Supplemental information:
MERS-CoV recombination: implications about the reservoir and potential for adaptation

Gytis Dudas[1] **and Andrew Rambaut**[1,2,3]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [2]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA, [3]Centre for Immunology, Infection and Evolution at the University of Edinburgh, Edinburgh, UK

October 28, 2015

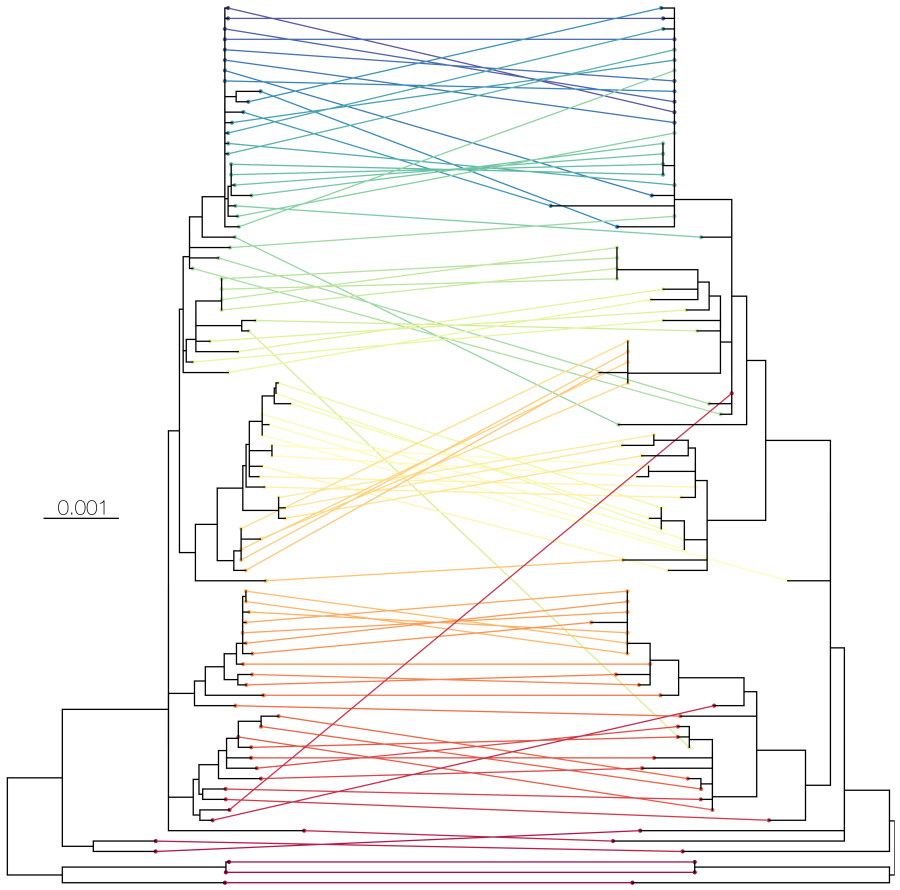

Figure S1. Trees recovered using GARD across the breakpoint at position 23722. NJ trees reconstructed by GARD across the first identified breakpoint. Tree from positions 1-23722 on the left and positions 23723-30126 on the right. The same tips in both trees are connected by coloured lines to indicate phylogenetic incongruity.

