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

April 30, 2015

Abstract

Recombination is a process that unlinks neighbouring loci allowing for independent evolutionary trajectories within genomes of many organisms. If not properly accounted for, recombination can compromise many evolutionary analyses. In addition, when dealing with organisms that are not obligately sexually reproducing, recombination gives insight into the rate at which distinct genetic lineages come into contact. Since June, 2012, Middle East respiratory syndrome coronavirus (MERS-CoV) has caused 1106 laboratory-confirmed infections, with 421 MERS-CoV associated deaths as of April 16, 2015. Although bats are considered as the likely ultimate source of zoonotic betacoronaviruses, dromedary camels have been consistently implicated as the source of current human infections in the Middle East. In this paper we use phylogenetic methods and simulations to show that MERS-CoV genome has likely undergone numerous recombinations recently. Recombination in MERS-CoV implies frequent co-infection with distinct lineages of MERS-CoV, probably in camels given the current understanding of MERS-CoV epidemiology.

Introduction

Recombination is an important process which expedites selection in many organisms (Muller, 1932) by unlinking loci. It also leads to different parts of recombining genomes to have different histories which, if not properly accounted for, can interfere with many genetic analyses, of which phylogenetic methods are amongst the most sensitive. Not accounting for recombination in phylogenetic analyses leads to incorrect (Schierup and Hein, 2000) and poorly supported genealogies (Posada and Crandall, 2002) and false inference of selection (Anisimova et al., 2003; Shriner et al., 2003).

With rising sequence availability during outbreaks of viral infectious disease, phylogenetic methods have been used to supplement our knowledge of epidemics in real time (Smith et al., 2009; Rambaut and Holmes, 2009; Lemey et al., 2009; Drosten et al., 2013; Cotten et al., 2013, 2014; Drosten et al., 2014; Gire et al., 2014). For some outbreaks there is little reason to suspect recombination, *e.g.* negative sense single stranded RNA viruses are thought to recombine over evolutionary, not population-level, time scales (Chare et al., 2003). Observable recombination in RNA viruses requires that two conditions are met: that viruses from distinct lineages co-infect a host and that a mechanism for recombination exists. For example, even though influenza A virus co-infection is extremely common in birds based on genome segment reassortment patterns (Li et al., 2004; Dong et al., 2011; Lu et al., 2014), recombination is extremely rare or absent (Chare et al., 2003; Boni et al., 2008). This is thought to be because template switching (Kirkegaard and Baltimore, 1986; Baric et al., 1987), the main mechanism of recombination in RNA viruses, is mechanistically difficult for single stranded negative sense RNA viruses (see Chare et al. 2003), and for influenza A viruses has only been convincingly shown in cell culture under extreme conditions (Mitnaul et al., 2000). When the genomic architecture of a virus is permissive to recombination, *i.e.* template switching occurs and is detectable, the extent of recombination is informative of co-infection and/or duration of infection.

Here we focus our attention on the Middle East respiratory syndrome coronavirus (MERS-CoV) (Zaki et al., 2012), a recent zoonotic infection with a relatively high case fatality ratio (Cauchemez et al., 2014; Memish et al., 2013; Assiri et al., 2013). Most human infections with MERS-CoV are thought to be the result of contact with *Camelus dromedarius* L., the dromedary camel, which is the presumed host of the virus. MERS-CoV, much like Severe acute respiratory syndrome coronavirus (SARS-CoV), is likely ultimately derived from bats (Corman et al., 2014a). MERS-CoV, along with Murine hepatitis virus and SARS-CoV, belongs to the Betacoronavirus genus. Betacoronavirus, as well as two other genera (Alpha- and Gammacoronavirus) out of four within the subfamily *Coronavirinae* have been shown to recombine in cell culture, *in vivo* and in eggs (Lai et al., 1985; Makino et al., 1986; Keck et al., 1988; Kottier et al., 1995; Herrewegh et al., 1998). Additionally, a coronavirus lineage related to MERS-CoV which was isolated from bats appears to have recombined around the spike (S) protein (Corman et al., 2014a). In this paper we show that although the genome of MERS-CoV contains considerable amounts of rate heterogeneity between genomic regions that can interfere with detection of recombination, we do nonetheless find evidence of sustained recombination that cannot be explained by rate heterogeneity alone. This has two important consequences, one is that care has to be taken when constructing

phylogenetic trees of MERS-CoV, as a single tree cannot accurately describe the complete history of all loci within a recombining genome. Secondly and more importantly, the observed rates of recombination in the MERS-CoV genome is evidence of a large number of MERS-CoV co-infections in some hosts, which has implications for understanding the dynamics of the virus in the animal reservoir.

Methods

Overview

Recombination leaves several characteristic clues in genomes:

- Alternative topologies (Robertson et al., 1995a,b; Holmes et al., 1999). In some scenarios, for example if there has been a single large-scale recombination event, it is possible to clearly identify recombining fragments based on phylogenetic incongruity. Strong support for incompatible phylogenetic trees across well-defined breakpoints is usually the most convincing evidence of recombination.
- Excessive homoplasies (Maynard Smith and Smith, 1998). The transfer of genetic material from one genetic background to another will result in apparent repeat mutations in different parts of a phylogenetic tree. However, it is possible for the same locus to undergo mutation independently, especially if the locus in question is under Darwinian selection. Detecting homoplasies alone is not sufficient to infer recombination, but should be demonstrated to occur in excess of expectation.
- Linkage disequilibrium (LD) decay (Meunier and Eyre-Walker, 2001). Linkage disequilibrium or LD is the non-random association of alleles at different loci. This is a statistic often reported for contemporaneous sequence data. In clonally (*i.e.* non-recombining) evolving organisms every allele is linked to every other allele in the genome and requires mutation to break linkage. In recombining organisms there is an expectation that LD will decay with distance between the loci, *i.e.* that loci further away from each other are more likely to be unlinked via recombination.

We test for each of these hallmarks of recombination in the MERS coronavirus genome using a combination of phylogenetic and linkage disequilibrium metrics.

Alternative topologies

We use the Genetic Algorithm for Recombination Detection (GARD) method (Kosakovsky Pond et al., 2006), as implemented in the software package HyPhy (Pond et al., 2005), to look for alternative tree topologies in sequence data. Briefly, the method compares a model where a single tree is derived from the whole alignment and alternative models where breakpoints are introduced into the alignment and phylogenetic trees are derived independently from the resulting fragments. The presence of recombination, especially if it is recent and concentrated in some parts of the alignment, will result in two or more phylogenetic trees fitting the data better than a single tree model. We use GARD under a

GTR (Tavaré, 1986) substitution model with Γ_4-distributed rate heterogeneity amongst sites (Yang, 1994) on a dataset of 85 MERS-CoV sequences. GARD was run repeatedly until no more breakpoints could be identified in the resulting fragments.

