EMBO reports

Manuscript EMBOR-2012-36362

Genomic evidence for elevated mutation rates in highly expressed genes

Chungoo Park, Wenfeng Qian and Jianzhi Zhang

Corresponding author: Jianzhi Zhang, University of Michigan

Review timeline:

Submission date:	05 July 2012
Editorial Decision:	09 August 2012
Revision received:	10 September 2012
Accepted:	05 October 2012

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 09 August 2012

Thank you very much for the submission of your research manuscript to our editorial office. I would like to apologize for the unusual delay in getting back to you with a decision on your manuscript, which was due to the summer season in which we allow referees more time to submit their reports. We have now received the full set of reviews on your manuscript.

As the detailed reports are pasted below I will only repeat the main points here. You will see that all reviewers acknowledge the potential interest of the findings and feel the study is suitable for publication in EMBO reports once their (rather minor) concerns have been addressed. They point out some instances in which further clarifications and discussions are needed. In addition, referee 2 raises two aspects of your analysis that would benefit from additional data (his/her points 4 and 5).

Overall, given these evaluations, the reviewers constructive comments and the potential interest of the study, I would like to give you the opportunity to revise your manuscript, with the understanding that the main concerns of the referees (as outlined above and in their reports) must be addressed. Acceptance of the manuscript will depend on a positive outcome of a second round of review and I should also remind you that it is EMBO reports policy to allow a single round of revision only and that therefore, acceptance or rejection of the manuscript will depend on the completeness of your responses included in the next, final version of the manuscript.

Revised manuscripts should be submitted within three months of a request for revision; they will otherwise be treated as new submissions. If you feel that this period is insufficient for a successful

submission of your revised manuscript I can potentially extend this period slightly. Also, the length of the revised manuscript should not exceed roughly 35,000 characters (including spaces). It currently exceeds this limit and it would therefore be good if the revised manuscript could be shortened slightly. This could, for example, be done by combining the results and discussion section, which would avoid redundancies. Should you find the length constraints to be a problem, you may consider including some peripheral data in the form of Supplementary information. However, materials and methods essential for the repetition of the key experiments should be described in the main body of the text and may not be displayed as supplemental information only.

I look forward to seeing a revised form of your manuscript when it is ready. Should you in the meantime have any questions, please do not hesitate to contact me.

Yours sincerely

Editor
EMBO Reports

REFEREE REPORTS:

Referee #1:

Comments on Park et al. "Genomic evidence for elevated mutation rates in highly expressed genes"

In this interesting manuscript the authors consider whether transcription is mutagenic or not, since it has been shown that transcription both elevates the rate of DNA lesions and the rate at which they are repaired. Using mutation accumulation lines and the divergence between species they show that the inferred rate of mutation is higher in highly transcribed genes. They also show that this pattern can be detected in mammals.

1. The authors show a significant correlation between expression and intron divergence (between human and macaque) but I wonder what the slope of this relationship is, because the difference in substitution rate between transcribed and non-transcribed DNA appears to be very slight (as in figure 2 in Hodgkinson and Eyre-Walker (2011 NRG)). If the slope of the line is very low then the relationship between transcription and mutation in mammals is only mildly interesting. Of course there appears to be quite a strong relationship between the two variables in the somatic tissue, but here the mutation rate of highly expressed genes is reduced relative to lowly expressed genes. The effect of transcription on the rate of mutation appears to be substantial in yeast.
2. In their first analysis they consider the expression level of sites that have accumulated mutations versus those which have not. They find that mutated sites have about twice the mutation rate of unmutated sites. It wasn't clear to me why they randomly chose 190 sites to get their random expectation; divide the sites into mutated and unmutated and compare their expression levels using a t-test or MW; this is likely to be a much more powerful test, even controlling for base composition. It may also give them enough power to control for whether the site is in a coding sequence or not; so it might be the case that coding sequences have a higher mutation rate than non-coding sites whether or not they are transcribed, so then mutated sites will have higher expression than non mutated sites, but this has nothing to with the level of transcription.
3. They consider the conditions under which a modifier can affect the rate of mutation. They consider the case of a modifier that affects a single linked locus, but I wonder how relevant this is, given that the processes they are discussing are more general - e.g. TCR. A modifier of the rate of TCR is likely to affect all expressed loci and selection against the modifier will be closer to delta_u*2*L*s_bar, where s_bar is the average effect of new mutations and L is the length of the genome. This actually raises the intriguing question as to when locus specific versus genome wide modifiers are more likely to be successfully selected; if L*s_bar > M, where M is the length of the locus then selection on genome wide modifiers will be stronger than on locus specific modifiers.