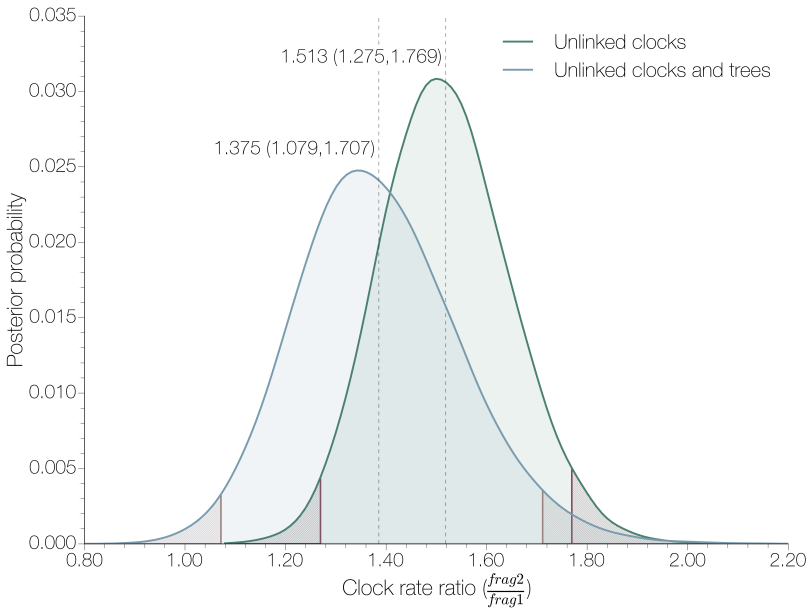

Figure S2. Empirical rate heterogeneity in MERS-CoV genome. Posterior estimates of the ratio between the molecular clock rates estimated independently from GARD-inferred fragment 2 (positions 23723-30126) and fragment 1 (positions 1-23722) under independent or linked tree models derived from 3 independent marginal likelihood analyses. Dotted lines indicate the mean of the distribution and numbers next to the line show the median and the 95% highest posterior density intervals.

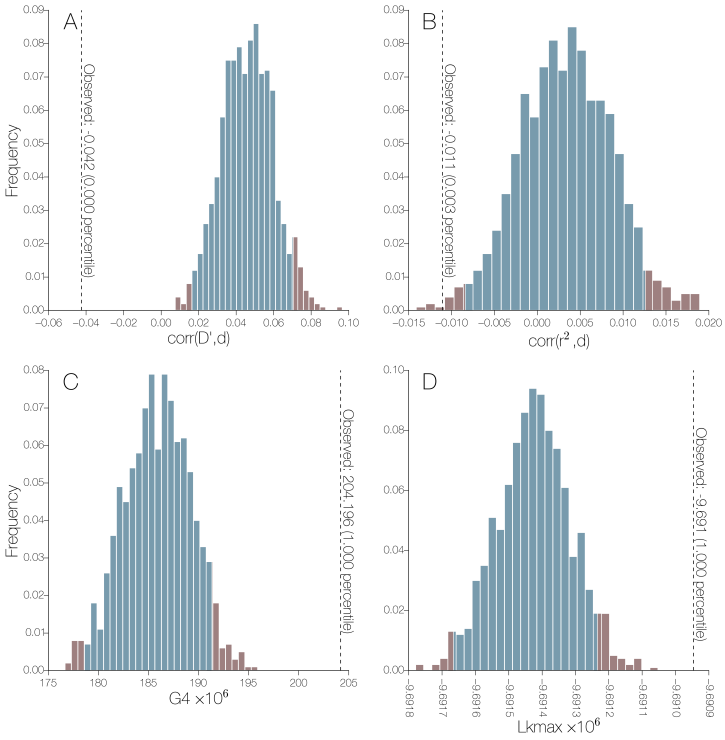

Figure S3. LDhat permutation test results for MERS-CoV. All 4 observed LD decay statistics (A - corr(D',d), B - corr(r^2,d), C - G4, D - Lkmax) for MERS-CoV data fall outside the distribution generated by permuting sites in ways consistent with recombination.