In addition to this test, we run BEAST (Drummond et al., 2012) on partitioned coding sequences derived from the first well-supported breakpoint inferred by GARD. We extracted the coding sequences from nucleotide positions 1-23722 and 23723-30126 (referred to as fragment 1 and 2, respectively) of MERS-CoV genomes. Independent HKY+Γ_4 (Hasegawa et al., 1985; Yang, 1994) nucleotide substitution models were specified for codon positions 1+2 and 3 and the analyses were run under an uncorrelated relaxed lognormal clock with an uninformative CTMC reference prior (Ferreira and Suchard, 2008) on the substitution rate for 100 million states, subsampling every 10000 states. The molecular clocks and trees of each genomic partition were either linked or unlinked, giving a total of 4 models. We used the multi-locus skygrid (Gill et al., 2013) as the demographic model for all analyses. Path-sampling and stepping stone sampling (Baele et al., 2012) were used to calculate marginal likelihoods and test the fit of each of the 4 models, under default parameters. In addition, 4 similar analyses were set up, but with strict molecular clocks, in order to contrast the performance of relaxed molecular clocks.

Excessive homoplasies

Testing for recombination by looking for homoplasic mutations in phylogenetic trees requires that two conditions are met. One, that recombination is rare enough, so that there is sufficient phylogenetic signal to reconstruct the "correct" phylogeny otherwise known as the clonal frame (Milkman and Bridges, 1990). Two, that alternative explanations for homoplasic mutations can be dismissed with some certainty. There is no straightforward way of testing for the former, but the latter is usually dictated by the underlying biology. For example, repeat amino acid substitutions are a well documented response of influenza viruses and HIV to drug treatment (Gubareva et al., 2001; Tisdale et al., 1993; Boucher et al., 1993).

We employ two methods to test for excessive homoplasies. First, we use a maximum likelihood phylogeny inferred using PhyML (Guindon and Gascuel, 2003) under a GTR+Γ_4 (Tavaré, 1986; Yang, 1994) nucleotide substitution model to recover a single tree using a MERS-CoV dataset comprised of 85 sequences. We then reconstruct ancestral sequences at each internal node and identify the mutations that have taken place along each branch using ClonalFrameML (Didelot and Falush, 2007). Mutations are then classified as either synapomorphies, shared variation derived via common descent or apparent homoplasies, shared variation derived from convergence, depending on how many times a given mutation has arisen in the phylogeny. The drawback of this method is that it necessarily conditions on a single tree with the highest likelihood.

We also employ BEAST (Drummond et al., 2012) to circumvent the limitation of conditioning the ancestral state reconstruction on a single tree. In addition to sampling various phylogenetic parameters from the posterior distribution BEAST is also able to map substitutions onto the branches of each MCMC-sampled phylogeny (O'Brien et al., 2009). This method is thus capable of estimating the posterior probability of a given mutation

being synapomorphic or homoplasic by integrating over different tree topologies. Homoplasy analyses were performed on the concatenated coding sequences of MERS-CoV after partitioning the alignment into all 3 codon positions, each with an HKY nucleotide substitution model (Hasegawa et al., 1985) and no Γ-distributed rate heterogeneity amongst sites. A relaxed uncorrelated molecular clock with lognormally distributed rates (Drummond et al., 2006) under a CTMC reference prior (Ferreira and Suchard, 2008) and the flexible multi-locus skygrid as the demographic model (Gill et al., 2013) were used. The MCMC chain was run for 100 million steps, sampling every 10000 steps.

Throughout the paper we will refer to the number of branches that have experienced a given mutation as homoplasy degree. We define the homoplasy degree to be the number of times a given mutation has originated independently minus one. For example a homoplasy degree of 1 indicates that a mutation has occured on 2 different branches in the phylogeny. That is, we assume that one of the mutations has arisen through replication error, whereas the other has potential to have been introduced via recombination and thus can be thought of as excessive. Synapomorphies, on the other hand, are states that are shared by two or more taxa through common descent and thus necessarily are those mutations that have occured exactly once in the phylogeny. They have a homoplasy degree of 0 in all figures.

Linkage disequilibrium decay

In the absence of recombination every allele should exhibit a high degree of linkage with other alleles in the genome. Under two extremes - clonal reproduction without recombination and free recombination - there is no correlation between LD and genomic distance and loci should be interchangeable. This is the basis of several non-parametric permutation tests for recombination that are implemented in the software package LDhat (McVean et al., 2002), which we used in combination with a dataset of 109 MERS-CoV genomes. Other, more complicated tests, such as composite likelihood methods, are also available but in our experience were incompatible with temporal sampling and rate heterogeneity.

Sequence simulations

To test the performance of some of the methods we simulated two sets of sequences. We use fastsimcoal2 (Excoffier et al., 2013) to simulate 10 replicate datasets that have the same dates of isolation and similar diversity to the MERS dataset with 85 sequences under no recombination.

Additionally, we use πBUSS (Bielejec et al., 2014) to simulate sequences down an MCMC-sampled phylogeny drawn at random from a linked-tree unlinked-clocks BEAST analysis described above. We modelled region-specific rate heterogeneity by simulating a 30kb "genome" and setting the molecular clock rate for the first 20kb to be 9.5×10^{-4} substitutions site^{-1}year^{-1} and the last 10kb to be 2.85×10^{-3}, 1.9×10^{-3} or 1.3×10^{-3} substitutions site^{-1}year^{-1}, corresponding to roughly 3-, 2- or 1.3-fold rate heterogeneity between the two parts of the simulated genome. Two replicate datasets were generated for each category of rate heterogeneity. Other than that all simulations were run under a relaxed lognormal molecular clock (Drummond et al., 2006) with standard deviation set to 7.42×10^{-7},

HKY substitution model (Hasegawa et al., 1985) with the transition/transversion ratio parameter (κ) set to 6.0 and Γ-distributed rate heterogeneity with 4 categories and shape parameter 0.04 and empirical nucleotide frequencies, all derived from the results of the marginal likelihood analyses described earlier. A MERS-CoV sequence isolated from a camel (NRCE-HKU270) was provided as the starting state at the root. As these sequences were simulated on a tree of MERS-CoV, we refer to this dataset as being empirically simulated.

We also reconstructed ancestral states for these sequences using ClonalFrameML, as described above, to arrive at a null expectation for the number of homoplasies we expect to observe under rate heterogeneity but without recombination.

Investigating the effects of temporal sampling and rate heterogeneity

All 10 sequence datasets simulated with fastsimcoal2 and 6 sequences empirically simulated in πBUSS were analyzed using LDhat (McVean et al., 2002) to ascertain the effects of temporal sampling, and in the case of πBUSS-simulated sequences, the effects of rate heterogeneity. Additionally, empirically simulated sequence datasets were run through GARD (Kosakovsky Pond et al., 2006), since the method considers both differences in tree topology and branch lengths when calculating the likelihoods of trees. Stark rate heterogeneity could thus easily be mis-interpreted as evidence for recombination by GARD.

