MERS-CoV spillover at the camel-human interface

Gytis Dudas[1*], **Luiz Max Carvalho**[2], **Andrew Rambaut**[2,3] **& Trevor Bedford**[1]

***For correspondence:** gdudas@fredhutch.org (GD)

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [2]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, UK, [3]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA

Abstract

Middle East respiratory syndrome coronavirus (MERS-CoV) is a zoonotic virus from camels causing significant mortality and morbidity in humans in the Arabian Peninsula. The epidemiology of the virus remains poorly understood, and while case-based and seroepidemiological studies have been employed extensively throughout the epidemic, viral sequence data have not been utilised to their full potential. Here we use existing MERS-CoV sequence data to explore its phylodynamics in two of its known major hosts, humans and camels. We employ structured coalescent models to show that long-term MERS-CoV evolution occurs exclusively in camels, whereas humans act as a transient, and ultimately terminal host. By analysing the distribution of human outbreak cluster sizes and zoonotic introduction times we show that human outbreaks in the Arabian peninsula are driven by seasonally varying zoonotic transfer of viruses from camels. Without heretofore unseen evolution of host tropism, MERS-CoV is unlikely to become endemic in humans.

Introduction

Middle East respiratory syndrome coronavirus (MERS-CoV), endemic in camels in the Arabian Peninsula, is the causative agent of zoonotic infections and limited outbreaks in humans. The virus, first discovered in 2012 (***Zaki et al., 2012***; ***Boheemen et al., 2012***), has caused more than 2000 infections and over 700 deaths, according to the World Health Organization (WHO) (***World Health Organization, 2017***). Its epidemiology remains obscure, largely because infections are observed among the most severely affected individuals, such as older males with comorbidities (***Assiri et al., 2013a***; ***The WHO MERS-CoV Research Group, 2013***). While contact with camels is often reported, other patients do not recall contact with any livestock, suggesting an unobserved community contribution to the outbreak (***The WHO MERS-CoV Research Group, 2013***). Previous studies on MERS-CoV epidemiology have used serology to identify factors associated with MERS-CoV exposure in potential risk groups (***Reusken et al., 2015***, ***2013***). Such data have shown high seroprevalence in camels (***Müller et al., 2014***; ***Corman et al., 2014***; ***Chu et al., 2014***; ***Reusken et al., 2013***, ***2014***) and evidence of contact with MERS-CoV in workers with occupational exposure to camels (***Reusken et al., 2015***; ***Müller et al., 2015***). Separately, epidemiological modelling approaches have been used to look at incidence reports through time, space and across hosts (***Cauchemez et al., 2016***).

Although such epidemiological approaches yield important clues about exposure patterns and potential for larger outbreaks, much inevitably remains opaque to such approaches due to difficulties in linking cases into transmission clusters in the absence of detailed information. Where sequence data are relatively cheap to produce, genomic epidemiological approaches can fill this critical gap in outbreak scenarios (***Liu et al., 2013***; ***Gire et al., 2014***; ***Grubaugh et al., 2017***). These data often yield a highly detailed picture of an epidemic when complete genome sequencing is performed consistently and appropriate metadata collected (***Dudas et al., 2017***). Sequence data can help discriminate between multiple and single source scenarios (***Gire et al., 2014***; ***Quick et al., 2015***), which are fundamental to quantifying risk (***Grubaugh et al., 2017***). Sequencing MERS-CoV has been performed as part of initial attempts to link human infections with the camel reservoir (***Memish et al., 2014***), nosocomial outbreak investigations (***Assiri et al., 2013b***) and routine surveillance (***Park et al., 2015***). A large portion of MERS-CoV sequences come from outbreaks within hospitals, where sequence data have been used to determine whether infections were isolated introductions or were part of a larger hospital-associated outbreak (***Fagbo et al., 2015***). Similar studies on MERS-CoV have taken place at broader geographic scales, such as cities (***Cotten et al., 2013***).

It is widely accepted that recorded human MERS-CoV infections are a result of at least several introductions of the virus into humans (***Cotten et al., 2013***) and that contact with camels is a major risk factor for developing MERS, per WHO guidelines (***World Health Organization, 2016***). Previous studies attempting to quantify the actual number of spillover infections have either relied on case-based epidemiological approaches (***Cauchemez et al., 2016***) or employed methods agnostic to signals of population structure within sequence data (***Zhang et al., 2016***). Here we use a dataset of 274 MERS-CoV genomes to investigate transmission patterns of the virus between humans and camels.

Here, we use an explicit model of metapopulation structure and migration between discrete subpopulations, referred to here as demes (***Vaughan et al., 2014***), derived from the structured coalescent (***Notohara, 1990***). Unlike approaches that model host species as a discrete phylogenetic trait of the virus using continuous-time Markov processes (or simpler, parsimony based, approaches) (***Faria et al., 2013***; ***Lycett et al., 2016***), population structure models explicitly incorporate contrasts in deme effective population sizes and migration between demes. By estimating independent coalescence rates for MERS-CoV in humans and camels, as well as migration patterns between the two demes, we show that long-term viral evolution of MERS-CoV occurs exclusively in camels. Our results suggest that spillover events into humans are seasonal and might be associated with the calving season in camels. However, we find that MERS-CoV, once introduced into humans, follows

transient transmission chains that soon abate. Using Monte Carlo simulations we show that R_0 for MERS-CoV circulating in humans is much lower than the epidemic threshold of 1.0 and that correspondingly the virus has been introduced into humans hundreds of times.

Results

MERS-CoV is predominantly a camel virus

The structured coalescent approach we employ (see Methods) identifies camels as a reservoir host where most of MERS-CoV evolution takes place (Figure 1), while human MERS outbreaks are transient and terminal with respect to long-term evolution of the virus (Figure 1-Figure supplement 1). Across 174 MERS-CoV genomes collected from humans, we estimate a median of 56 separate camel-to-human cross-species transmissions (95% highest posterior density (HPD): 48–63). While we estimate a median of 3 (95% HPD: 0–12) human-to-camel migrations, the 95% HPD interval includes zero and we find that no such migrations are found in the maximum clade credibility tree (Figure 1). Correspondingly, we observe substantially higher camel-to-human migration rate estimates than human-to-camel migration rate estimates (Figure 1-Figure supplement 2). This inference derives from the tree structure wherein human viruses appear as clusters of highly related sequences nested within the diversity seen in camel viruses, which themselves show significantly higher diversity and less clustering. This manifests as different rates of coalescence with camel viruses showing a scaled effective population size $N_e\tau$ of 3.49 years (95% HPD: 2.71–4.38) and human viruses showing a scaled effective population of 0.24 years (95% HPD: 0.14–0.34).