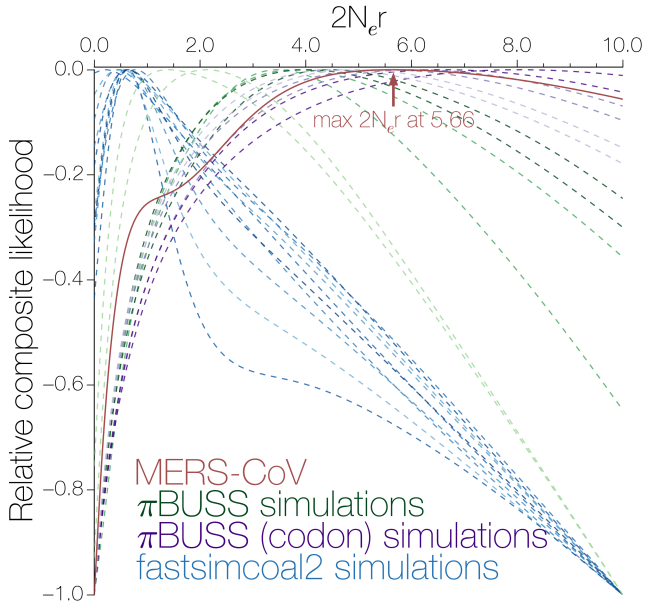

Figure S4. Relative composite likelihood surface. Composite likelihoods for the recombination rate estimates were rescaled to be within the range [-1,0]. Surfaces are coloured by data source: MERS-CoV estimate is in red, πBUSS simulations under a nucleotide substitution model, πBUSS simulations under a codon substitution model in purple and fastsimcoal2 simulations in blue. Colour scheme is identical to figure 2 in the main text. Maximum composite likelihood for MERS-CoV data is achieved at ρ=5.66, all other datasets have an inferred recombination rate above 0 despite being simulated without recombination.

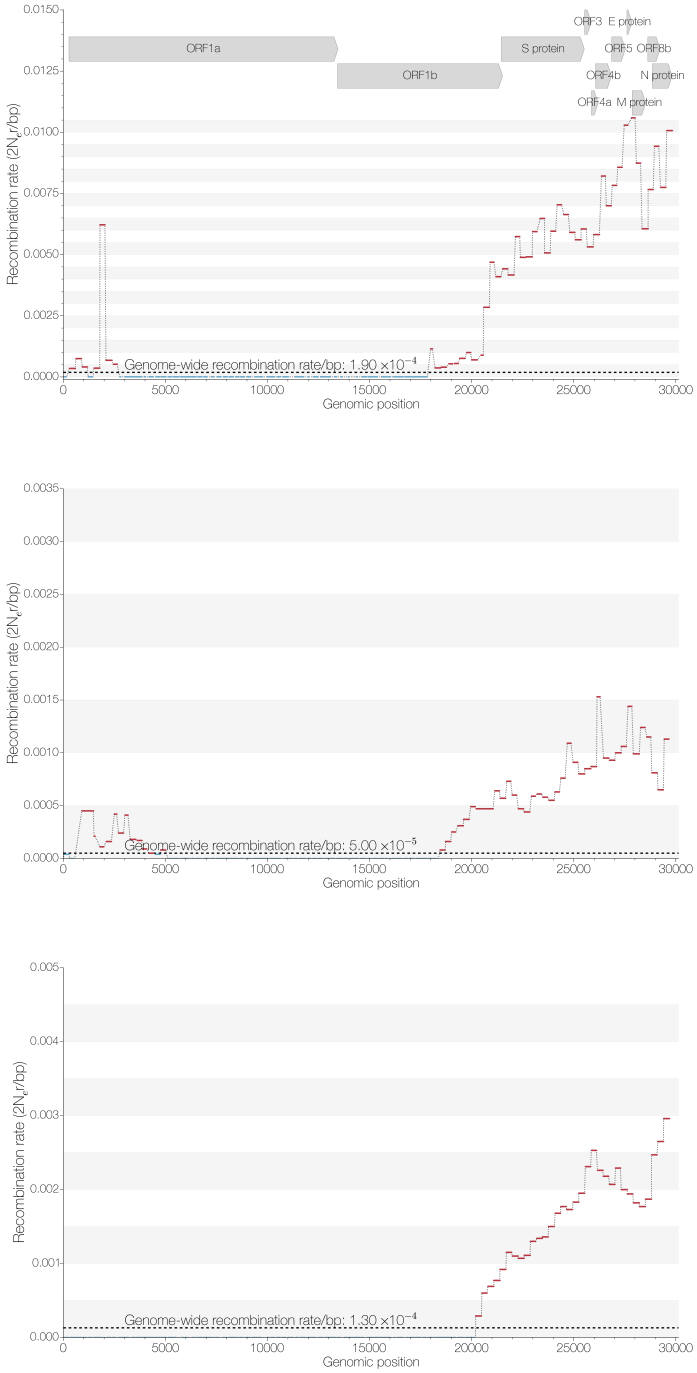

Figure S5. Window-based estimates of recombination rate. Inferred recombination rates for 300 nucleotide-long windows in MERS-CoV genome (top), πBUSS-simulated sequences with 1.3× rate heterogeneity (middle) and 3× rate heterogeneity (bottom) under a nucleotide substitution model. Recombination rates that are above the inferred genome-wide recombination rate are in red. Simulated rate heterogeneity is sufficient to mislead this method, although the inferred recombination rates in the last third of the MERS-CoV genome are much greater than those inferred from the simulated data.