Host-association alleles

In order to test for the presence of alleles associated with host shifts (presumably camel to human) we adapt the χ^2_{df} (Zhao et al., 2005) statistic of LD to estimate the association between host (camel or human) and alleles at polymorphic loci. Briefly, we consider the host to act as a polymorphic site (encoded as H or C, for human and camel, respectively) and compare the association between the "allele" or host and alleles at polymorphic sites. A perfect association of 1.0 could mean, for example, that a biallelic site has one allele that is only found in camel viruses and the other allele only in human viruses.

Results

MERS-CoV genome shows evidence of alternative tree topologies

GARD identified a breakpoint at nucleotide position 23722 (corrected ΔAIC=103.6 between single versus two tree model), roughly in the middle of the coding sequence for the S (spike) protein. Running the resulting fragment 1 (positions 1-23722) and fragment 2 (positions 23723-30126) through GARD again yielded a further breakpoint in fragment 1 at position 12257 (corrected ΔAIC=33.7), near the boundary between ORF1a and ORF1b genes. No more breakpoints could be identified by GARD in the resulting fragments 1.1 (positions 1-12257), 1.2 (positions 12258-23722) and 2 (positions 23723-30126).

5 out of 6 empirical simulation alignments simulated without recombination, were identified by GARD as having breakpoints around position 20000, where the clock rate for the rest of the "genome" was increased to be 1.3, 2 or 3 times higher than the first 20kb. Corrected ΔAIC values decreased with decreasing rate heterogeneity, indicating loss of statistical power to detect differences between genomic regions. Analyses in BEAST where the MERS-CoV genome is partitioned into positions 1-23722 and positions 23723-30126 (corresponding to the first GARD-inferred breakpoint) with each partition having an independent molecular clock rate but the same tree or both independent molecular clock rates and independent trees, showed that rate heterogeneity as expressed by the ratio of second fragment rate to first fragment rate to be on the order of 1.513 (95% highest posterior density 1.275, 1.769) for unlinked clocks and 1.375 (95% HPDs 1.079, 1.707) for unlinked clocks and trees (see figure S1).

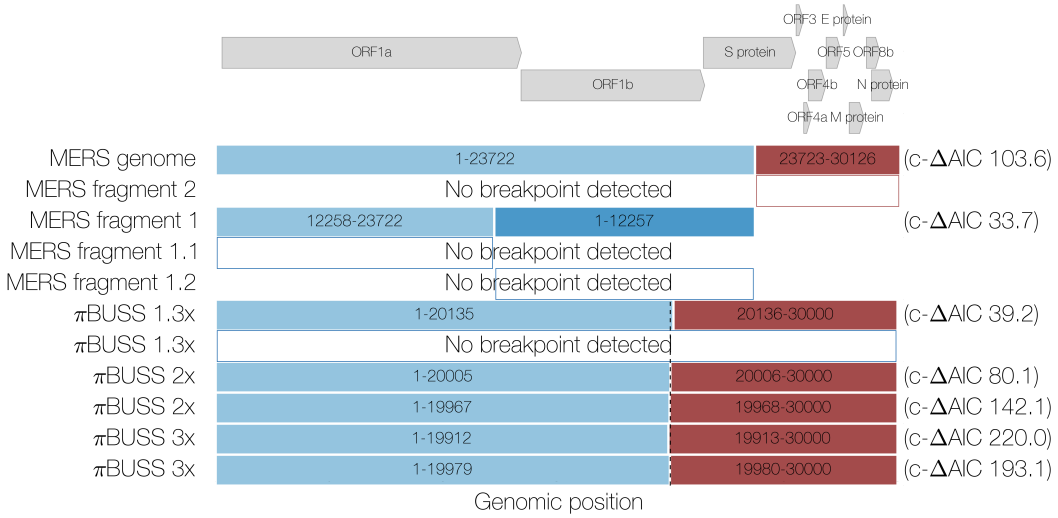

Figure 1. Summary of GARD results. Coloured boxes indicate fragments resulting from GARD-inferred breakpoints with corrected Δ-AIC values shown on the right. Dashed line indicates the actual position where the evolutionary model for simulated sequences under 3 levels of rate heterogeneity is changed. Arrows at the top indicate the positions and names of coding sequences within the MERS-CoV genome.

MERS-CoV genome exhibits linkage disequilibrium decay

Permutation tests as implemented in LDhat work under the assumption that loci are interchangeable only when there is free recombination or no recombination at all. The tests compare 4 statistics estimated from the actual data to 1000 permutations of the data where site numbers for each locus are reshuffled. Correlation coefficient between two measures of LD, r^2 (Hill and Robertson, 1968) and D' (Lewontin, 1964), are expected to show a negative correlation with increasing distance between loci if there is recombination. Permutation of recombining loci will produce a distribution skewed towards more positive

values for these two LD statistics and the percentile of the actual observed value can then be used to assess significance.

G4 is the sum of distances between pairs of loci with 4 observed haplotypes, which can only occur if there is repeat mutation or recombination at one of the loci. Under recombination the observed G4 statistic should take a statistically higher value in a distribution of G4 values derived from permuted data. Lkmax is the composite likelihood of pairs of loci under an estimated recombination rate and a given level of sequence diversity. Like the G4 statistic, this statistic is expected to fall in the upper tail of the distribution derived from permuted data in the presence of recombination.

All 4 permutation tests show a consistent signal of recombination in the MERS-CoV genome (see figures 2 and S2). Data from fastsimcoal2 simulations, which did not have rate heterogeneity, produced values for these statistics which mostly fell inside the range of values generated by permuting the simulated data, as expected (figure 2). On 1 occassion this is not the case – simulation 9 passed the Lkmax test and failed the other three. Empirically simulated data, on the other hand, tended to exhibit extreme values, that is the observed value fell below the 2.5th or above the 97.5th percentile of the permuted data, but in ways which were not consistent with recombination. For example, replicate 1 of simulation with 3-fold rate heterogeneity exhibits extreme values for all 4 tests, but only one of these – corr(r^2,d) is consistent with recombination.