We believe that the small number of inferred human-to-camel migrations are induced by the migration rate prior, while some are derived from phylogenetic proximity of human sequences to the apparent "backbone" of the phylogenetic tree. This is most apparent in lineages in early-mid 2013 that lead up to sequences comprising the MERS-CoV clade dominant in 2015, where owing to poor sampling of MERS-CoV genetic diversity from camels the model cannot completely dismiss humans as a potential alternative host. The first sequences of MERS-CoV from camels do not appear in our data until November 2013. Our finding of negligible human-to-camel transmission is robust to choice of prior (Figure 1-Figure supplement 2).

The repeated and asymmetric introductions of short-lived clusters of MERS-CoV sequences from camels into humans leads us to conclude that MERS-CoV epidemiology in humans is dominated by zoonotic transmission (Figure 1 and 1-Figure supplement 1). We observe dense terminal clusters of MERS-CoV circulating in humans that are of no subsequent relevance to the evolution of the virus. These clusters of presumed human-to-human transmission are then embedded within extensive diversity of MERS-CoV lineages inferred to be circulating in camels, a classic pattern of source-sink dynamics. Our findings suggest that instances of human infection with MERS-CoV are more common than currently thought, with exceedingly short transmission chains mostly limited to primary cases that might be mild and ultimately not detected by surveillance or sequencing. Structured coalescent analyses recover the camel-centered picture of MERS-CoV evolution despite sequence data heavily skewed towards non-uniformly sampled human cases and are robust to choice of prior. Comparing these results with a currently standard discrete trait analysis (***Lemey et al., 2009***) approach for ancestral state reconstruction shows dramatic differences in host reconstruction at internal nodes (Figure 1-Figure supplement 3). Discrete trait analysis reconstruction identifies both camels and humans as important hosts for MERS-CoV persistence, but with humans as the ultimate source of camel infections. A similar approach has been attempted previously (***Zhang et al., 2016***), but this interpretation of MERS-CoV evolution disagrees with lack of continuing human transmission chains outside of Arabian peninsula, low seroprevalence in humans and very high seroprevalence in camels across Saudi Arabia. We suspect that this particular discrete trait analysis reconstruction is false due to biased data, *i.e.* having nearly twice as many MERS-CoV sequences from humans ($n = 174$)

than from camels ($n = 100$) and the inability of the model to account for and quantify vastly different rates of coalescence in the phylogenetic vicinity of both types of sequences. We can replicate these results by either applying the same model to current data (Figure 1-Figure supplement 3) or by enforcing equal coalescence rates within each deme in the structured coalescent model (Figure 1-Figure supplement 4).

MERS-CoV shows seasonal introductions

We use the posterior distribution of MERS-CoV introduction events from camels to humans (Figure 1) to model seasonal variation in zoonotic transfer of viruses. We identify four months (April, May, June, July) when the odds of MERS-CoV introductions are increased (Figure 2) and four when the odds are decreased (August, September, November, December). Camel calving is reported to occur from October to February (***Almutairi et al., 2010***), with rapidly declining maternal antibody levels in calves within the first weeks after birth (***Wernery, 2001***). It is possible that MERS-CoV sweeps through each new camel generation once critical mass of susceptibles is reached (***Martinez-Bakker et al., 2014***), leading to a sharp rise in prevalence of the virus in camels and resulting in increased force of infection into the human population. Strong influx of susceptibles and subsequent sweeping outbreaks in camels may explain evidence of widespread exposure to MERS-CoV in camels from seroepidemiology (***Müller et al., 2014***; ***Corman et al., 2014***; ***Chu et al., 2014***; ***Reusken et al., 2013***, ***2014***).

Although we find evidence of seasonality in zoonotic spillover timing, no such relationship exists for sizes of human sequence clusters (Figure 2B). This is entirely expected, since little seasonality in human behaviour that could facilitate MERS-CoV transmission is expected following an introduction. Similarly, we do not observe any trend in human sequence cluster sizes over time (Figure 2B, Spearman $\rho = 0.06$, $p = 0.68$), suggesting that MERS-CoV outbreaks in humans are neither growing nor shrinking in size. This is not surprising either, since MERS-CoV is a camel virus that has to date, experienced little-to-no selective pressure to improve transmissibility between humans.

MERS-CoV is poorly suited for human transmission

Structured coalescent approaches clearly show humans to be a terminal host for MERS-CoV, implying poor transmissibility. However, we wanted to translate this observation into an estimate of the basic reproductive number R_0 to provide an intuitive epidemic behaviour metric that does not require expertise in reading phylogenies. The parameter R_0 determines expected number of secondary cases in a single infections as well as the distribution of total cases that result from a single introduction event into the human population (Equation 1, Methods). We estimate R_0 along with other relevant parameters via Monte Carlo simulation in two steps. First, we simulate case counts across multiple outbreaks totaling 2000 cases using Equation 1 and then we subsample from each case cluster to simulate sequencing of a fraction of cases. Sequencing simulations are performed via a multivariate hypergeometric distribution, where the probability of sequencing from a particular cluster depends on available sequencing capacity (number of trials), numbers of cases in the cluster (number of successes) and number of cases outside the cluster (number of failures). In addition, each hypergeometric distribution used to simulate sequencing is concentrated via a bias parameter, that enriches for large sequence clusters at the expense of smaller ones (for its effects see Figure 3-Figure supplement 1). This is a particularly pressing issue, since *a priori* we expect large hospital outbreaks of MERS to be overrepresented in sequence data, whereas sequences from primary cases will be sampled exceedingly rarely. We record the number, mean, standard deviation and skewness (third standardised moment) of sequence cluster sizes in each simulation (left-hand panel in Figure 3) and extract the subset of Monte Carlo simulations in which these summary statistics fall within the 95% highest posterior density observed in the empirical MERS-CoV data from structured coalescent analyses. We record R_0 values, as well as the number

of case clusters (equivalent to number of zoonotic introductions), for these empirically matched simulations. A schematic of this Monte Carlo procedure is shown in Figure 3-Figure supplement 2. Since the total number of cases is fixed at 2000, higher R_0 results in fewer larger transmission clusters, while lower R_0 results in many smaller transmission clusters.

We find that observed phylogenetic patterns of sequence clustering strongly support R_0 below 1.0 (middle panel in Figure 3). Mean R_0 value observed in matching simulations is 0.72 (95% percentiles 0.57–0.90), suggesting the inability of the virus to sustain transmission in humans. Lower values for R_0 in turn suggest relatively large numbers of zoonotic transfers of viruses into humans (right-hand panel in Figure 3). Median number of cross-species introductions observed in matching simulations is 592 (95% percentiles 311-811). Our results suggest a large number of unobserved MERS primary cases. Given this, we also performed simulations where the total number of cases is doubled to 4000 to explore the impact of dramatic underestimation of MERS cases. In these analyses R_0 values tend to decrease even further as numbers of introductions go up, although very few simulations match currently observed MERS-CoV sequence data (Figure 3-Figure supplement 3).