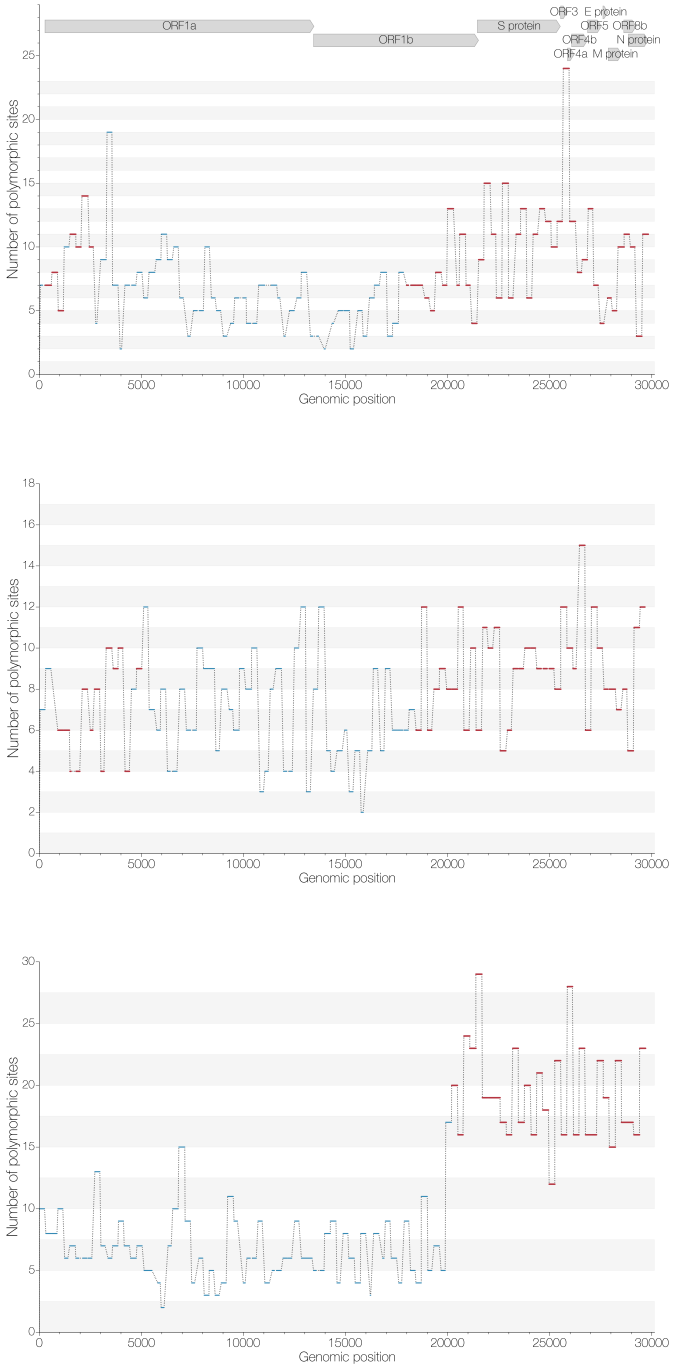

Figure S6. Window-based estimates of polymorphic site density. Inferred polymorphic site densities for 300 nucleotide-long windows in MERS-CoV genome (top), πBUSS-simulated sequences with 1.3× rate heterogeneity (middle) and 3× rate heterogeneity (bottom) under a nucleotide substitution model. Windows are coloured red if their recombination rate is above the inferred genome-wide recombination rate. Extreme rate heterogeneity (3×) results in a higher density of polymorphic sites in the region with the higher rate.