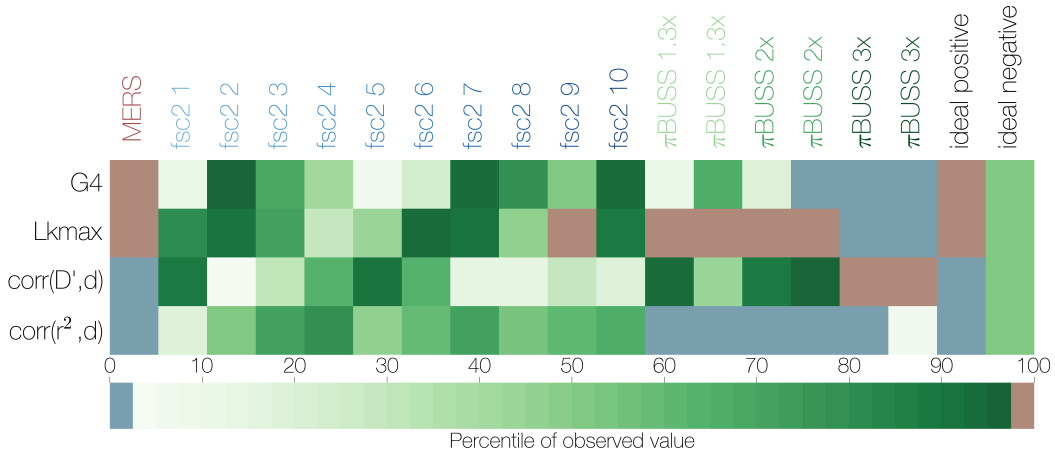

Figure 2. Summary of non-parametric tests for recombination. The percentile of the observed value for 4 statistics of LD decay (y axis) in the distribution of permuted datasets is indicated by colour. Sequence datasets are shown on the x axis, starting with MERS-CoV sequences, followed by 10 fastsimcoal2-simulated datasets and 6 empircally simulated datasets with different degrees of rate heterogeneity. Expected values for ideal datasets are shown in the last two columns, an ideal positive corresponds to the presence of recombination. Values falling between the 2.5th and 97.5th percentile are shown in green, values falling below the 2.5th percentile are in blue, those that are above the 97.5th percentile in red.

Composite likelihood methods are susceptible to rate heterogeneity

The composite likelihood method, which finds the composite likelihood surface of recombination rate, inferred non-zero recombination rates for all simulated datasets (see figure S3), revealing some degree of susceptibility to both temporal sampling and rate heterogeneity. A window-based approach of this test shows a sharp increase in the recombination rate estimated within 300 nucleotide windows around nucleotide 21000, close to the breakpoint inferred by GARD (see figure S4). We recovered a qualitatively similar pattern when analyzing empirically simulated sequences. It is important to note, however, that none of the simulated data, even under extreme heterogeneity, reproduced the same scale of the estimated recombination rate. Whereas in MERS-CoV data the vast majority of 300 nucleotide windows after position 23000 have a recombination rate per base consistently higher than 0.005, only data simulated under extreme rate heterogeneity approach values as high as that.

In addition to the apparently higher recombination rate in regions with higher rates we expect rate heterogeneity to produce a higher density of polymorphic sites in regions that are evolving faster. This is quite obvious in empirically simulated data with 3-fold rate heterogeneity – the region with higher rate also contains, on average, more polymorphic loci per window in the last third of the "genome" than the first 20kb (see figure S5). We only see hints of this in the actual MERS-CoV genome, with an apparent decline in polymorphism density from position 5000 to 15000 which resembles that of the simulated data with 1.3-fold rate heterogeneity.

Homoplasies in MERS-CoV genomes are ubiquitous

Homoplasy analyses suggests that the MERS-CoV genome is rife with apparent homoplasies. Both maximum likelihood and Bayesian approaches to ancestral sequence reconstruction converge on similar patterns of homoplasy density (figures 3 and S6). Both methods identify the region around the S (spike) protein as having a high density of synonymous homoplasies.

Empirically simulated sequences showed that homoplasies are not that unlikely in the absence of recombination. All sequences empirically simulated in πBUSS had 2-fold homoplasies ranging in frequency from 0.0222 to 0.0550 of all polymorphic sites, with sequences simulated under higher levels of rate heterogeneity having more homoplasies and higher homoplasy degrees (figure 4). However, even under a caricature model of rate heterogeneity we did not reach the same degree of homoplasy as that observed in MERS-CoV, where homoplasic sites comprise as much as 0.1447 of all polymorphic sites and reach homoplasy degrees as high as 4.

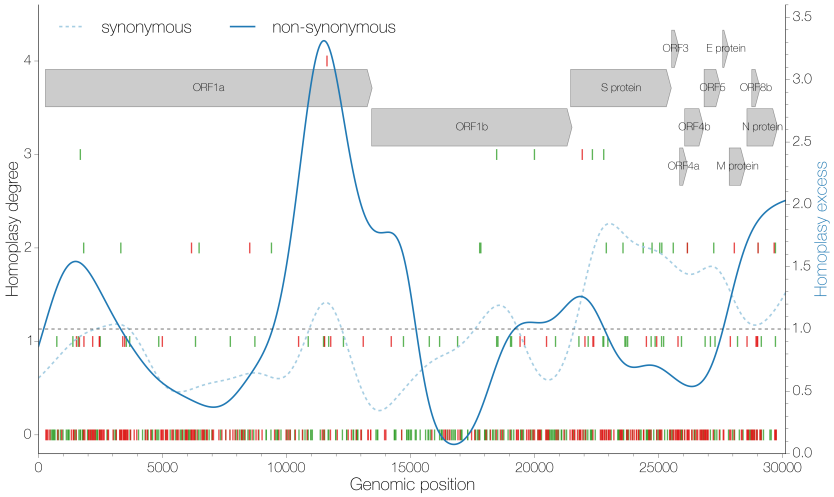

Figure 3. Distribution of apparent homoplasies. Position along the genome is shown on the x axis and homoplasy degree, the number of times a particular mutation has occured in excess in the tree as inferred by maximum likelihood, is shown on the y axis (left). Individual mutations are marked by vertical lines, synonymous ones in green and non-synonymous in red. 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.

Model testing supports a model including rate heterogeneity, but not alternative tree topologies

A model including rate heterogeneity alone across breakpoints inferred by the GARD method (*i.e.* linked trees, unlinked relaxed clocks) performs best when applied to MERS-CoV data (figure 5 log marginal likelihoods: -48137.86 and -48138.91, using path and stepping stone sampling, respectively). The next best-performing model (log Bayes factor ≈18) is linked trees and relaxed clocks. Overall, unlinking molecular clock rates between the two genomic partitions and using relaxed clocks appear sufficient to dramatically improve model fit.