Overall, our analyses indicate that MERS-CoV is poorly suited for human-to-human transmission, with an estimated $R_0 < 0.90$ and sequence sampling likely to be biased towards large hospital outbreaks (Figure 3-Figure supplement 1). All matching simulations exhibit highly skewed distributions of case cluster sizes with long tails (Figure 3-Figure supplement 4), which is qualitatively similar to the results of (***Cauchemez et al., 2016***). We find that simulated sequence cluster sizes resemble observed sequence cluster sizes in the posterior distribution, with a slight reduction in mid-sized clusters in simulated data (Figure 3-Figure supplement 5). Given these findings, and the fact that large outbreaks of MERS occurred in hospitals, the combination of frequent spillover of MERS-CoV into humans and occasional outbreak amplification via poor hygiene practices (***Assiri et al., 2013b***; ***Chen et al., 2017***) appear sufficient to explain observed epidemiological patterns of MERS-CoV.

Recombination shapes MERS-CoV diversity

Recombination has been shown to occur in all genera of coronaviruses, including MERS-CoV (***Lai et al., 1985***; ***Makino et al., 1986***; ***Keck et al., 1988***; ***Kottier et al., 1995***; ***Herrewegh et al., 1998***). In order to quantify the degree to recombination has shaped MERS-CoV genetic diversity we used two recombination detection approaches across partitions of taxa corresponding to inferred MERS-CoV clades. Both methods rely on sampling parental and recombinant alleles within the alignment, although each quantifies different signals of recombination. One hallmark of recombination is the ability to carry alleles derived via mutation from one lineage to another, which appear as repeated mutations taking place in the recipient lineage, somewhere else in the tree. The PHI (pairwise homoplasy index) test quantifies the appearance of these excessive repeat mutations (homoplasies) within an alignment (***Bruen et al., 2006***). Another hallmark of recombination is clustering of alleles along the genome, due to how template switching, the primary mechanism of recombination in RNA viruses, occurs. 3Seq relies on the clustering of nucleotide similarities along the genome between sequence triplets – two potential parent-donors and one candidate offspring-recipient sequences (***Boni et al., 2007***).

Both tests can give spurious results in cases of extreme rate heterogeneity and sampling over time (***Dudas and Rambaut, 2016***), but both tests have not been reported to fail simultaneously. PHI and 3Seq methods consistently identify most of the apparent 'backbone' of the MERS-CoV phylogeny as encompassing sequences with evidence of recombination (Figure 4-Figure supplement 1). Neither method can identify where in the tree recombination occurred, but each full asterisk in Figure 4-Figure supplement 1 should be interpreted as the minimum partition of data that still captures both donor and recipient alleles involved in a recombination event. This suggests a non-negligible contribution of recombination in shaping existing MERS-CoV diversity. As done previously (***Dudas and Rambaut, 2016***), we show large numbers of homoplasies in MERS-CoV data

(Figure 4-Figure supplement 2) with some evidence of genomic clustering of such alleles. These results are consistent with high incidence of MERS-CoV in camels (***Müller et al., 2014***; ***Corman et al., 2014***; ***Chu et al., 2014***; ***Reusken et al., 2014***; ***Ali et al., 2017***), allowing for co-infection with distinct genotypes and thus recombination to occur.

Owing to these findings, we performed a sensitivity analysis in which we partitioned the MERS-CoV genome into two fragments and identified human outbreak clusters within each fragment. We find strong similarity in the grouping of human cases into outbreak clusters between fragments (Figure 4A). Between the two trees in figure 4B four (out of 54) 'human' clades are expanded where either singleton introductions or two-taxon clades in fragment 2 join other clades in fragment 1. For the reverse comparison there are five 'human' clades (out of 53) in fragment 2 that are expanded. All such clades have low divergence (figure 4B) and thus incongruences in human clades are more likely to be caused by differences in deme assignment rather than actual recombination. And while we observe evidence of distinct phylogenetic trees from different parts of the MERS-CoV genome (Figure 4B), human clades are minimally affected and large portions of the posterior probability density in both parts of the genome are concentrated in shared clades (Figure 4-Figure supplement 3). Additionally, we observe the same source-sink dynamics between camel and human populations in trees constructed from separate genomic fragments as were observed in the original full genome tree (Figures 1, 4B).

Observed departures from strictly clonal evolution suggest that while recombination is an issue for inferring MERS-CoV phylogenies, its effect on the human side of MERS outbreaks is minimal, as expected if humans represent a transient host with little opportunity for co-infection. MERS-CoV evolution on the reservoir side is complicated by recombination, though is nonetheless still largely amenable to phylogenetic methods. Amongst other parameters of interest, recombination is expected to interfere with molecular clocks, where transferred genomic regions can give the impression of branches undergoing rapid evolution, or branches where recombination results in reversions appearing to evolve slow. In addition to its potential to influence tree topology, recombination in molecular sequence data is an erratic force with unpredictable effects. We suspect that the effects of recombination in MERS-CoV data are reigned in by a relatively small effective population size of the virus in Saudi Arabia (see next section) where haplotypes are fixed or nearly fixed, thus preventing an accumulation of genetic diversity that would then be reshuffled via recombination. Nevertheless, we choose not to report on any particular estimates for times of common ancestors (tMRCAs), even though these are expected to be somewhat robust for dating human clusters, and we do not report on the evolutionary rate of the virus, even though it appears to fall firmly within the expected range for RNA viruses: regression of nucleotide differences to Jordan-N3/2012 genome against sequence collection dates yields a rate of 4.59×10^{-4} subs/site/year, Bayesian structured coalescent estimate from primary analysis 9.57×10^{-4} (95% HPDs: $8.28 - 10.9 \times 10^{-4}$) subs/site/year.

MERS-CoV shows population turnover in camels

Here we attempt to investigate MERS-CoV demographic patterns in the camel reservoir. We supplement camel sequence data with a single earliest sequence from each human cluster, treating viral diversity present in humans as a sentinel sample of MERS-CoV diversity circulating in camels. This removes conflicting demographic signals sampled during human outbreaks, where densely sampled closely related sequences from humans could be misconstrued as evidence of demographic crash in the viral population.

Despite lack of convergence, neither of the two demographic reconstructions show evidence of fluctuations in the relative genetic diversity ($N_e\tau$) of MERS-CoV over time (Figure 5). We recover a similar demographic trajectory when estimating $N_e\tau$ over time with a skygrid tree prior across the genome split into ten fragments with independent phylogenetic trees to account for confounding

effects of recombination (Figures 5, 5-Figure supplement 1). However, we do note that coalescence rate estimates are high relative to the sampling time period, with a mean estimate of $N_e\tau$ at 3.49 years (95% HPD: 2.71–4.38), and consequently MERS-CoV phylogeny resembles a ladder, as often seen in human influenza A virus phylogenies (***Bedford et al., 2011***).

