Dear Virus Evolution editorial board,

Please find attached our revised manuscript entitled "MERS-CoV recombination: implications about the reservoir and potential for adaptation". This is a resubmission of manuscript VEVOLU-2015-032. The editorial assessment stated that:

> Both reviewers were a little underwhelmed by the evidence of recombination and think that you might be misinterpreting complex patterns of rate heterogeneity (accross sites and also possibly accross lineages) as evidence of recombination. You should either more rigorously test this possibility or be a bit more cautious. Maybe use maciej's 3Seq method or try Bruin's Phitest - the latter seems to be particularly robust to rate hetergeneity and is quite powereful (IMHO more so than the non-parametric tests in LDHat).

We thank the editor and the two reviewers for comments and suggestions. We agree with the reviewers and the editor that rate heterogeneity can, and most likely is complex in the MERS-CoV genome. We have performed additional sequence simulations under a codon model, per reviewer #1s suggestion. As expected, the codon position specific constraint has resulted in a much higher proportion of homoplasic sites for clonally simulated codon sequences. We would like to point out, however, that even under this model MERS-CoV sequences are highly aberrant, having a considerably higher proportion of homoplasic sites. Only when rate heterogeneity in simulated sequences is extreme and unrealistic does the proportion of homoplasic sites starts approaching values observed in the actual MERS-CoV data. Even then, the distribution of homoplasy degrees is much shallower for simulated sequences. The most parsimonious explanation for this is the existence of a diverse population of MERS-CoV from which alleles have been drawn and recombined into genomes that eventually get sampled. The alternative, which is highly implausible, is that MERS-CoV has highly specific mutational preferences and undergoes identical mutations multiple times. We would also like to point out that although we have shown that many recombination tests fail spuriously, we have also shown that the tests appear to fail inconsistently, i.e. a simulated dataset that is recognized as being recombinant by one test might not be identified as such by another. In contrast, MERS-CoV has passed every test that we have tried so far, including 3Seq and PHI. We have additionally included more supplementary figures showing phylogenetic trees derived from portions of the MERS-CoV genome and trees recovered by GARD and highlighted the incongruity between them.

Point-by-point responses to the reviewers are attached.

Sincerely,
Gytis Dudas

## Reviewer responses

Original reviewer criticisms are in plain text. Our responses follow in **bold**.

## Reviewer 1

1. "If recombination really is frequent enough to break up long tracts of homoplasies, shouldnt it also be common enough that one would be likely to observe at least a few recent events that have not yet been broken up? Ones (like the bat sequence in Corman 2014a) that show phylogenetic evidence of different histories for different genomic regions?

The reason I mention this point is that although I find the methods used here to be very impressive, I remain somewhat skeptical about the strong conclusions of the authors that frequent recombination is certainly at play."

**We expected to see, and looked hard for, unbroken recent tracts of recombination but did not find any in the data we have. However, a study by Wang et al. (2015) was published whilst our manuscript was in review and identified the common ancestor of the Korean MERS outbreak as a recent recombinant, where around 6500 nucleotides have been transferred across a considerable genetic distance in mid-to-late 2014. One potential explanation for why we observe only short stretches of potentially recombinant alleles is that recombination tracts themselves are short. We list this as another possibility in the manuscript now:**

> **Thus in MERS-CoV we interpret extremely short homoplasy tracts as evidence of relatively frequent recombination. Alternatively, recombination tracts might be short and thus unable to transfer multiple informative sites across lineages.**

2. "For example, the piBUSS simulations partition rate heterogeneity into two contiguous genomic regions. But we know that there are complex site specific rate heterogeneity patterns throughout viral genomes, quite distinct from the simplistic rate heterogeneity modeled in these simulated sequences (not least because of differing constraints on different codon positions and different selective pressures within domains and even single aa sites within genes).

Given that the conclusion is that homoplasies are scattered not clustered because of rampant recombination, it seems that more realistic simulations are warranted. And, even then, I would suggest the authors insert something to the effect of,

> "we know that even minor mis-specifications in the modeling of simulated data can lead to highly significant false positive results. Given that the one thing we can be sure of is that the real evolutionary model was different from the ones used to simulate, we retain a degree of caution in our interpretation of the results. On the other hand, it is difficult to discern how such compelling results indicative of recombination could be spurious. At the very least, frequent

recombination may be a feature of MERS CoV evolution given our results, and further characterization of its role and the biological implications are warranted."