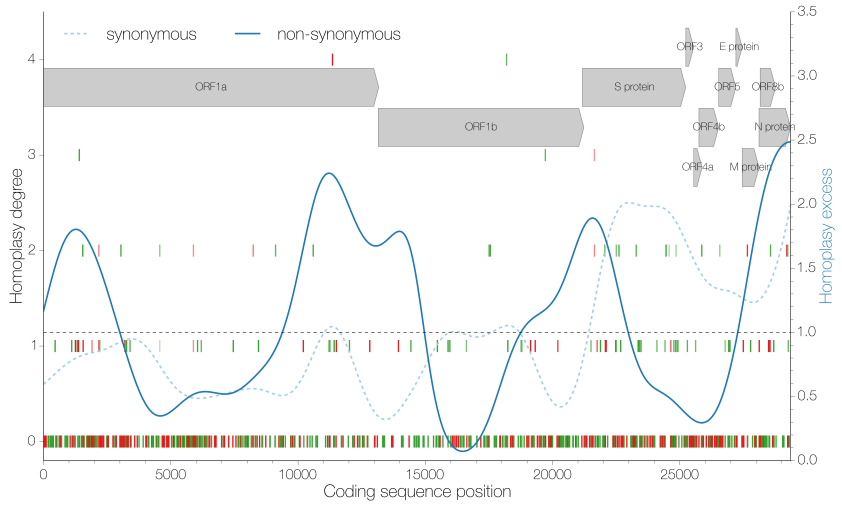

Figure S7. Homoplasy degrees inferred by BEAST. Position along the genome is shown on the x axis and homoplasy degree, the number of times a particular mutation has occured in excess in the tree, is shown on the y axis. Individual mutations are marked by vertical lines, synonymous ones in green and non-synonymous in red with transparency representing the posterior probability of a given homoplasy degree for each mutation. The ratio of apparent homoplasy over synapomorphy kernel density estimates (bandwidth=0.1) is shown in blue for synonymous (dashed) and non-synonymous (solid) sites separately. Arrows at the top indicate the positions and names of coding sequences within the MERS-CoV genome.

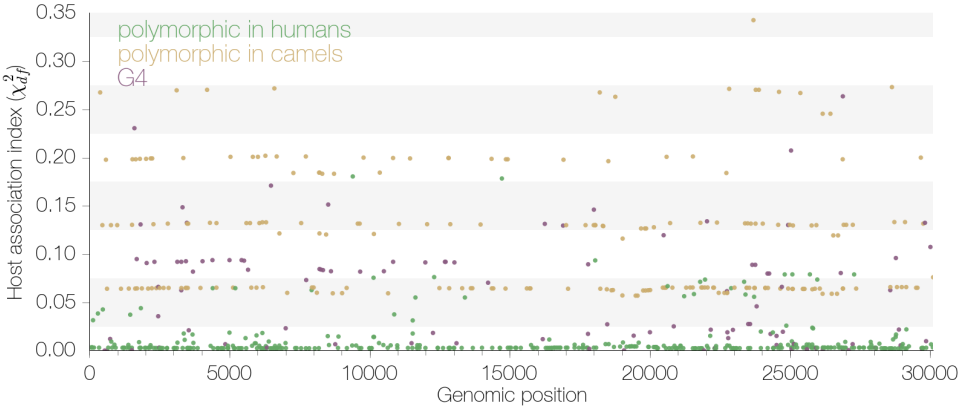

Figure S8. Host association indices for variable sites. Estimates for the association between particular alleles and host. The association index is an adapted version of the χ^2_{df} statistic of LD, and quantifies how well one can predict the allele at any given polymorphic site, given the host it was isolated from. No perfect associations (association index = 1.0) between particular alleles and host (human or camel) were found.