Discussion

Recombination tests consistently point to recombination in MERS-CoV

The majority of methods we used (with the exception of marginal likelihood model testing) point to the combined effects of recombination and rate heterogeneity in the genome of MERS coronavirus. GARD (figure 1) identified 2 breakpoints in the genome with high support. Estimating the rate heterogeneity across the first breakpoint in BEAST gave an empirical rate heterogeneity ratio between MERS-CoV genome positions 23723-

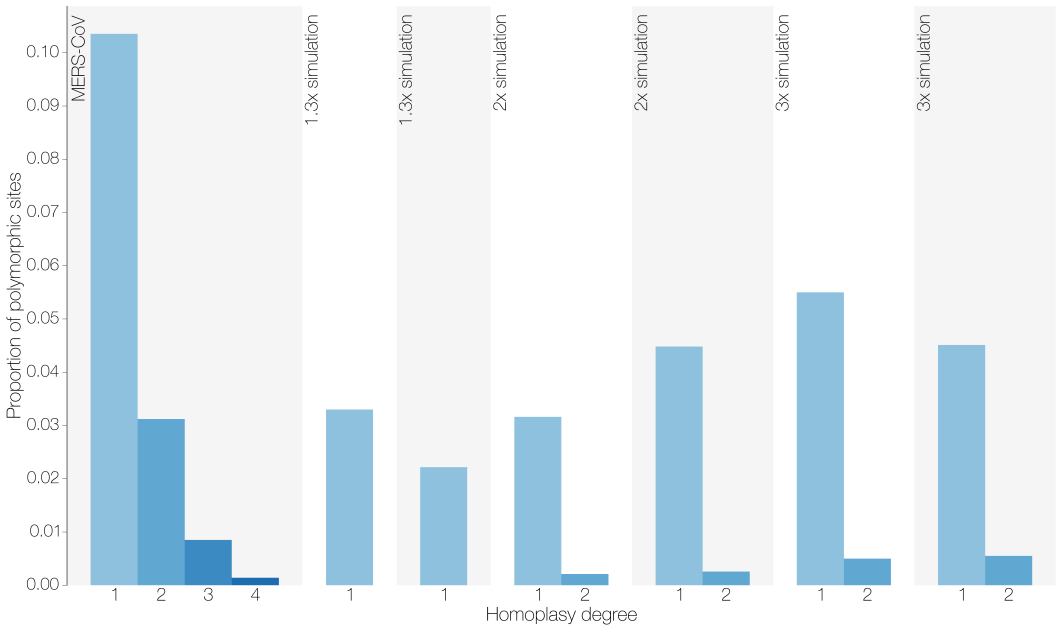

Figure 4. Homoplasy prevalence in MERS-CoV and simulated datasets. Bars show the proportion of all polymorphic sites that are homoplasic, split by homoplasy degree as inferred by maximum likelihood, in MERS-CoV and datasets simulated with different degrees of rate heterogeneity. Homoplasy degree indicates how many times a given mutation has occured in excess in the phylogenetic tree.

30126 and 1-23722 to be somewhere between 1.3 and 1.5 (see figure S1). However, the support for this first breakpoint in MERS-CoV is comparable to support for empircally simulated sequences with 2-fold rate heterogeneity. In the latter case high support reflected changes in likelihood induced by contrast in branch lengths between genomic partitions, not topology. This would suggest that evidence for differences in tree topology between MERS fragments 1 and 2 is beyond what would be expected from rate heterogeneity alone.

Permutation tests show that statistics related to LD decay derived from MERS-CoV sequence data are outliers compared to permuted data (figures 2 and S2). Sequences simulated empirically and with varying levels of rate heterogeneity also have a tendency to exhibit extreme values for these statistics. However, only MERS-CoV data has values for all 4 tests that are in the direction consistent with recombination. This implies that permutation tests are to some extent robust to rate heterogeneity, but only when combined together.

Homoplasy analyses regardless of inference method show that MERS-CoV sequences contain a large number of homoplasic sites with high homoplasy degrees (figures 3 and S6). Through sequence simulation we also confirmed that both the numbers of homoplasic sites and their homoplasy degrees in MERS-CoV genomes are excessive, even when compared to unrealistic scenarios (*e.g.* 3-fold rate heterogeneity, see figure 4). 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).

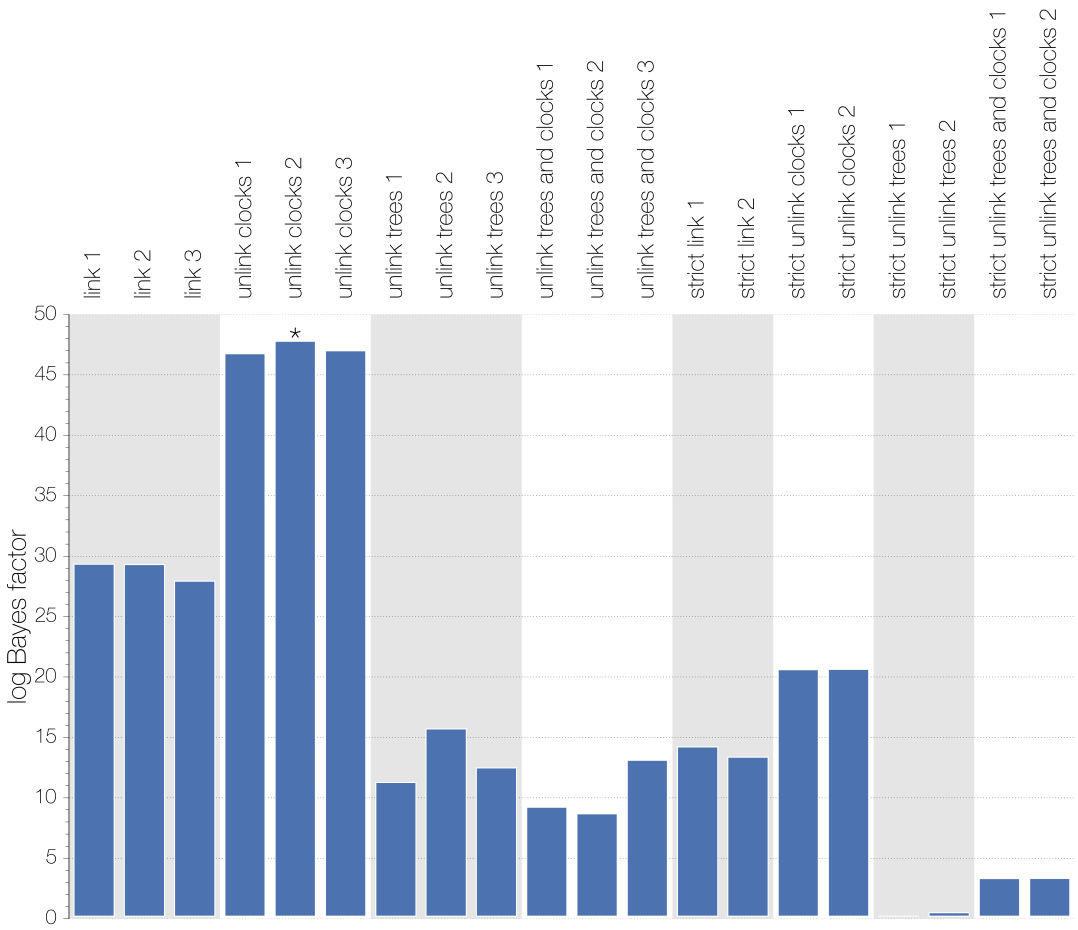

Figure 5. Summary of model comparisons. Difference in marginal likelihoods (Bayes factor) estimated by path-sampling between the worst model (linked strict molecular clock, unlinked trees) and all others. Asterisks indicate the best-performing model (unlinked relaxed clocks, linked trees, run 2) for MERS-CoV data. Analyses employing a relaxed molecular clock were run independently 3 times, those with a strict molecular clock 2 times. Marginal likelihoods estimated using stepping stone sampling gave identical results.