"

**We definitely agree with the reviewer that real life site specific rate heterogeneity is far more complicated than our piBUSS simulations. We intended piBUSS simulations to be deliberately exaggerated so as to highlight the pitfalls of the methods that we used. After further consideration we included more sequence simulations under a slightly more realistic model (codon model with purifying selection), which improves the analyses we present and allows for a more nuanced discussion. We thank the reviewer for bringing this up. We have added the following sentences to discuss additional results:**

> **Homoplasies become increasingly more prevalent when a more realistic codon model is used, due to differences in codon position constraint. Even then, MERS-CoV genomes possess mutations with much higher homoplasy degrees, surpassing simulated datasets with caricature levels of rate heterogeneity. This makes sense under a recombination scenario, as alleles persist in a diverse population and get recombined into novel backgrounds repeatedly, giving an appearance of highly repeatable mutations. Nevertheless, substitution patterns in real genomes are often highly complex and homoplasy-based methods have been shown to be susceptible to rate heterogeneity across sites, especially under higher levels of sequence divergence (Posada and Crandall, 2001). Although rate heterogeneity certainly exists in MERS-CoV data, the divergence levels are still quite low ($\theta$/site = 0.0047), giving us some degree of certainty in our inference of homoplasies. It is also reassuring that both maximum likelihood and Bayesian sequence reconstruction converged on similar patterns of homoplasy and synapomorphy across the genome (figures 3 and S6).**

3. "Is there biological evidence of dual infections in camels from sequence data? For example, only single sequence types (apparently) were recovered from camels in a high prevalence slaughterhouse setting (Infection Ecology and Epidemiology 2015, 5: 28305 - http://dx.doi.org/10.3402/iee.v5.28305). It may be worth following up with those authors to see if their molecular methods have have missed dual infections. But, if there really are only single infections in camels even in populations with antibody prevalence >50%, it should be noted in the current manuscript. My reading of that paper, where every camel virologically studied had just one, not >1, of the five sequence types circulating in the slaughterhouse, raises my suspicion a bit about the conclusion of rampant recombination. If you can't find dual infections in camels with the highest prevalence rates found anywhere in the middle east, then perhaps frequent recombination in camels is not likely. Maybe bats are different and more prone to dual infections? Does anyone know? "

**The reviewer raises a very good point. Dual infections should be detectable through sequencing, but we are not aware of any study reporting this. At the same time, however, sequencing of camels has been extremely limited in**

**comparison to the sequencing performed in humans. We have added this as a point to the discussion:**

> **A study by Adney et al. (2014) has shown that camels only suffer mild symptoms from MERS-CoV infection and numerous other studies indicate an extremely high prevalence of antibodies specific against MERS-CoV in camels (Müller et al., 2014; Corman et al., 2014b; Chu et al., 2014; Reusken et al., 2013, 2014). At the same time, however, sequencing has not indicated the presence of multiple infection in camels, or any other animal. We believe that individual MERS-CoV co-infections are rare, but given the size of the epidemic in camels, as inferred from serology, the total number of co-infections is high. In addition, MERS-CoV infection is transient in camels (Adney et al., 2014) and thus sequencing efforts, which have been insufficient and very limited in camels, are highly unlikely to capture a co-infection.**

4. "Minor point: comma splice in the sentence "This has two important consequences, one is that care has to be taken..."""

**We have changed the punctuation in this sentence. Thank you.**

## Reviewer 2

1. "Is there recombination in MERS-CoV?

Figures 3 and 4 seem to say yes.

But Figure 2 shows that the piBUSS-simulated clonal datasets seem to give positive recombination signals when rate heterogeneity is high."

**Figure 2 shows that piBUSS-simulated datasets tend to exhibit *extreme* values for each of the statistics tested, but none of the simulated datasets exhibit values for all 4 statistics that would be consistent with recombination. We have added the ideal positive and ideal negative cells to the figure to show that under recombination there is an expected direction for each statistic. Only MERS-CoV data exhibits values that are extreme for all 4 statistics *and* consistent with recombination. We have added additional text to make this more explicit.**

2. "And on page 10, a GARD analysis tells us that a Bayesian model comparison supports rate heterogeneity over recombination as an explanation for the alignment.

And, on the top of page 7: "5 out of 6 empirical simulation alignment simulated without recombination were identified by GARD as having having breakpoints..." If this is true, it means that GARD has a high likelihood of detecting false positive recombination signals."

