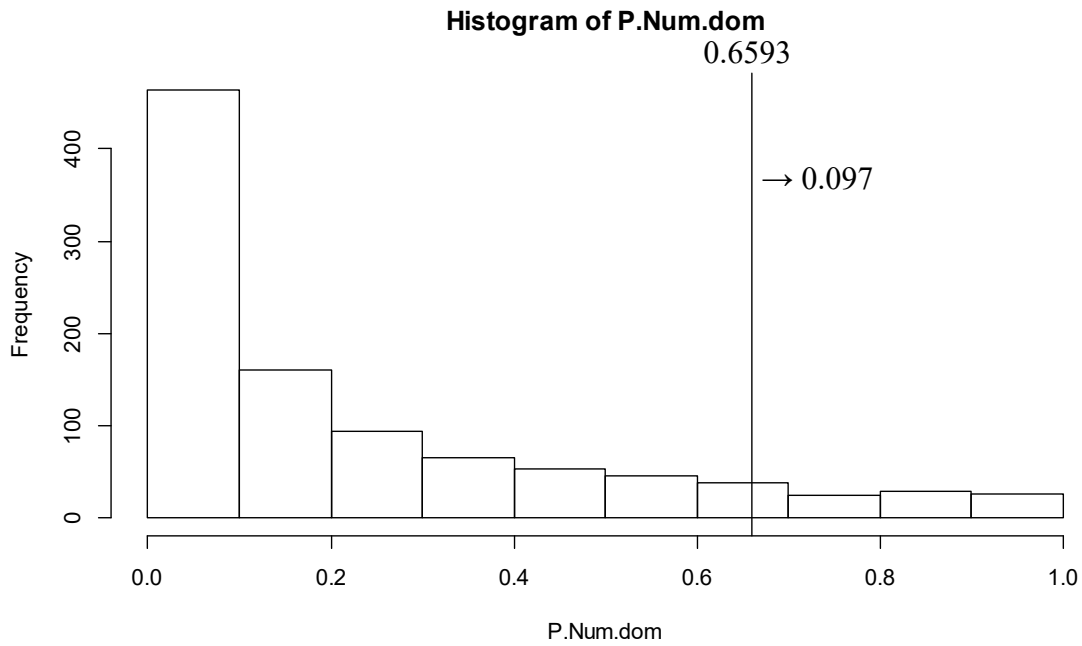
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value: 0.016



**Num.dom**

Synonymous model:  $P = 0.6593$

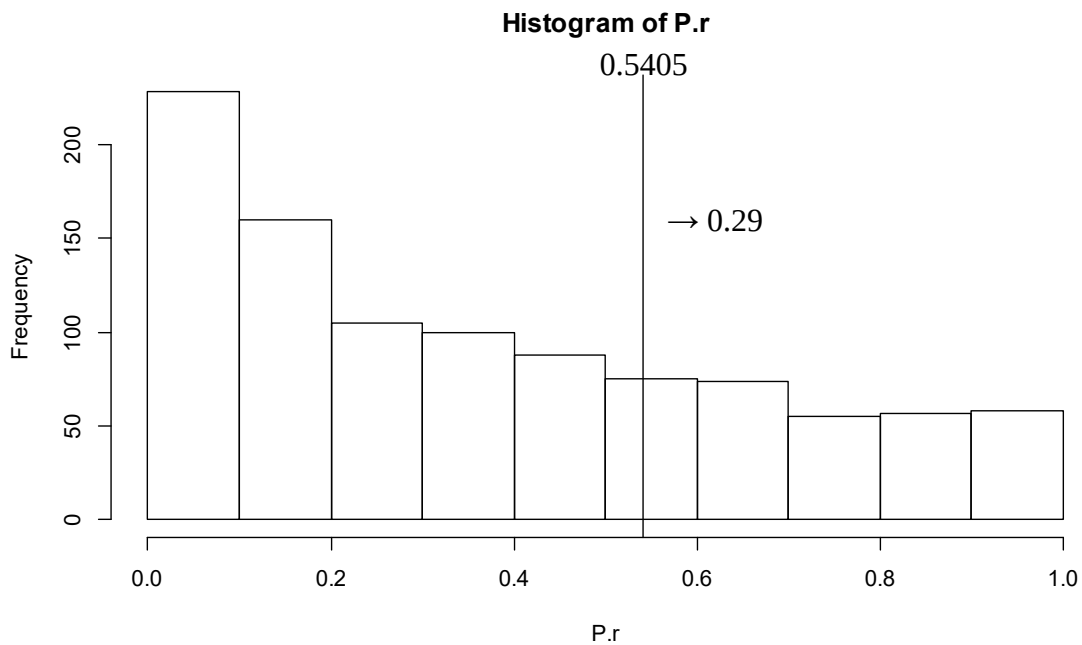
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value:  
0.097