In summary I think this is an interesting manuscript.

It is my standard policy to sign my reviews

Adam Eyre-Walker

Referee #2:

The authors investigate the contributions of transcription associated mutagenesis and transcription coupled repair on mutation spectrum and rate which is of significant interest to the biological community and has implications on the selective effect of a mutation. A majority of the analysis is focused on mutations at intron sites in yeast and humans, and finds elevated mutation rates in highly expressed genes. Although the patterns provided may suggest transcription-associated mutagenesis, the significance is low, and there are certain factors contributing to these patterns that are not fully addressed. This paper is well written and does a good job of summarizing the previous work, however, due to the limited amount of data - especially in yeast, the significance of these results may be improved by extending to four-fold degenerate sites.

Major comments:

1) The authors first analyze mutations derived from a uracil-DNA glycosylase knockout line that has undergone mutation accumulation. The authors are concerned about selection so they filter the data set by expression knockout profiles of wt lines, and use the RNA-seq from another set of lines. One potential issue here is that the knockout of UNG1 has altered the expression profile of the line. Although this may not alter the genomic wide profile, there is very little control of expression here and needs to be commented on.
2) In the same section, the authors use a mean expression level to determine significance (standard deviation? -- should be in figure1). A single mutation in a highly expressed gene can skew the mean expression of the 190 mutations.
3) UNG1 KO increases the number of C->T mutations from spontaneous deamination, biasing the mutations towards G/C rich genes. Are those genes more highly expressed? It is unclear what the randomized 190 sites are selected from --- genes, IG? These issues can explain the observations without TAM or TCR and needs to be clarified.
4) The second analysis involves standard phylogenetic comparison between Saccharomyces species. The authors are again concerned about selection so they analyze only intron sites, but in the same paragraph suggest that selection increases with transcription in introns. If both introns and syn sites have some signature of selection in yeast, why not use 4-fold degenerate sites as well here? Same point for the human analysis.
5) There is limited information about the human analysis. Number of introns analyzed? In addition to clarity, one issue here is that by using a set distance from 3' UTR it is possible to capture enhancer regions that can skew the mean rate difference for across all tissues. The difference observed may be simply the selection differences in different tissues on enhancers 5kb downstream from genes. Surveying multiple untranscribed regions (5kb/10kb/20kb) for each intron can provide more confidence than a set region.

Minor changes.

Pg3:Non-temperate->non-template

Referee #3:

In the present manuscript, the authors investigated the correlation between gene expression levels and mutation rates in Saccharomyces species as well as primates, in order to clarify the net impact of transcription on a mutation rate at the genomic scale.
First, on the basis of the comparison between a wild strain and a mutation accumulation line of yeast, the authors showed that the mean expression level of the 190 mutated sites is significantly greater than the random expectation (fig. 1). This indicates that the rate of point mutation in a gene increases with the expression level of the gene. Next, they conducted inter-specific comparison of genome sequences of two Saccharomyces species and showed the positive

correlation between the expression level of a gene and the mutation rate that was estimated from a nucleotide substitution rate of its intron (table 1).

Furthermore, to understand what kind of Transcription-Associated Mutagenesis (TAM) mechanism regulates the mutation rates, the authors inferred all single nucleotide substitutions that occurred in the S. cerevisiae lineage since its separation from S. paradoxus.

Then, they calculated the difference between the frequency of each mutation type in introns and that in untranscribed regions of the genome. While four common mutation types all show significantly higher frequencies on the non-transcribed strand, only the C->T frequency is higher on the transcribed strand, indicating that two different TAM mechanisms simultaneously act on each strand.

Finally, using alignments of human and macaque genome sequences and human RNA-Seq data, the authors estimated the difference between the mutation rate of all introns of the gene and that of its flanking untranscribed regions. They found this difference to be positively correlated with the expression level (fig. 3). Based on these results, the authors claimed that at the genomic scale, the effect of TAM overwhelms that of transcription-coupled repair (TCR), and therefore, transcription is overall mutagenic in yeast and human.

In the manuscript, the authors focused on mutation rates in constitutive introns that are less affected by natural selection than synonymous substitution sites, and they used the data from not artificial reporter gene assays but the comparative analysis of actual genomes. These attempts by the authors clearly showed that transcription induces the mutations with different mechanisms of TAM depending on each DNA strand, implying that it opposes the previous reports.

These findings are highly meaningful, and therefore I would recommend acceptance of this manuscript for publication in EMBOR after correction of the following typos:

On page 3, Line 12: "non-temperate" should be "non-template".
On page 15, 5th line from the bottom: "Yeats" should be "Yeasts".
Figure 3: "Kindey" should be "Kidney".