This empirically estimated effectived population can be compared to the expected effective population size in a simple epidemiological model. At endemic equilibrium, we expect scaled effective population size $N_e\tau$ to follow $I / 2\beta$, where β is the equilibrium rate of transmission and I is the equilibrium number of infecteds (***Frost and Volz, 2010***). We assume that β is constant and is equal to the rate of recovery. Given a 20 day duration of infection in camels (***Adney et al., 2014***), we arrive at $\beta = 365/20 = 18.25$ infections per year. Given extremely high seroprevalence estimates within camels in Saudi Arabia (***Müller et al., 2014***; ***Corman et al., 2014***; ***Chu et al., 2014***; ***Reusken et al., 2013***, ***2014***), we expect camels to usually be infected within their first year of life. Therefore we can estimate the rough number of camel infections per year as the number of calves produced each year. We find there are $830\,000$ camels in Saudi Arabia (***Abdallah and Faye, 2013***) and that female camels in Saudi Arabia have an average fecundity of 45% (***Abdallah and Faye, 2013***). Thus, we expect $830\,000 \times 0.50 \times 0.45 = 186\,750$ new calves produced yearly and correspondingly $186\,750$ new infections every year, which spread over 20 day intervals gives an average prevalence of $I = 10\,233$ infections. This results in an expected scaled effective population size $N_e\tau = 280.4$ years.

Comparing expected $N_e\tau$ to empirical $N_e\tau$ we arrive at a ratio of 80.3 (64.0–103.5). This is less than the equivalent ratio for human measles virus (***Bedford et al., 2011***) and is in line with the expectation from neutral evolutionary dynamics plus some degree of transmission heterogeneity (***Volz et al., 2013***) and seasonal troughs in prevalence. Thus, we believe that the ladder-like appearance of the MERS-CoV tree can likely be explained by entirely demographic factors.

Discussion

MERS-CoV epidemiology

In this study we aimed to understand the drivers of MERS coronavirus transmission in humans and what role the camel reservoir plays in perpetuating the epidemic in the Arabian peninsula by using sequence data collected from both hosts (174 from humans and 100 from camels). We showed that currently existing models of population structure (***Vaughan et al., 2014***) can identify distinct demographic modes in MERS-CoV genomic data, where viruses continuously circulating in camels repeatedly jump into humans and cause small outbreaks doomed to extinction (Figures 1, 1-Figure supplement 1). This inference succeeds under different choices of priors for unknown demographic parameters (Figure 1-Figure supplement 2) and in the presence of strong biases in sequence sampling schemes (Figure 3). When rapid coalescence in the human deme is not allowed (Figure 1-Figure supplement 4) structured coalescent inference loses power and ancestral state reconstruction is nearly identical to that of discrete trait analysis (Figure 1-Figure supplement 3). When allowed different deme-specific population sizes, the structured coalescent model succeeds because a large proportion of human sequences fall into tightly connected clusters, which informs a low estimate for the population size of the human deme. This in turn informs the inferred state of long ancestral branches in the phylogeny, *i.e.* because these long branches are not immediately coalescing, they are most likely in camels.

From sequence data we identify at least 50 zoonotic introductions of MERS-CoV into humans from the reservoir (Figure 1), from which we extrapolate that hundreds more such introductions must have taken place (Figure 3). Although we recover migration rates from our model (Figure 1-Figure supplement 2), these only pertain to sequences and in no way reflect the epidemiologically relevant *per capita* rates of zoonotic spillover events. We also looked at potential seasonality in MERS-CoV spillover into humans. Our analyses indicated a period of three months where the odds of a

sequenced spillover event are increased, with timing consistent with an enzootic amongst camel calves (Figure 2). As a result of our identification of large and asymmetric flow of viral lineages into humans we also find that the basic reproduction number for MERS-CoV in humans is well below the epidemic threshold (Figure 3). Having said that, there are highly customisable coalescent methods available that extend the methods used here to accommodate time varying migration rates and population sizes, integrate alternative sources of information and fit to stochastic nonlinear models (***Rasmussen et al., 2014***), which would be more appropriate for MERS-CoV. Some distinct aspects of MERS-CoV epidemiology could not be captured in our methodology, such as hospital outbreaks where R_0 is expected to be consistently closer to 1.0 compared to community transmission of MERS-CoV. Outside of coalescent-based models there are population structure models that explicitly relate epidemiological parameters to the branching process observed in sequence data (***Kühnert et al., 2016***), but often rely on specifying numerous informative priors and can suffer from MCMC convergence issues.

Strong population structure in viruses often arises through limited gene flow, either due to geography (***Dudas et al., 2017***), ecology (***Smith et al., 2009***) or evolutionary forces (***Turner et al., 2005***; ***Dudas et al., 2015***). On a smaller scale population structure can unveil important details about transmission patterns, such as identifying reservoirs and understanding spillover trends and risk, much as we have done here. When properly understood naturally arising barriers to gene flow can be exploited for more efficient disease control and prevention, as well as risk management.

Transmissibility differences between zoonoses and pandemics

Severe acute respiratory syndrome (SARS) coronavirus, a Betacoronavirus like MERS-CoV, caused a serious epidemic in humans in 2003, with over 8000 cases and nearly 800 deaths. Since MERS-CoV was also able to cause significant pathogenicity in the human host it was inevitable that parallels would be drawn between MERS-CoV and SARS-CoV at the time of MERS discovery in 2012. Although we describe the epidemiology of MERS-CoV from sequence data, indications that MERS-CoV has poor capacity to spread human-to-human existed prior to any sequence data. SARS-CoV swept through the world in a short period of time, but MERS cases trickled slowly and were restricted to the Arabian Peninsula or resulted in self-limiting outbreaks outside of the region, a pattern strongly indicative of repeat zoonotic spillover. Infectious disease surveillance and control measures remain limited, so much like the SARS epidemic in 2003 or the H1N1 pandemic in 2009, zoonotic pathogens with $R_0 > 1.0$ are probably going to be discovered after spreading beyond the original location of spillover. Even though our results show that MERS-CoV does not appear to present an imminent global threat, we would like to highlight that its epidemiology is nonetheless concerning.

Pathogens *Bacillus anthracis*, Andes hantavirus (***Martinez et al., 2005***), monkeypox (***Reed et al., 2004***) and influenza A are able to jump species barriers but only influenza A viruses have historically resulted in pandemics (***Lipsitch et al., 2016***). MERS-CoV may join the list of pathogens able to jump species barriers but not spread efficiently in the new host. Since its emergence in 2012, MERS-CoV has caused self-limiting outbreaks in humans with no evidence of worsening outbreaks over time. However, sustained evolution of the virus in the reservoir and continued flow of viral lineages into humans provides the substrate for a more transmissible variant of MERS-CoV to possibly emerge. Previous modeling studies (***Antia et al., 2003***; ***Park et al., 2013***) suggest a positive relationship between initial R_0 in the human host and probability of evolutionary emergence of a novel strain which passes the supercritical threshold of $R_0 > 1.0$. This leaves MERS-CoV in a precarious position; on one hand its current R_0 of ~0.7 is certainly less than 1, but its proximity to the supercritical threshold raises alarm. With very little known about the fitness landscape or adaptive potential of MERS-CoV, it is incredibly challenging to predict the likelihood of the emergence more transmissible variants. In light of these difficulties, we encourage continued genomic surveillance of MERS-CoV in the camel reservoir and from sporadic human cases to rapidly identify a supercritical variant, if one

does emerge.