**r**

Synonymous model:  $P = 0.5405$

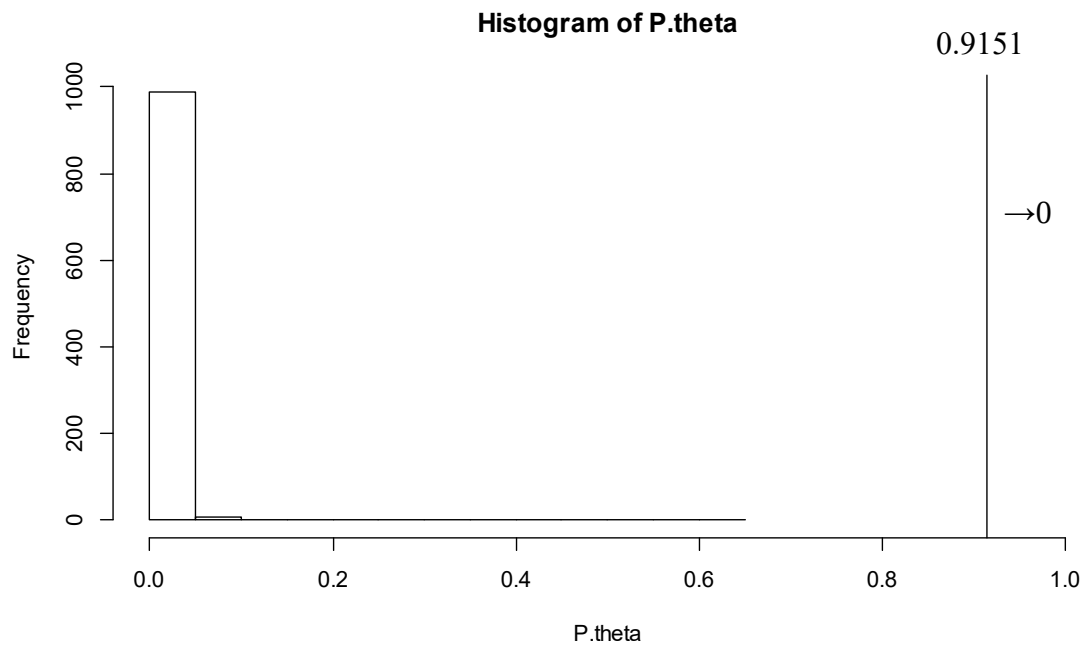
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value:  
0.29



**theta**

Synonymous model:  $P = 0.9151$

Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value:  
0



---

## Revision round #1

2017-08-17

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.

Preprint DOI: <https://doi.org/10.1101/118695>

Reviewed by [anonymous](#),..... 2017-11-28 15:59

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](#), 2017-11-28 15:59

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 than 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'd 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:  $\lambda S = \text{constant} * (Li)^\alpha$ ": I think in other parts of the manuscript alpha was  $\alpha l$ .

*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 Lenski'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.NBPC, 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

**Author's reply: See next page**



We thank the recommender and two reviewers for providing their constructive comments. We have now revised our manuscript with an aim to address those comments, focusing on improving the connection between the introduction, and the rest of the manuscript with respect to ideas about parallel evolution. Responses to specific comments can be seen below. Recommender/ reviewer comments are shown in blue, followed by our responses in black.

### **Response to comments from Stephanie Bedhomme**

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.

We think there is a valid point here. We recognize the potential for disconnection and we have now added an additional figure (Fig 4) and a section in the discussion (lines 450-460) to better connect the paper results and discussion to the idea of parallel evolution. We argue that our modeling approach makes testable predictions regarding the expected level of parallel evolution. We illustrate this on the dataset by contrasting predicted and observed degree of parallelism (estimated as the pairwise Jaccard Index, J).