1st Revision - authors' response 10 September 2012

Response to the reviewers

Referee #1

Comment 1:
The authors show a significant correlation between expression and intron divergence (between human and macaque) but I wonder what the slope of this relationship is, because the difference in substitution rate between transcribed and non-transcribed DNA appears to be very slight (as in figure 2 in odgkinson and Eyre-Walker (2011 NRG)). If the slope of the line is very low then the relationship between transcription and mutation in mammals is only mildly interesting. Of course there appears to be quite a strong relationship between the two variables in the somatic tissue, but here the mutation rate of highly expressed genes is reduced relative to lowly expressed genes. The effect of transcription on the rate of mutation appears to be substantial in yeast.

Response:
This is an excellent point. But, we cannot directly estimate the slope of the relationship between transcription level and mutation rate in humans due to the use of partial correlations to control multiple confounding factors. Instead, a multiple linear regression analysis was performed. Based on this analysis, doubling the expression level of an averagely expressed gene increases the mutation rate difference between its introns and flanking untranscribed regions by 15%. We have now added this result to page 9 of the main text and the methodological details to supplemental methods (page 4).

Comment 2:
In their first analysis they consider the expression level of sites that have accumulated mutations versus those which have not. They find that mutated sites have about twice the mutation rate of unmutated sites. It wasn't clear to me why they randomly chose 190 sites to get their random expectation; divide the sites into mutated and unmutated and compare their expression levels using a t-test or MW; this is likely to be a much more powerful test, even controlling for base composition. It may also give them enough power to control for whether the site is in a coding sequence or not; so it might be the case that coding sequences have a higher mutation rate than non-coding sites whether or not they are transcribed, so then mutated sites will have higher expression than non mutated sites, but this has nothing to with the level of transcription.

Response:
We believe that our original test is most accurate, because we randomly sampled 190 sites to evaluate the probability that the mean expression of the actual mutated sites is greater than that of randomly selected sites. This test requires no assumption of the underling distribution of the expression level among sites. The test suggested by the reviewer is approximate, because of the assumption of a *t* distribution of the mean expression. Nevertheless, we followed the suggestion to compare the mean expression of the mutated sites with that of the rest of the genome (after controlling the GC content). We found a *P*-value of 0.053 using Welch two-sample *t*-test. We also followed the reviewer's comment to compare coding and non-coding regions. Mutation rate is actually significantly lower in coding regions than non-coding regions ($P = 0.0018$; chi-square test). Apparently, it is high expression, rather than being coding, that boosts the mutation rate.

Comment 3:
They consider the conditions under which a modifier can affect the rate of mutation. They consider the case of a modifier that affects a single linked locus, but I wonder how relevant this is, given that the processes they are discussing are more general - e.g. TCR. A modifier of the rate of TCR is likely to affect all expressed loci and selection against the modifier will be closer to delta_u*2*L*s_bar, where s_bar is the average effect of new mutations and L is the length of the genome. This actually raises the intriguing question as to when locus specific versus genome wide modifiers are more likely to be successfully selected; if L*s_bar > M, where M is the length of the locus then selection on genome wide modifiers will be stronger than on locus specific modifiers.

Response:
The reviewer may have misread this section. We were actually discussing a modifier of gene expression level (rather than TCR) that impacts the mutation rate. There are many gene-specific modifiers of gene expression levels (e.g., *cis*-regulatory elements). In terms of modifiers of TCR, we agree that they are more likely to be genomic rather than gene-specific.

Referee #2

Comment 1:
The authors first analyze mutations derived from a uracil-DNA glycosylase knockout line that has undergone mutation accumulation. The authors are concerned about selection so they filter the data set by expression knockout profiles of wt lines, and use the RNA-seq from another set of lines. One potential issue here is that the knockout of UNG1 has altered the expression profile of the line. Although this may not alter the genomic wide profile, there is very little control of expression here and needs to be commented on.

Response:
As we described in the manuscript, *UNG1* encodes uracil-DNA glycosylase, which is used for repair of uracil in DNA formed by spontaneous cytosine deamination. Thus, deleting *UNG1* is not expected to alter the expressions of many genes in the yeast genome. Of course, it is possible that the expressions of a small number of genes are altered by the deletion. But, we believe that the use of the expression data from the wild-type strain should have made our conclusion more conservative, because the expression level differences between the two strains should have lowered

the power of our test. We have added in the text our assumption of similar transcriptomes between the wild-type and *UNG1*-deletion strains (page 5).

Comment 2:
In the same section, the authors use a mean expression level to determine significance (standard deviation? -- should be in figure1). A single mutation in a highly expressed gene can skew the mean expression of the 190 mutations.