Methods

Sequence data

All MERS-CoV sequences were downloaded from GenBank and accession numbers are given in Supplementary File 1.Sequences without accessions were kindly shared by Ali M. Somily, Mazin Barry, Sarah S. Al Subaie, Abdulaziz A. BinSaeed, Fahad A. Alzamil, Waleed Zaher, Theeb Al Qahtani, Khaldoon Al Jerian, Scott J.N. McNabb, Imad A. Al-Jahdali, Ahmed M. Alotaibi, Nahid A. Batarfi, Matthew Cotten, Simon J. Watson, Spela Binter, and Paul Kellam prior to publication. Fragments of some strains submitted to GenBank as separate accessions were assembled into a single sequence. Only sequences covering at least 50% of MERS-CoV genome were kept, to facilitate later analyses where the alignment is split into two parts in order to account for effects of recombination (***Dudas and Rambaut, 2016***). Sequences were annotated with available collection dates and hosts, designated as camel or human, aligned with MAFFT (***Katoh and Standley, 2013***), and edited manually. Protein coding sequences were extracted and concatenated, reducing alignment length from 30130 down to 29364, which allowed for codon-partitioned substitution models to be used. The final dataset consisted of 174 genomes from human infections and 100 genomes from camel infections (Supplementary File 1).

Phylogenetic analyses

Primary analysis, structured coalescent

For our primary analysis, the MultiTypeTree module (***Vaughan et al., 2014***) of BEAST v2.4.3 (***Bouckaert et al., 2014***) was used to specify a structured coalescent model with two demes – humans and camels. At time of writing structured population models are available in BEAST v2 (***Bouckaert et al., 2014***) but not in BEAST v1 (***Drummond et al., 2012***). We use the more computationally intensive MultiTypeTree module (***Vaughan et al., 2014***) over approximate methods also available in BEAST v2, such as BASTA (***Maio et al., 2015***), MASCOT (***Mueller et al., 2017***), and PhyDyn (***Volz, 2011***). Structured coalescent model implemented in the MultiTypeTree module (***Vaughan et al., 2014***) estimates four parameters: rate of coalescence in human viruses, rate of coalescence in camel viruses, rate of migration from the human deme to the camel deme and rate of migration from the camel deme to the human deme. Analyses were run on codon position partitioned data with two separate HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) nucleotide substitution models specified for codon positions 1+2 and 3. A relaxed molecular clock with branch rates drawn from a lognormal distribution (***Drummond et al., 2006***) was used to infer the evolutionary rate from date calibrated tips. Default priors were used for all parameters except for migration rates between demes for which an exponential prior with mean 1.0 was used. All analyses were run for 200 million steps across ten independent Markov chains (MCMC runs) and states were sampled every $20\,000$ steps. Due to the complexity of multitype tree parameter space 50% of states from every analysis were discarded as burn-in, convergence assessed in Tracer v1.6 and states combined using LogCombiner distributed with BEAST v2.4.3 (***Bouckaert et al., 2014***). Three chains out of ten did not converge and were discarded altogether. This left $70\,000$ states on which to base posterior inference. Posterior sets of typed (where migrating branches from structured coalescent are collapsed, and migration information is left as a switch in state between parent and descendant nodes) trees were summarised using TreeAnnotator v2.4.3 with the common ancestor heights option (***Heled and Bouckaert, 2013***). A maximum likelihood phylogeny showing just the genetic relationships between MERS-CoV genomes from camels and humans was recovered using PhyML (***Guindon et al., 2003***) under a HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) nucleotide substitution model and is shown in Figure 1-Figure supplement 5.

Control, structured coalescent with different prior

As a secondary analysis to test robustness to choice of prior, we set up an analysis where we increased the mean of the exponential distribution prior for migration rate to 10.0. All other parameters were identical to the primary analysis and as before 10 independent MCMC chains were run. In this case, two chains did not converge. This left $80\,000$ states on which to base posterior inference. Posterior sets of typed trees were summarised using TreeAnnotator v2.4.3 with the common ancestor heights option (***Heled and Bouckaert, 2013***).

Control, structured coalescent with equal deme sizes

To better understand where statistical power of the structured coalescent model lies we set up a tertiary analysis where a model was set up identically to the first structured coalescent analysis, but deme population sizes were enforced to have the same size. This analysis allowed us to differentiate whether statistical power in our analysis is coming from effective population size contrasts between demes or the backwards-in-time migration rate estimation. Five replicate chains were set up, two of which failed to converge after 200 million states. Combining the three converging runs left us with $15\,000$ trees sampled from the posterior distribution, which were summarised in TreeAnnotator v2.4.3 with the common ancestor heights option (***Heled and Bouckaert, 2013***).

Control, structured coalescent with more than one tree per genome

Due to concerns that recombination might affect our conclusions (***Dudas and Rambaut, 2016***), as an additional secondary analysis, we also considered a scenario where alignments were split into two fragments (fragment 1 comprised of positions 1-21000, fragment 2 of positions 21000-29364), with independent clocks, trees and migration rates, but shared substitution models and deme population sizes. Fragment positions were chosen based on consistent identification of the region around nucleotide 21000 as a probable breakpoint by GARD (***Pond et al., 2006***) by previous studies into SARS and MERS coronaviruses (***Hon et al., 2008***; ***Dudas and Rambaut, 2016***). All analyses were set to run for 200 million states, subsampling every $20\,000$ states. Chains not converging after 200 million states were discarded. 20% of the states were discarded as burn-in, convergence assessed with Tracer 1.6 and combined with LogCombiner. Three chains out of ten did not converge. This left $70\,000$ states on which to base posterior inference. Posterior sets of typed trees were summarised using TreeAnnotator v2.4.3 with the common ancestor heights option (***Heled and Bouckaert, 2013***).