**Technically GARD does not explicitly test for recombination, but for the presence of trees in the alignment whose combined likelihoods fit the data better than a single tree. For example, in the presence of rate heterogeneity alone the difference between the trees would mostly come down to branch lengths, not**

**tree topology. This is not a problem as long as GARD results are interpreted with caution. We have included this point in the text.**

3. "If rate heterogeneity is a better explanation for the data, then there does not seem to be enough evidence to support the presence recombination in the history of this sample.

( please let me know if I have misunderstood something here ).

I realize the following comment comes with some bias (which is why I am signing this review) but here it is anyway: the most sensitive and most specific way to detect presence or absence of recombination in a data set like this is to use the exact tests in 3SEQ (http://mol.ax/3seq). See Figure 2 in Boni et al (Genetics, 2007, 176:1035). I am happy to get the authors to set up on this analysis; no authorhip or acknowledgement needed."

**This is a very useful suggestion. We have included a new supplementary figure (figure S10) showing the results from 3Seq for MERS and simulated data.**

4. "The best and simplest way to demonstrate recombination is to build three phylogenies separated at breakpoints 12257 and 23722. This should be the central figure in the paper."

**We agree with the reviewer that clear phylogenetic incongruity is by far the most convincing evidence of recombination, however, Figure 6 of our manuscript highlights why we chose not to go down this path. Homoplasies within the MERS-CoV genome are common, most often occurring on their own and not in tracts, and do not appear to exhibit any specific patterns of allele flow between branches of the tree, hence any attempts at producing phylogenies from short fragments of the genome will only result in essentially random phylogenies that do not convey any meaningful information. Nonetheless, we include an additional figure in supplementary material showing phylogenies recovered from several 4000 nucleotide fragments of the MERS-CoV genome and highlight the inconsistencies in topology between these trees (figure S9).**

5. "You should state the drawbacks of homoplasy methods in the Discussion section. They are more susceptible to false positives than other methods when rate hetergeneity is high. See Posada & Crandall, PNAS, 2001, 98:13761."

**This is an important point, especially in light of the intuitive results from piBUSS simulations under a codon model. However, we also note that the overall low diversity of MERS-CoV should mitigate this somewhat. We discuss this in response to question 2 from reviewer #1.**

6. "Second paragraph of intro states that although co-infection is common for avian influenza, recombination is extremely rare of absent. This is true. The more appropriate citation for this is Boni et al (2010, PLoS One). The 2008 paper looks at human influenza viruses."

**Thank you for the updated reference.**

7. "For the list at the beginning of the methods section, you can cite Posada, Crandall, and Holmes (Annual Rev Genetics, 2002, 36:75) and explain that this is a modified version of their list. I understand that the techniques have changed since 2002, so some are omitted

here, but a fourth one worth adding in is "varying pattern of sequence identity along the genome" which is what many of the methods in RDP use. This is item (b) in the Posada-2002 paper."

**We have added a reference to the review by Posada et al in the methods section. However, we feel that a varying pattern of sequence identity along the genome, as a hallmark of recombination, is the same as excessive homoplasy.**

8. "Also, for the first bullet point in this list, you may want to explain to readers that a minumum of two trees are necessary for this analysis. Obvious, I know. But some readers need to hear this."

**We agree. We have modified the methods section accordingly:**

> **Recombination can be inferred by reconstructing two or more phylogenetic trees from a partitioned alignment and looking for topological incongruity between them. Strong support for at least 2 incompatible phylogenetic trees across well-defined breakpoints is usually the most convincing evidence of recombination.**

9. "Second paragraph on page 14 is not convincing. I think once point #1 is addressed, this paragraph should be re-written."

**Our follow up to the GARD analysis, using marginal likelihood estimates from BEAST, does not present anything surprising. As we mention in the text, a speckled pattern of recombination does not alter tree topology in a substantial way and topologies between neighboring genomic regions remain highly correlated. As such, a model including two phylogenies is heavily penalised for over-parameterization.**

10. "Host association aanalysis is intruiging but not discusses in the main text."

**Yes, our apologies for this. We discuss the figure briefly in its legend and decided against discussing it any further, partly because it is a negative result and partly because we do not believe that there is sufficient data, especially from camels, to pursue such an analysis and have much confidence in the outcome. We have previously included a sentence in the discussion, which we have since expanded:**

> **Furthermore, we do not detect any strong associations between host and particular alleles (figure S8), although we do not believe that there is a sufficient number of sequences from camels to have much confidence in this result.**