This is important, since homoplasies inferred using BEAST are integrated over all possible tree topologies, whereas homoplasies inferred by maximum likelihood were conditioned on a single tree. The convergence between these two methods suggests that the data contain enough phylogenetic signal to recover what could be called a "true" tree and that homoplasies, for the most part, can be correctly identified as such.

One major concern surrounding the inference of homoplasies is host adaptation. There are a number of canonical mutations associated with host shifts, *e.g.* the glutamic acid to lysine amino acid substitution at position 627 in the PB2 protein of avian influenza A viruses confers the ability of the virus to replicate in mammals (Subbarao et al., 1993) and a small number of amino acid substitutions in Parvoviruses are associated with adaptations to different hosts (Chang et al., 1992). If MERS-CoV is repeatedly emerging in humans convergent mutations would be expected to arise that might allow the virus to adapt to

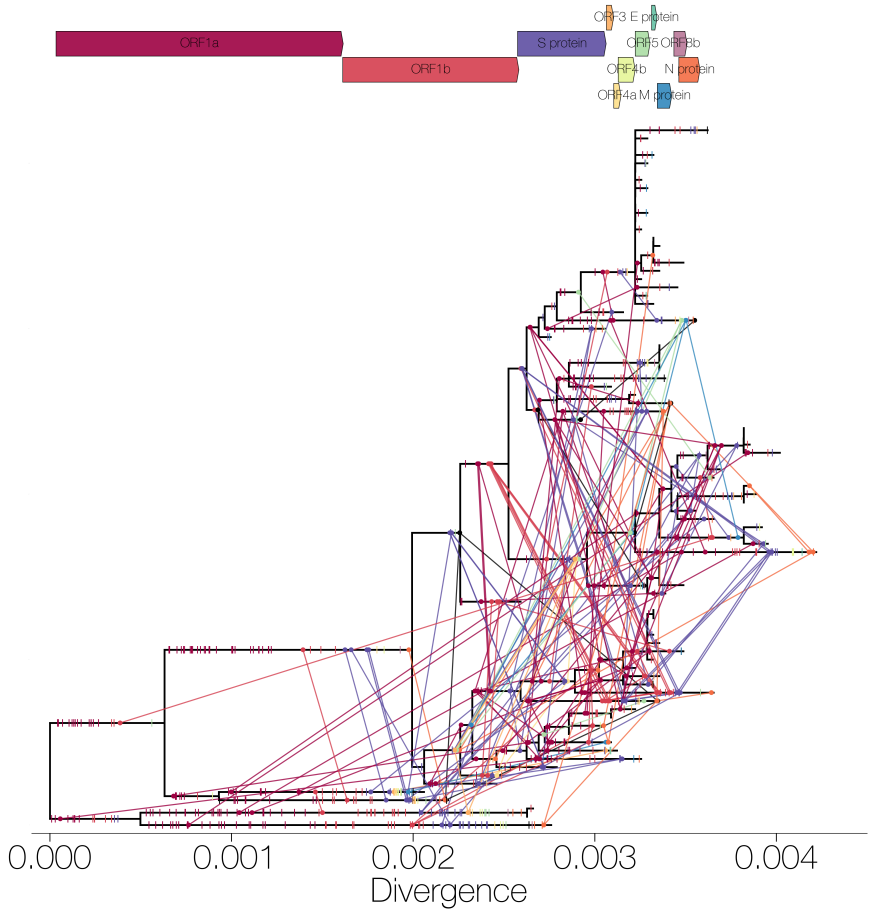

Figure 6. Mutations mapped onto a ML phylogeny. A maximum likelihood phylogeny of 85 MERS-CoV sequences with maximum likelihood-mapped mutations. Synapomorphies are shown as coloured ticks (coloured by coding sequence in which they occur) on branches where they occur. Homoplasies are shown as circles connected with coloured lines, colour corresponds with the coding sequence in which the mutation has occured. Mutations are positioned on the branches in proportion to where the mutation occurs in the genome, *e.g.* mutations shown towards the end of a branch correspond to mutations near the 3' terminus of the genome. Arrows at the top indicate the order, relative length and names of coding sequences within the MERS-CoV genome.

humans.

However, we expect most host-adaptation mutations to be non-synonymous, whereas we detect both non-synonymous and synonymous homoplasies. This implies the action of recombination, rather than repeated selection for the same host-specific mutations. Furthermore, we do not detect any strong associations between host and particular alleles (figure S7).

The overall phylogenetic and genomic patterns of homoplasies are consistent with fairly frequent recombination through time (figure 6). Recent recombination should result in long homoplasy tracts shared across branches in the phylogenetic tree. At most we observe 2 stretches of adjacent homoplasies, one encompassing 3 homoplasies and another encompassing 2 homoplasies, that are shared between taxa, and likely to be caused by recent recombination. The vast majority of homoplasies that we observe, however, occur on their own. Recombination tracts, rather than single template switches are not uncommon in other coronaviruses (Keck et al., 1988; Kottier et al., 1995; Herrewegh et al., 1998). Thus in MERS-CoV we interpret extremely short homoplasy tracts as evidence of relatively frequent recombination.

Unlike all other tests we performed model testing through marginal likelihoods indicates that models including rate heterogeneity explain MERS-CoV data partitioned across a well-supported breakpoint better than models including independent trees. At first this may seem paradoxical, but we believe this result is due to the combined effects of the way homoplasious sites are distributed across the genome and phylogenetic tree of MERS-CoV (figure 6) and the number of parameters involved. A speckled pattern of homoplasious sites without phylogenetic signal could easily be overwhelmed by the signal coming from the sites that support what could be called "the one true tree", *i.e.* the clonal frame, in the data. Secondly, each phylogenetic tree contains at least n-1 free parameters, so it is not surprising then that models attempting to recover 2 independent trees for both genomic fragments resulting from alternative tree topology analysis of MERS-CoV with highly correlated genealogies are penalized for the extra parameters introduced by a second tree.

Implications for future analyses

Recombination aside, MERS-CoV genomes exhibit a significant degree of rate heterogeneity amongst sites. Marginal likelihood analyses indicate that estimating independent molecular clocks after partitioning the MERS-CoV genome into 2 fragments alone substantially increases model fit over a completely linked (trees and clocks) model (log Bayes factor $\approx$18). This highlights the advantage of employing relaxed molecular clocks, as in our case the method is clearly capable of accomodating recombination in an otherwise entirely clonal analysis framework. In addition, previous studies of SARS-like coronaviruses in bats have identified recombination breakpoints in small numbers of isolates falling close to the "transition zone" around site 22000 (Hon et al., 2008; Lau et al., 2010) which in our analysis of MERS-CoV is where GARD, LDhat and BEAST identify changes in the underlying model of evolution (figures 1, S4 and 3).