Control, discrete trait analysis

A currently widely used approach to infer ancestral states in phylogenies relies on treating traits of interest (such as geography, host, *etc.*) as features evolving along a phylogeny as continuous time Markov chains with an arbitrary number of states (***Lemey et al., 2009***). Unlike structured coalescent methods, such discrete trait approaches are independent from the tree (*i.e.* demographic) prior and thus unable to influence coalescence rates under different trait states. Such models have been used in the past to infer the number of MERS-CoV host jumps (***Zhang et al., 2016***) with results contradicting other sources of information. In order to test how a discrete trait approach compares to the structured coalescent we used our 274 MERS-CoV genome data set in BEAST v2.4.3 (***Bouckaert et al., 2014***) and specified identical codon-partitioned nucleotide substitution and molecular clock models to our structured coalescent analysis. To give the most comparable results we used a constant population size coalescent model, as this is the demographic function for each deme in the structured coalescent model. Five replicate MCMC analyses were run for 200 million states, three of which converged onto the same posterior distribution. The converging chains were combined after removing 20 million states as burn-in to give a total of $27\,000$ trees drawn from the

posterior distribution. These trees were then summarised using TreeAnnotator v2.4.5 with the common ancestor heights option (***Heled and Bouckaert, 2013***).

Introduction seasonality

We extracted the times of camel-to-human introductions from the posterior distribution of multitype trees. This distribution of introduction times was then discretised as follows: for sample $k = 1, 2, \ldots, L$ from the posterior, Z_{ijk} was 1 if there as an introduction in month i and year j and 0 otherwise. We model the sum of introductions at month i and year j across the posterior sample $Y_{ij} = \sum_{k=1}^{L} Z_{ijk}$ with the hierarchical model:

$$
\begin{aligned}
Y_{ij} &\sim \text{Binomial}(L, \theta_{ij}) \\
\theta_{ij} &= \text{logistic}(\alpha_j + \beta_i) \\
\alpha_j &\sim \text{Normal}(\mu_y, \sigma_y) \\
\mu_y &\sim \text{Normal}(0, 1) \\
\sigma_y &\sim \text{Cauchy}(0, 2.5) \\
\beta_i &\sim \text{Normal}(0, \sigma_m) \\
\sigma_m &\sim \text{Cauchy}(0, 2.5),
\end{aligned}
$$

where α_j represents the contribution of year to expected introduction count and β_i represents the contribution of month to expected introduction count. Here, $\text{logistic}(\alpha_j + \beta_i) = \frac{\exp(\alpha_j + \beta_i)}{\exp(\alpha_j + \beta_i) + 1}$. We sampled posterior values from this model via the Markov chain Monte Carlo methods implemented in Stan (***Carpenter et al., 2016***). Odds ratios of introductions were computed for each month i as $\text{OR}_i = \exp(\beta_i)$.

Epidemiological analyses

Here, we employ a Monte Carlo simulation approach to identify parameters consistent with observed patterns of sequence clustering (Figure 3-Figure supplement 2). Our structured coalescent analyses indicate that every MERS outbreak is a contained cross-species spillover of the virus from camels into humans. The distribution of the number of these cross-species transmissions and their sizes thus contain information about the underlying transmission process. At heart, we expect fewer larger clusters if fundamental reproductive number R_0 is large and more smaller clusters if R_0 is small. A similar approach was used in ***Grubaugh et al.*** (***2017***) to estimate R_0 for Zika introductions into Florida.

Branching process theory provides an analytical distribution for the number of eventual cases j in a transmission chain resulting from a single introduction event with R_0 and dispersion parameter ω (***Blumberg and Lloyd-Smith, 2013***). This distribution follows

$$
\Pr(j | R_0, \omega) = \frac{\Gamma(\omega j + j - 1)}{\Gamma(\omega j)\,\Gamma(j+1)} \, \frac{(\frac{R_0}{\omega})^{j-1}}{(1 + \frac{R_0}{\omega})^{\omega j + j - 1}}. \tag{1}
$$

Here, R_0 represents the expected number of secondary cases following a single infection and ω represents the dispersion parameter assuming secondary cases follow a negative binomial distribution (***Lloyd-Smith et al., 2005***), so that smaller values represent larger degrees of heterogeneity in the transmission process.

As of 10 May 2017, the World Health Organization has been notified of 1952 cases of MERS-CoV (***World Health Organization, 2017***). We thus simulated final transmission chain sizes using Equation 1 until we reached an epidemic comprised of $N = 2000$ cases. $10\,000$ simulations were run for 121

uniformly spaced values of R_0 across the range [0.5–1.1] with dispersion parameter ω fixed to 0.1 following expectations from previous studies of coronavirus behavior (***Lloyd-Smith et al., 2005***). Each simulation results in a vector of outbreak sizes $\mathbf{c}$, where c_i is the size of the *i*th transmission cluster and $\sum_{i=1}^{K} c_i = 2000$ and K is the number of clusters generated.

Following the underlying transmission process generating case clusters $\mathbf{c}$ we simulate a secondary process of sampling some fraction of cases and sequencing them to generate data analogous to what we empirically observe. We sample from the case cluster size vector $\mathbf{c}$ without replacement according to a multivariate hypergeometric distribution (Algorithm 1). The resulting sequence cluster size vector $\mathbf{s}$ contains K entries, some of which are zero (*i.e.* case clusters not sequenced), but $\sum_{i=1}^{K} s_i = 174$ which reflects the number of human MERS-CoV sequences used in this study. Note that this "sequencing capacity" parameter does not vary over time, even though MERS-CoV sequencing efforts have varied in intensity, starting out slow due to lack of awareness, methods and materials and increasing in response to hospital outbreaks later. As the default sampling scheme operates under equiprobable sequencing, we also simulated biased sequencing by using concentrated hypergeometric distributions where the probability mass function is squared (bias = 2) or cubed (bias = 3) and then normalized. Here, bias enriches the hypergeometric distribution so that sequences are sampled with weights proportional to $(c_1^{\text{bias}}, c_2^{\text{bias}}, \ldots, c_k^{\text{bias}})$. Thus, bias makes larger clusters more likely to be 'sequenced'.

After simulations were completed, we identified simulations in which the recovered distribution of sequence cluster sizes $\mathbf{s}$ fell within the 95% highest posterior density intervals for four summary statistics of empirical MERS-CoV sequence cluster sizes recovered via structured coalescent analysis: number of sequence clusters, mean, standard deviation and skewness (third central moment). These values were 48-61 for number of sequence clusters, 2.87–3.65 for mean sequence cluster size, 4.84–6.02 for standard deviation of sequence cluster sizes, and 415.40–621.06 for skewness of sequence cluster sizes.

We performed a smaller set of simulations with 2500 replicates and twice the number of cases, *i.e.* $\sum_{i=1}^{K} C_i = 4000$, to explore a dramatically underreported epidemic. Additionally, we performed additional smaller set of simulations on a rougher grid of R_0 values (23 values, 0.50–1.05), with 5 values of dispersion parameter ω (0.002, 0.04, 0.1, 0.5, 1.0) and 3 levels of bias (1, 2, 3) to justify our choice of dispersion parameter ω that was fixed to 0.1 in the main analyses (Figure 3-Figure

supplement 6).