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).

This is now included on lines 156-159.

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.

We agree and we now explicitly test these assumptions. See lines 218-219 (methods), lines 296-301 (results), and supplementary info table S2.

All the manuscript is focused 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.

This has now been added to the discussion on lines 444-449.

### **Response to comments from anonymous**

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.

We agree that GC content of a gene is a genomic variable that could be a proxy for a variety of processes including mutation and (biased) gene conversion rate variation throughout the genome. We now report the effect of including this variable when modeling both synonymous and non synonymous mutation counts as well as additional tests. While the general point made by the reviewers is well taken, it turns out here that this covariate explains preciously little of the variation in mutation counts. We therefore chose to keep the discussion of this variable to a minimum (but see revised results lines 406-410).

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 disproportionately cause adjacent non-synonymous changes rather than adjacent synonymous changes. This can be addressed by counting MNM events as single events.

We do not think this is a problem in our study. 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  $> \sim 1000$  bps away from each other. We now mention this on lines 162-165.

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.

We are not sure we understand what the reviewer is suggesting here. However, it is not clear how we would separate relaxed selection from adaptation with our method. This might be feasible if one were explicitly fitting different fitness classes of mutations, but this approach is out of the scope of this study.

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.

Additional information has been added on lines 156 – 159, as well as an additional reference that details the experimental conditions.

Minor comments:

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

Changed, see line 190.

Line 199 reports a result in the methods section... omitted word “whether.”

There is actually no omitted word here.

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.

We now include this test. See supplementary table S2.

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

The yeast were grown as haploids without sex in the experiment that the data come from. Regarding the likelihood of BGC, we agree that scope for this process is probably more limited in haploids although yeast is known to undergo mitotic gene conversion (e.g. Bethke and Golin, 1994, Genetics) and there is certainly evidence from other haploid organisms, namely bacteria, that BGC can drive GC content in genomes (Lassalle et al. 2015. GC-content evolution in bacterial genomes: The biased gene conversion hypothesis expands. PLoS genetics).

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

Both the experimental evolution study and the expression study growth conditions are standard YPD growth media at 37 deg C. We now make this clear in the manuscript on lines 157, and lines 189-190.

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.

We now define these terms and include a citation in the introduction (lines 55-58)

### **Response to comments from Bastien Boussau**

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.

As mentioned above in our response to other reviewer comments, we have now added an additional figure (Fig 4) and a section in the discussion (lines 450-460) with an aim to better connect the paper results and discussion to the idea of parallel evolution. We argue that our modeling approach (including all the mutation data, both those that represent parallel evolution and those that do not) makes testable predictions regarding the expected level of parallel evolution. We also agree with the reviewer's point that several of the non-synonymous mutations in this data set may be neutral. In fact, this possibility fits well with our approach. In our statistical models, selective effects of non-synonymous mutations can range from deleterious, to neutral, to beneficial. We now state this explicitly on lines 125-127.

More specific comments:

p4 l66: "genes that exhibit a higher than expected number": than

Fixed.

"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'd prefer to see the name of the model according to PAML (e.g. M0, M1...).

We now include the following "We used a codon table model (i.e. seqtype = 1; CodonFreq = 3) with a fixed tree topology (i.e. runmode = 0)" on lines 180-181.

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?

We have expanded our explanation of the permutations on lines 233-237.

p12 l262 "and MS1:  $\lambda S = \text{constant} * (Li)^\alpha$ ": I think in other parts of the manuscript alpha was alpha 1.

Changed.

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

Fixed.

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

We have changed the wording here to “significantly improve models predicting the distribution of mutations”.

p14 l307: "mutation counts from Lenski's long term evolution experiment": Lenski's  
Fixed.

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.NBPC, then gene length is included in PC10 already, so I don't understand the argument.

This sentence was inadvertently left in the manuscript from a previous iteration of the models. It has now been removed.

p16 l355: what about the other correlations? Could the number of domains be another "summary variable"?  
Yes, this is a valid point. We now point this out on lines 383-284.

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.  
Fixed.

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?

We checked, and saw that GC content is not significant in our model. We also see no bias towards GC in the variants. We now explain this on lines 406-410.

p18 l408 "and move closer the goal of predicting which genes": closer to  
Fixed.

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