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' 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 that 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.

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

This is now fixed.

- l122: Is \pi*i really a probability? Given that \lambda*iN=\lambdai*S \times \pi*i, and that \lambdai*S 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.

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

- l276: "and the per nucleotide mutations does not vary significantly across the genome": mutation

This is now fixed on line 222.

- 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.

Great, we will be sure to do this.

- 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.

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

- 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.

Reference is now added.

- 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.

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

- 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.

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) <u>Look for significant differences between parameter estimates</u> for nonsynonymous versus synonymous mutation counts

Parameters of interest: *constant, α1, α2, α3, θ*

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

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 different if one tests each parameter in isolation. Doing the same comparison focused on each variable, we obtain the following P-values:
α1 (parameter for *length*): P = 0.1863
α3 (parameter for *num.dom*): P = 0.5933
α3 (parameter for *r*): P = 0.1529

1B) <u>Test for significant differeneces between parameter estimates in the principal component model:</u>
Model 1: $NB(\ \lambda(N) = constant * L_i * \mathrm{PC10}_i^{\ \alpha 1}, \theta\ )$
Model 2: $NB(\ \lambda(N) = \underline{constant_N} * L_i * \mathrm{PC10}_i^{\alpha 1}, \theta\ ) + NB(\ \lambda(S) = \underline{constant_S} * L_i * \mathrm{PC10}_i^{\alpha 1}, \theta\ )$
Model 3: $NB(\ \lambda(N) = \underline{constant_N} * L_i * \mathrm{PC10}_i^{\underline{\alpha 1N}}, \theta\ ) + NB(\ \lambda(S) = \underline{constant_S} * L_i * \mathrm{PC10}_i^{\underline{\alpha 1S}}, \theta\ )$

```
          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) <u>Reduce the sample size of the non-synonymous counts</u> 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.

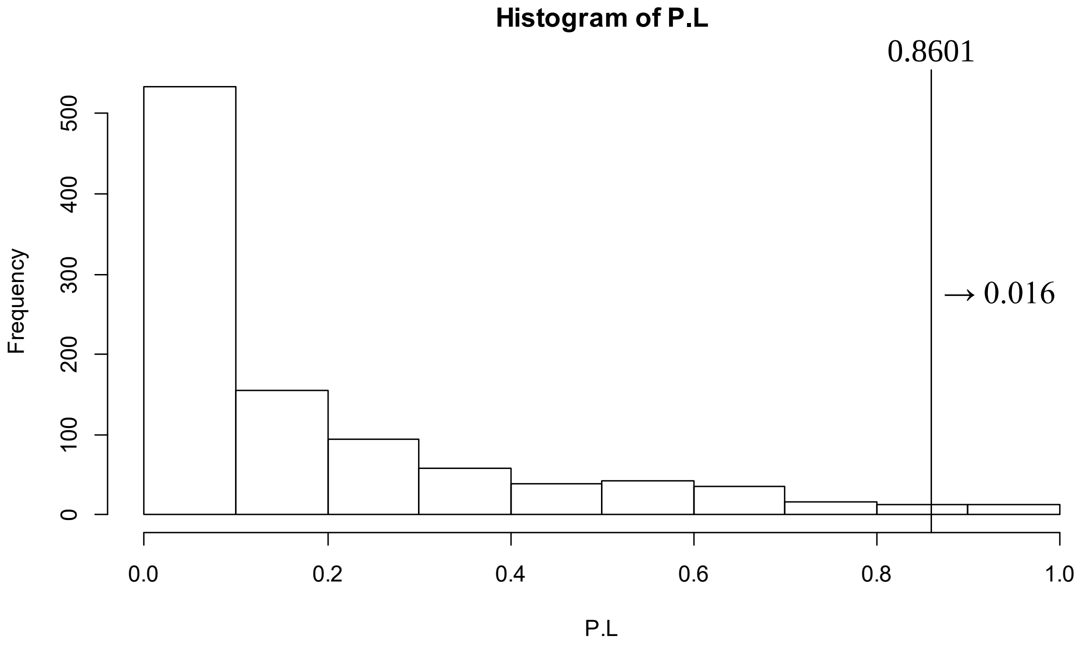
Results:

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 nonsynoymous 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

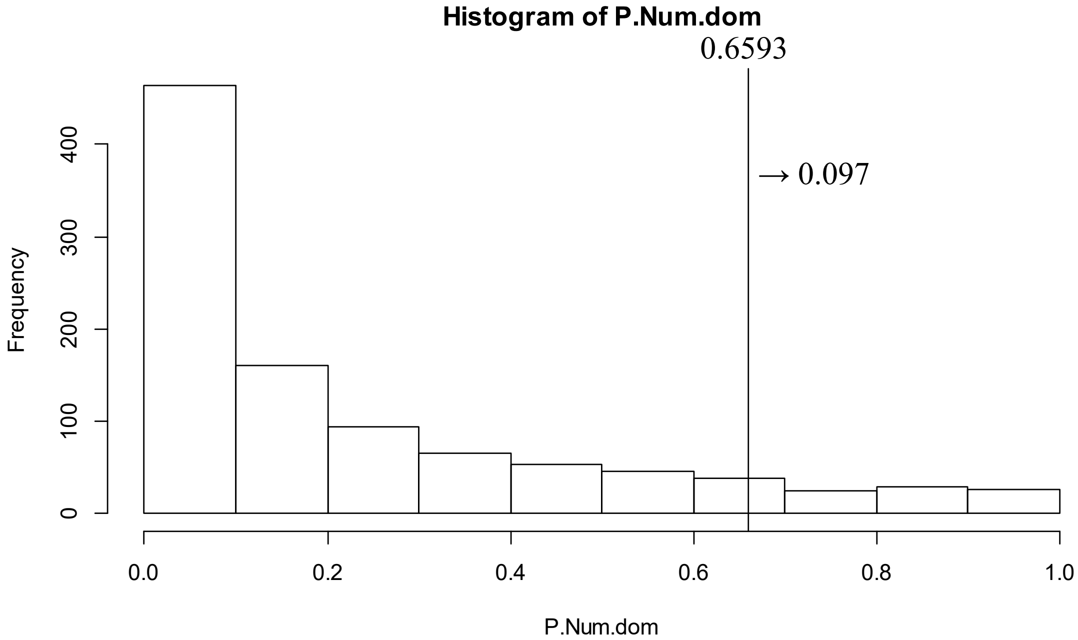
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value: 0.016



Num.dom

Synonymous model: P = 0.6593

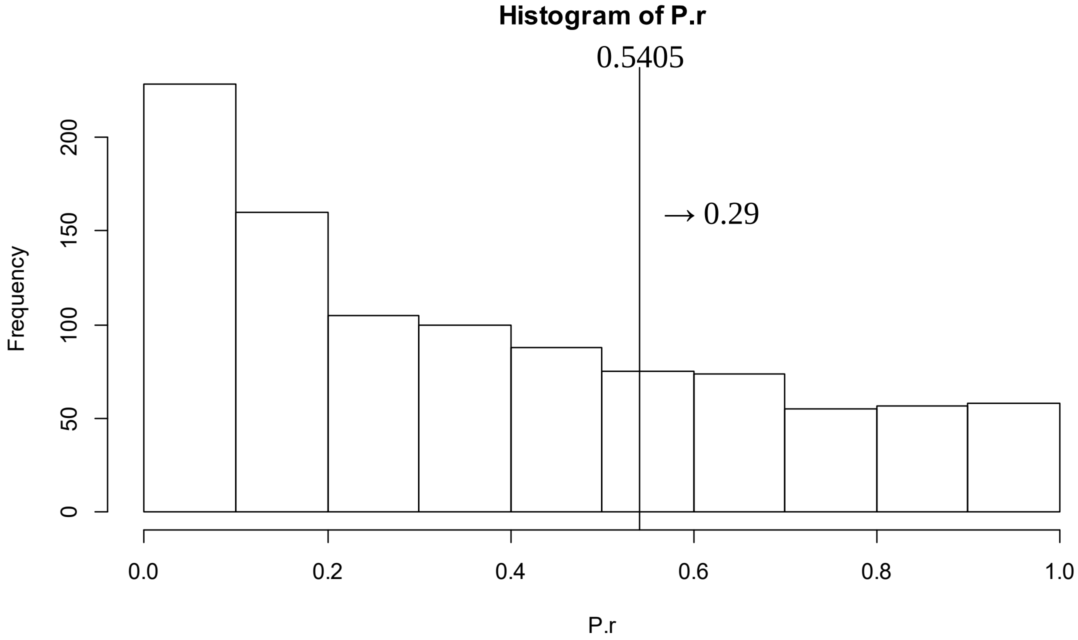
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value: 0.097



r

Synonymous model: P = 0.5405

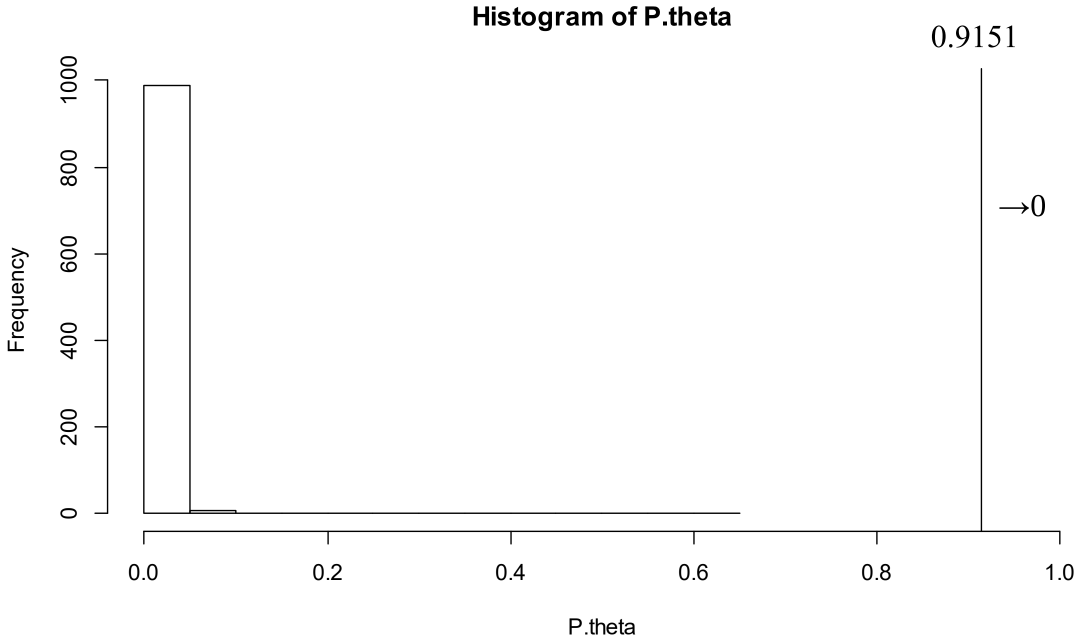
Proportion of re-sampled non-synonymous model P-values that are less than the synonymous P-value: 0.29



theta

Synonymous model: P = 0.9151

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


Histogram of P.theta
→0
Frequency
P.theta