We also expect that as more sequences of MERS-CoV become available more homoplasies will be detected, some contributing to the homoplasy degree of the homoplasies already reported here, some previously unknown and some turning mutations currently thought of as synapomorphies into homoplasies. Although new sequences are likely to come from human cases, we think that sequencing MERS-CoV circulating in dromedary camels is of extreme importance from both surveillance and epidemiological points of view.

Implications about the virus population structure and infection dynamics

Our results point towards frequent recombination in MERS-CoV in the recent history of the MERS-CoV outbreak. For this to occur different lineages of the virus must encounter each other often and implies frequent co-infection with MERS-CoV. To date it is difficult to ascertain whether the human infections with MERS-CoV are a result of substantial asymptomatic transmission amongst humans, or repeated zoonosis of the virus from camels to humans or a combination thereof. Given the severity of MERS we find it unlikely that humans could be sufficiently frequently co-infected with two or more different lineages of the virus. Previous serological studies have failed to find evidence of prevalent past MERS-CoV infections of humans (Gierer et al., 2013; Aburizaiza et al., 2013), although a recent nation-wide study in Saudi Arabia has detected non-negligible numbers of individuals with antibodies against MERS-CoV, especially amongst shepherds and slaughterhouse workers (Müller et al., 2015). We thus propose that MERS-CoV mostly infects, and recombines, in camels. 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 (Müller et al., 2014; Corman et al., 2014b; Chu et al., 2014; Reusken et al., 2013, 2014), which creates nearly ideal conditions for the virus to recombine.

Another point worth considering is that alleles that have arisen through mutation in MERS-CoV can be recombined, increasing the genetic variation of the virus (Muller, 1932). Whether this is of epidemiological importance for humans depends entirely on what alleles are circulating in the reservoir, although there is no evidence that MERS-CoV is particularly likely to become as transmissible as common human pathogens or even SARS-CoV.

Data availability

Python scripts used to process trees and sequences are available at:
https://github.com/evogytis/MERS_recombination/tree/master/scripts.

Input and output files for programs used are publicly available at:
https://github.com/evogytis/MERS_recombination.

Acknowledgements

GD was supported by a Natural Environment Research Council studentship D76739X. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864.

References

Aburizaiza AS, Mattes FM, Azhar EI, Hassan AM, Memish ZA, Muth D, Meyer B, Lattwein E, Müller M, Drosten C. 2013. Investigation of anti-MERS-coronavirus antibodies in blood donors and abbatoir workers in jeddah and makkah, kingdom of saudi arabia, fall 2012. Journal of Infectious Diseases. p. jit589.

Adney DR, van Doremalen N, Brown VR, Bushmaker T, Scott D, de Wit E, Bowen RA, Munster VJ. 2014. Replication and shedding of MERS-CoV in upper respiratory tract of inoculated dromedary camels. Emerging Infectious Diseases. 20:1999–2005.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics. 164:1229–1236.

Assiri A, Al-Tawfiq JA, Al-Rabeeah AA, et al. (13 co-authors). 2013. Epidemiological, demographic, and clinical characteristics of 47 cases of middle east respiratory syndrome coronavirus disease from saudi arabia: a descriptive study. The Lancet Infectious Diseases. 13:752–761.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution. 29:2157–2167.

Baric RS, Shieh CK, Stohlman SA, Lai MMC. 1987. Analysis of intracellular small RNAs of mouse hepatitis virus: evidence for discontinuous transcription. Virology. 156:342–354.

Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A, Suchard MA. 2014. πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. BMC Bioinformatics. 15:133.

Boni MF, Zhou Y, Taubenberger JK, Holmes EC. 2008. Homologous recombination is very rare or absent in human influenza a virus. Journal of Virology. 82:4807–4811.

Boucher CA, Cammack N, Schipper P, Schuurman R, Rouse P, Wainberg MA, Cameron JM. 1993. High-level resistance to (-) enantiomeric 2'-deoxy-3'-thiacytidine in vitro is due to one amino acid substitution in the catalytic site of human immunodeficiency virus type 1 reverse transcriptase. Antimicrobial Agents and Chemotherapy. 37:2231–2234.

Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enouf V, van der Werf S, Ferguson NM. 2014. Middle east respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. The Lancet Infectious Diseases. 14:50–56.

Chang SF, Sgro JY, Parrish CR. 1992. Multiple amino acids in the capsid structure of canine parvovirus coordinately determine the canine host range and specific antigenic and hemagglutination properties. Journal of Virology. 66:6858–6867.

Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of

homologous recombination in negative-sense RNA viruses. Journal of General Virology. 84:2691–2703.

Chu DK, Poon LL, Gomaa MM, et al. (13 co-authors). 2014. MERS coronaviruses in dromedary camels, egypt. Emerging Infectious Diseases. 20:1049–1053.

Corman VM, Ithete NL, Richards LR, Schoeman MC, Preiser W, Drosten C, Drexler JF. 2014a. Rooting the phylogenetic tree of middle east respiratory syndrome coronavirus by characterization of a conspecific virus from an african bat. Journal of Virology. 88:11297–11303.

Corman VM, Jores J, Meyer B, et al. (13 co-authors). 2014b. Antibodies against MERS coronavirus in dromedary camels, kenya, 1992–2013. Emerging Infectious Diseases. 20.

Cotten M, Watson SJ, Kellam P, et al. (22 co-authors). 2013. Transmission and evolution of the middle east respiratory syndrome coronavirus in saudi arabia: a descriptive genomic study. The Lancet. 382:1993–2002.

Cotten M, Watson SJ, Zumla AI, et al. (20 co-authors). 2014. Spread, circulation, and evolution of the middle east respiratory syndrome coronavirus. mBio. 5:e01062–13.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence data. Genetics. 175:1251–1266.

Dong G, Luo J, Zhang H, Wang C, Duan M, Deliberto TJ, Nolte DL, Ji G, He H. 2011. Phylogenetic diversity and genotypical complexity of h9n2 influenza a viruses revealed by genomic sequence analysis. PLoS ONE. 6:e17212.

Drosten C, Muth D, Corman V, et al. (15 co-authors). 2014. An observational, laboratory-based study of outbreaks of MERS-coronavirus in jeddah and riyadh, kingdom of saudi arabia, 2014. Clinical Infectious Diseases. p. ciu812.

Drosten C, Seilmaier M, Corman VM, et al. (22 co-authors). 2013. Clinical features and virological analysis of a case of middle east respiratory syndrome coronavirus infection. The Lancet Infectious Diseases. 13:745–751.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution. 29.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9:e1003905.

Ferreira MAR, Suchard MA. 2008. Bayesian analysis of elapsed times in continuous-time markov chains. Canadian Journal of Statistics. 36:355–368.