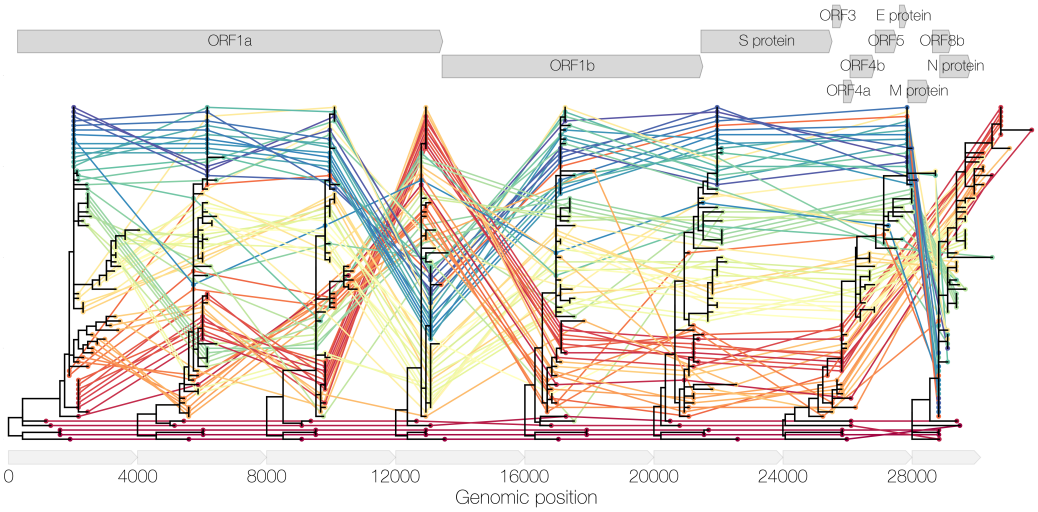

Figure S9. Maximum likelihood phylogenies across MERS-CoV genome. Maximum likelihood phylogenies recovered with PhyML (Guindon and Gascuel, 2003) under GTR+Γ_4 (Tavaré, 1986; Yang, 1994) nucleotide substitution model across 4000 nucleotide fragments derived from the MERS-CoV genome. Each tip is connected to its counterpart in phylogenies of neighboring fragments and coloured sequentially according to the order in which tips appear in the first fragment. Arrows at the top indicate the relative positions, lengths and names of coding sequences in the MERS-CoV genome, arrows at the bottom indicate the relative lengths of fragments used to produce the phylogenies.

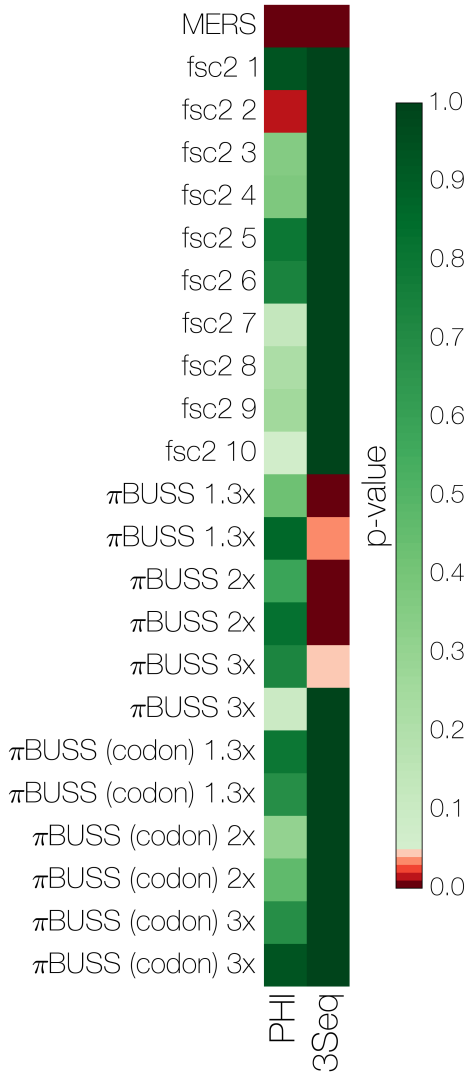

Figure S10. Results (p-values) from pairwise homoplasy index (PHI) and 3Seq analyses on MERS and simulated datasets. Both PHI and 3Seq analyses indicate that there is strong evidence of recombination in MERS-CoV (PHI p-value and Bonferroni-corrected 3Seq p-values <0.05, in red). Some simulated datasets are spuriously identified as recombinant by either PHI or 3Seq, but not both.

References

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology. 52:696–704.

Tavaré S. 1986. Some Mathematical Questions in Biology: DNA Sequence Analysis. Lectures on Mathematics in the Life Sciences, volume 17, 57-86. American Mathematical Society.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. Journal of Molecular Evolution. 39:306–314.