Data: Array of case cluster sizes in outbreak $\mathbf{c} = (c_1, c_2, \ldots, c_K)$, sequences available M, total outbreak size N, where $N = \sum_{i=1}^{K} c_i$.
Result: Array of sequence cluster sizes sampled: $\mathbf{s} = (s_1, s_2, \ldots, s_K)$.
Draw s_i from a hypergeometric distribution with c_i successes, $N - c_i$ failures after M trials;
while $i < K$ **do**
  $i = i + 1$;
  $M = M - s_{i-1}$;
  Compute the probability mass function (pmf) for all possible values of s_i, $\mathbf{p} = (p(0)^{\text{bias}}, p(1)^{\text{bias}}, \ldots, p(c_i)^{\text{bias}}) \times (\sum_i p_i^{\text{bias}})^{-1}$, where $p(\cdot)$ is the pmf for a hypergeometric distribution with c_i successes, $N - c_i$ failures after M trials;
  Draw a sequence cluster size s_i from array of potential sequence cluster sizes $(0, 1, \ldots, c_i)$ according to $\boldsymbol{p}$;
end
Add remaining sequences to last sequence cluster $c_K = M - s_{K-1}$;

Algorithm 1: Multivariate hypergeometric sampling scheme. Pseudocode describes the multivariate hypergeometric sampling scheme that simulates sequencing. Probability of sequencing a given number of cases from a case cluster depends on cluster size and sequences left (*i.e.* "sequencing capacity"). The bias parameter determines how probability mass function of the hypergeometric distribution is concentrated.

Demographic inference of MERS-CoV in the reservoir

In order to infer the demographic history of MERS-CoV in camels we used the results of structured coalescent analyses to identify introductions of the virus into humans. The oldest sequence from each cluster introduced into humans was kept for further analysis. This procedure removes lineages coalescing rapidly in humans, which would otherwise introduce a strong signal of low effective population size. These subsampled MERS-CoV sequences from humans were combined with existing sequence data from camels to give us a dataset with minimal demographic signal coming from epidemiological processes in humans. Sequences belonging to the outgroup clade where most of MERS-CoV sequences from Egypt fall were removed out of concern that MERS epidemics in Saudi Arabia and Egypt are distinct epidemics with relatively poor sampling in the latter. Were more sequences of MERS-CoV available from other parts of Africa we speculate they would fall outside of the diversity that has been sampled in Saudi Arabia and cluster with early MERS-CoV sequences from Jordan and sequences from Egyptian camels. However, currently there are no indications of what MERS-CoV diversity looks like in camels east of Saudi Arabia. A flexible skygrid tree prior (***Gill et al., 2013***) was used to recover estimates of relative genetic diversity ($N_e\tau$) at 50 evenly spaced grid points across six years, ending at the most recent tip in the tree (2015 August) in BEAST v1.8.4 (***Drummond et al., 2012***), under a relaxed molecular clock with rates drawn from a lognormal distribution (***Drummond et al., 2006***) and codon position partitioned (positions $1 + 2$ and 3) HKY+Γ_4 (***Hasegawa et al., 1985***; ***Yang, 1994***) nucleotide substitution models. At time of writing advanced flexible coalescent tree priors from the skyline family, such as skygrid (***Gill et al., 2013***) are available in BEAST v1 (***Drummond et al., 2012***) but not in BEAST v2 (***Bouckaert et al., 2014***). We set up five independent MCMC chains to run for 500 million states, sampling every $50\,000$ states. This analysis suffered from poor convergence, where two chains converged onto one stationary distribution, two to another and the last chain onto a third stationary distribution, with high effective sample sizes. Demographic trajectories recovered by the two main stationary distributions are very similar and differences between the two appear to be caused by convergence onto subtly different tree topologies. This non-convergence effect may have been masked previously by the use of all available MERS-CoV sequences from humans which may have lead MCMC towards one of the multiple stationary distributions.

To ensure that recombination was not interfering with the skygrid reconstruction we also split our MERS-CoV alignment into ten parts 2937 nucleotides long. These were then used as separate partitions with independent trees and clock rates in BEAST v1.8.4 (***Drummond et al., 2012***). Nucleotide substitution and relaxed clock models were set up identically to the whole genome skygrid analysis described above (***Drummond et al., 2006***; ***Hasegawa et al., 1985***; ***Yang, 1994***). Skygrid coalescent tree prior (***Gill et al., 2013***) was used jointly across all ten partitions for demographic inference. Five MCMC chains were set up, each running for 200 million states, sampling every $20\,000$ states.

Data availability

Sequence data and all analytical code is publicly available at github.com/blab/structured-mers (***Dudas, 2017***).

Acknowledgements

We would like to thank Allison Black for useful discussion and advice. GD is supported by the Mahan postdoctoral fellowship from the Fred Hutchinson Cancer Research Center. TB is a Pew Biomedical Scholar and is supported by NIH R35 GM119774-01. AR was supported in part by the European Union Seventh Framework Programme for research, technological development and demonstration under Grant Agreement no. 278433-PREDEMICS and no. 725422-RESERVOIRDOCS, and the Wellcome Trust through project 206298/Z/17/Z.

We gratefully acknowledge the contribution of the following scientists for sharing MERS-CoV sequence data before publication:

Ali M. Somily[1], Mazin Barry[1], Sarah S. Al Subaie[1], Abdulaziz A. BinSaeed[1], Fahad A. Alzamil[1], Waleed Zaher[1], Theeb Al Qahtani[1], Khaldoon Al Jerian[1], Scott J.N. McNabb[2], Imad A. Al-Jahdali[3], Ahmed M. Alotaibi[4], Nahid A. Batarfi[5], Matthew Cotten[6], Simon J. Watson[6], Spela Binter[6], Paul Kellam[6].

[1]College of Medicine, King Saud University, Riyadh, Kingdom of Saudi Arabia
[2]Rollins School of Public Health, Emory University, Atlanta, GA, USA
[3]Deputy Minister. Ex. General Director King Fahad General Hospital, Jeddah and Occupational and environmental medicine, Um AlQura University, Kingdom of Saudi Arabia
[4]Department of Intensive Care, Prince Mohammed bin Abdulaziz Hospital, Riyadh, Kingdom of Saudi Arabia
[5]Epidemiology section, Command and Control Center (CCC) Ministry of Health, Jeddah
[6]Wellcome Trust Sanger Institute, Hinxton, United Kingdom

References

Abdallah H, Faye B. Typology of camel farming system in Saudi Arabia. Emirates Journal of Food and Agriculture. 2013; 25(4):250.

Adney DR, van Doremalen N, Brown VR, Bushmaker T, Scott D, de Wit E, Bowen RA, Munster VJ. Replication and Shedding of MERS-CoV in Upper Respiratory Tract of Inoculated Dromedary Camels. Emerging Infectious Diseases. 2014 Dec; 20(12):1999–2005. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4257817/, doi:10.3201/eid2012.141280.