Gierer S, Hofmann-Winkler H, Albuali WH, et al. (11 co-authors). 2013. Lack of MERS coronavirus neutralizing antibodies in humans, eastern province, saudi arabia. Emerging Infectious Diseases. 19:2034–2036.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. Molecular Biology and Evolution. 30:713–724.

Gire SK, Goba A, Andersen KG, et al. (58 co-authors). 2014. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. Science. 345:1369–1372.

Gubareva LV, Kaiser L, Matrosovich MN, Soo-Hoo Y, Hayden FG. 2001. Selection of influenza virus mutants in experimentally infected volunteers treated with oseltamivir. Journal of Infectious Diseases. 183:523–531.

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

Hasegawa M, Kishino H, Yano Ta. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution. 22:160–174.

Herrewegh AAPM, Smeenk I, Horzinek MC, Rottier PJM, Groot RJd. 1998. Feline coronavirus type II strains 79-1683 and 79-1146 originate from a double recombination between feline coronavirus type i and canine coronavirus. Journal of Virology. 72:4508–4514.

Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik. 38:226–231.

Holmes EC, Urwin R, Maiden MC. 1999. The influence of recombination on the population structure and evolution of the human pathogen neisseria meningitidis. Molecular Biology and Evolution. 16:741–749.

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FCC. 2008. Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory Syndrome (SARS)-Like Coronavirus and Its Implications on the Direct Ancestor of SARS Coronavirus. Journal of Virology. 82:1819–1826.

Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM. 1988. In vivo RNA-RNA recombination of coronavirus in mouse brain. Journal of Virology. 62:1810–1813.

Kirkegaard K, Baltimore D. 1986. The mechanism of RNA recombination in poliovirus. Cell. 47:433–443.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Molecular Biology and Evolution. 23:1891–1901.

Kottier SA, Cavanagh D, Britton P. 1995. Experimental evidence of recombination in coronavirus infectious bronchitis virus. Virology. 213:569–580.

Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA. 1985. Recombination between nonsegmented RNA genomes of murine coronaviruses. Journal of Virology. 56:449–456.

Lau SKP, Li KSM, Huang Y, et al. (14 co-authors). 2010. Ecoepidemiology and Complete Genome Comparison of Different Strains of Severe Acute Respiratory Syndrome-Related Rhinolophus Bat Coronavirus in China Reveal Bats as a Reservoir for Acute, Self-Limiting Infection That Allows Recombination Events. Journal of Virology. 84:2808–2819.

Lemey P, Suchard M, Rambaut A. 2009. Reconstructing the initial global spread of a human influenza pandemic. PLoS Currents. 1.

Lewontin RC. 1964. The interaction of selection and linkage. i. general considerations; heterotic models. Genetics. 49:49–67. PMID: 17248194 PMCID: PMC1210557.

Li KS, Guan Y, Wang J, et al. (22 co-authors). 2004. Genesis of a highly pathogenic and potentially pandemic h5n1 influenza virus in eastern asia. Nature. 430:209–213.

Lu L, Lycett SJ, Brown AJL. 2014. Reassortment patterns of avian influenza virus internal segments among different subtypes. BMC Evolutionary Biology. 14:16.

Makino S, Keck JG, Stohlman SA, Lai MM. 1986. High-frequency RNA recombination of murine coronaviruses. Journal of Virology. 57:729–737.

Maynard Smith J, Smith NH. 1998. Detecting recombination from gene trees. Molecular Biology and Evolution. 15:590–599.

McVean G, Awadalla P, Fearnhead P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics. 160:1231–1241.

Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. 2013. Family cluster of middle east respiratory syndrome coronavirus infections. New England Journal of Medicine. 368:2487–2494.

Meunier J, Eyre-Walker A. 2001. The correlation between linkage disequilibrium and distance: Implications for recombination in hominid mitochondria. Molecular Biology and Evolution. 18:2132–2135.

Milkman R, Bridges MM. 1990. Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. Genetics. 126:505–517.

Mitnaul LJ, Matrosovich MN, Castrucci MR, Tuzikov AB, Bovin NV, Kobasa D, Kawaoka Y. 2000. Balanced hemagglutinin and neuraminidase activities are critical for efficient replication of influenza a virus. Journal of Virology. 74:6015–6020.

Muller HJ. 1932. Some genetic aspects of sex. The American Naturalist. 66:118–138.

Müller MA, Corman VM, Jores J, et al. (12 co-authors). 2014. MERS coronavirus neutralizing antibodies in camels, eastern africa, 1983–1997. Emerging Infectious Diseases. 20.

Müller MA, Meyer B, Corman VM, et al. (19 co-authors). 2015. Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a nationwide, cross-sectional, serological study. The Lancet Infectious Diseases. 15:559–564.

O'Brien JD, Minin VN, Suchard MA. 2009. Learning to count: Robust estimates for labeled distances between molecular sequences. Molecular Biology and Evolution. 26:801–814. PMID: 19131426.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 21:676–679.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution. 54:396–402.

Rambaut A, Holmes E. 2009. The early molecular epidemiology of the swine-origin a/h1n1 human influenza pandemic. PLoS Currents. 1.

Reusken CB, Haagmans BL, Müller MA, et al. (24 co-authors). 2013. Middle east respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. The Lancet Infectious Diseases. 13:859–866.

Reusken CB, Messadi L, Feyisa A, et al. (17 co-authors). 2014. Geographic distribution of MERS coronavirus among dromedary camels, africa. Emerging Infectious Diseases. 20:1370–1374.

Robertson DL, Hahn BH, Sharp PM. 1995a. Recombination in AIDS viruses. Journal of Molecular Evolution. 40:249–259.

Robertson DL, Sharp PM, McCutchan FE, Hahn BH. 1995b. Recombination in HIV-1. Nature. 374:124–126.

Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics. 156:879–891.

Shriner D, Nickle DC, Jensen MA, Mullins JI. 2003. Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genetics Research. 81:115–121.

Smith GJD, Vijaykrishna D, Bahl J, et al. (13 co-authors). 2009. Origins and evolutionary genomics of the 2009 swine-origin h1n1 influenza a epidemic. Nature. 459:1122–1125.

Subbarao EK, London W, Murphy BR. 1993. A single amino acid in the PB2 gene of influenza a virus is a determinant of host range. Journal of Virology. 67:1761–1764.

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.

Tisdale M, Kemp SD, Parry NR, Larder BA. 1993. Rapid in vitro selection of human immunodeficiency virus type 1 resistant to 3'-thiacytidine inhibitors due to a mutation in the YMDD region of reverse transcriptase. Proceedings of the National Academy of Sciences of the United States of America. 90:5653–5656.

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

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in saudi arabia. New England Journal of Medicine. 367:1814–1820.

Zhao H, Nettleton D, Soller M, Dekkers JCM. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetics Research. 86:77–87.