Response:
The reviewer may have misunderstood our statistical test, because his/her concern has been taken care of by the statistical test used. Briefly, we compare the mean expression of the 190 mutated sites with that of 190 randomly sampled sites from the genome. Outliers can happen in both mutated sites and randomly sampled sites, and thus would not impact our test. The level of statistical significance is estimated by the fraction of 10,000 replications in which the mean expression of the randomly picked 190 sites equals to or exceeds the observed value. The standard deviation of the mean expression is not presented, because it is not used to calculate the *P*-value. Instead, the actual frequency distribution is presented (bars in Fig. 1).

Comment 3:
UNG1 KO increases the number of C->T mutations from spontaneous deamination, biasing the mutations towards G/C rich genes. Are those genes more highly expressed? It is unclear what the randomized 190 sites are selected from --- genes, IG? These issues can explain the observations without TAM or TCR and needs to be clarified.

Response:
We already controlled the number of G:C and A:T sites in the comparison of expression levels of mutated sites and randomly sampled sites (page 5). The random sites were sampled from the entire genome except the genic regions (from start to stop codons in gene sequences) of the genes under selection. We have clarified these points.

Comment 4:
The second analysis involves standard phylogenetic comparison between Saccharomyces species. The authors are again concerned about selection so they analyze only intron sites, but in the same paragraph suggest that selection increases with transcription in introns. If both introns and syn sites have some signature of selection in yeast, why not use 4-fold degenerate sites as well here? Same point for the human analysis.

Response:
Synonymous mutations are known to be subject to natural selection for preferred synonymous codons, especially in species with large population sizes (e.g., bacteria, yeast, and Drosophila). Although introns may contain regulatory sites, the expectation is that these sites constitute only a small fraction of intron sequences. In other words, the overall selection on introns is expected to be lower than that on synonymous sites in yeast. Thus, we analyze only introns in yeast. Note that the potential presence of constrained sites in introns is expected to reduce the positive correlation between mutation rate and expression level, suggesting that our conclusion is conservative.

Selection for biased codon usage is expected to be much weaker in humans than in yeast. We thus followed the reviewer's suggestion to examine synonymous sites. A positive correlation between d_S (synonymous substitution rate) and expression level is observed in the testis (new Fig. S4).

Comment 5:
There is limited information about the human analysis. Number of introns analyzed? In addition to clarity, one issue here is that by using a set distance from 3' UTR it is possible to capture enhancer regions that can skew the mean rate difference for across all tissues. The difference observed may be simply the selection differences in different tissues on enhancers 5kb downstream from genes. Surveying multiple untranscribed regions (5kb/10kb/20kb) for each intron can provide more confidence than a set region.

Response:

The introns of 3,538 human genes were subject to analysis (see Supplemental Methods). Although enhancers may occur in 3' regions at least 5 kb away from the end of 3UTR, the impact of such occurrences on our result is minimal because enhances would constitute only a tiny fraction of the 5 kb segment examined. Nevertheless, to satisfy the reviewer, we also examined 5 kb segments that are at least 10 kb away from the end of 3UTR. The results are similar (new Fig. S3).

Comment 6:
Minor changes.

Pg3:Non-temperate->non-template

Response:
Corrected.

Referee #3

Comment 1:
On page 3, Line 12: "non-temperate" should be "non-template".

Response:
Corrected.

Comment 2:
On page 15, 5th line from the bottom: "Yeats" should be "Yeasts".

Response:
Corrected.

Comment 3:
Figure 3: "Kindey" should be "Kidney".

Response:
Corrected.

2nd Editorial Decision 05 October 2012

I am very pleased to accept your manuscript for publication in the next available issue of EMBO reports. Thank you for your contribution to our journal.

At the end of this email I include important information about how to proceed. Please ensure that you take the time to read the information and complete and return the necessary forms to allow us to publish your manuscript as quickly as possible.

As part of the EMBO publication's Transparent Editorial Process, EMBO reports publishes online a Review Process File to accompany accepted manuscripts. As you are aware, this File will be published in conjunction with your paper and will include the referee reports, your point-by-point response and all pertinent correspondence relating to the manuscript.

If you do NOT want this File to be published, please inform the editorial office within 2 days, if you have not done so already, otherwise the File will be published by default [contact: emboreports@embo.org]. If you do opt out, the Review Process File link will point to the following statement: "No Review Process File is available with this article, as the authors have chosen not to make the review process public in this case."

Thank you again for your contribution to EMBO reports and congratulations on a successful

publication. Please consider us again in the future for your most exciting work.

Yours sincerely

Editor
EMBO Reports