Ali MA, Shehata MM, Gomaa MR, Kandeil A, El-Shesheny R, Kayed AS, El-Taweel AN, Atea M, Hassan N, Bagato O, Moatasim Y, Mahmoud SH, Kutkat O, Maatouq AM, Osman A, McKenzie PP, Webby RJ, Kayali G. Systematic, active surveillance for Middle East respiratory syndrome coronavirus in camels in Egypt. Emerg Microbes Infect. 2017; 6(1):e1.

Almutairi SE, Boujenane I, Musaad A, Awad-Acharari F. Non-genetic factors influencing reproductive traits and calving weight in Saudi camels. Trop Anim Health Prod. 2010; 42(6):1087–1092.

Antia R, Regoes RR, Koella JC, Bergstrom CT. The role of evolution in the emergence of infectious diseases. Nature. 2003 Dec; 426(6967):658–661. https://www.nature.com/nature/journal/v426/n6967/full/nature02104.html, doi: 10.1038/nature02104.

Assiri A, Al-Tawfiq JA, Al-Rabeeah AA, Al-Rabiah FA, Al-Hajjar S, Al-Barrak A, Flemban H, Al-Nassir WN, Balkhy HH, Al-Hakeem RF, Makhdoom HQ, Zumla AI, Memish ZA. Epidemiological, demographic, and clinical characteristics of 47 cases of Middle East respiratory syndrome coronavirus disease from Saudi Arabia: a descriptive study. Lancet Infect Dis. 2013; 13(9):752–761.

Assiri A, McGeer A, Perl TM, Price CS, Al Rabeeah AA, Cummings DAT, Alabdullatif ZN, Assad M, Almulhim A, Makhdoom H, Madani H, Alhakeem R, Al-Tawfiq JA, Cotten M, Watson SJ, Kellam P, Zumla AI, Memish ZA. Hospital outbreak of Middle East respiratory syndrome coronavirus. N Engl J Med. 2013; 369(5):407–416.

Bedford T, Cobey S, Pascual M. Strength and tempo of selection revealed in viral gene genealogies. BMC Evol Biol. 2011; 11:220.

Blumberg S, Lloyd-Smith JO. Inference of R_0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput Biol. 2013; 9(5):e1002993.

Boheemen Sv, Graaf Md, Lauber C, Bestebroer TM, Raj VS, Zaki AM, Osterhaus ADME, Haagmans BL, Gorbalenya AE, Snijder EJ, Fouchier RAM. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. mBio. 2012; 3(6):e00473–12.

Boni MF, Posada D, Feldman MW. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics. 2007; 176(2):1035–1047.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014; 10(4):e1003537.

Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006; 172(4):2665–2681.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. J Stat Softw. 2016; 20:1–37.

Cauchemez S, Nouvellet P, Cori A, Jombart T, Garske T, Clapham H, Moore S, Mills HL, Salje H, Collins C, Rodriquez-Barraquer I, Riley S, Truelove S, Algarni H, Alhakeem R, AlHarbi K, Turkistani A, Aguas RJ, Cummings DAT, Kerkhove MDV, et al. Unraveling the drivers of MERS-CoV transmission. Proc Natl Acad Sci USA. 2016; 113(32):9081–9086.

Chen X, Chughtai AA, Dyda A, MacIntyre CR. Comparative epidemiology of Middle East respiratory syndrome coronavirus (MERS-CoV) in Saudi Arabia and South Korea. Emerg Microbes Infect. 2017; 6(6):e51.

Chu DKW, Poon LLM, Gomaa MM, Shehata MM, Perera RAPM, Abu Zeid D, El Rifay AS, Siu LY, Guan Y, Webby RJ, Ali MA, Peiris M, Kayali G. MERS Coronaviruses in Dromedary Camels, Egypt. Emerg Infect Dis. 2014; 20(6):1049–1053.

Corman VM, Jores J, Meyer B, Younan M, Liljander A, Said MY, Gluecks I, Lattwein E, Bosch BJ, Drexler JF, Bornstein S, Drosten C, Müller MA. Antibodies against MERS Coronavirus in Dromedary Camels, Kenya, 1992–2013. Emerg Infect Dis. 2014; 20(8):1319–1322.

Cotten M, Watson SJ, Kellam P, Al-Rabeeah AA, Makhdoom HQ, Assiri A, Al-Tawfiq JA, Alhakeem RF, Madani H, AlRabiah FA, Hajjar SA, Al-nassir WN, Albarrak A, Flemban H, Balkhy HH, Alsubaie S, Palser AL, Gall A, Bashford-Rogers R, Rambaut A, et al. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. Lancet. 2013; 382(9909):1993–2002.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLoS Biol. 2006 Mar; 4(5):e88. http://dx.doi.org/10.1371/journal.pbio.0040088, doi: 10.1371/journal.pbio.0040088.

Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29(8):1969–1973.

Dudas G, mers-structure: Looking into MERS-CoV dynamics through the structured coalescent lens. Bedford Lab; 2017. https://github.com/blab/mers-structure. b1fe9abbd633222342f7850ec01a494812e2ca9b.

Dudas G, Bedford T, Lycett S, Rambaut A. Reassortment between influenza B lineages and the emergence of a coadapted PB1–PB2–HA gene complex. Mol Biol Evol. 2015; 32(1):162–172.

Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, Park DJ, Ladner JT, Arias A, Asogun D, Bielejec F, Caddy SL, Cotten M, D'Ambrozio J, Dellicour S, Di Caro A, Diclaro JW, Duraffour S, Elmore MJ, Fakoli LS, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017 Apr; 544(7650):309–315. https://www.nature.com/nature/journal/v544/n7650/abs/nature22040.html, doi: 10.1038/nature22040.

Dudas G, Rambaut A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. Virus Evol. 2016; 2(1):vev023.

Fagbo SF, Skakni L, Chu DKW, Garbati MA, Joseph M, Hakawi AM. Molecular Epidemiology of Hospital Outbreak of Middle East Respiratory Syndrome, Riyadh, Saudi Arabia, 2014. Emerg Infect Dis. 2015; 21(11):1981.

Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Phil Trans R Soc B. 2013; 368:20120196.

Frost SD, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philos Trans Royal Soc B Trans R Soc B. 2010; 365(1548):1879–1890.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. Mol Biol Evol. 2013; 30(3):713.

Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014 Sep; 345(6202):1369–1372. http://science.sciencemag.org/content/345/6202/1369, doi: 10.1126/science.1259657.

Grubaugh ND, Ladner JT, Kraemer MU, Dudas G, Tan AL, Gangavarapu K, Wiley MR, White S, Thézé J, Magnani DM, Prieto K, Reyes D, Bingham A, Paul LM, Robles-Sikisaka R, Oliveira G, Pronty D, Metsky HC, Baniecki ML, Barnes KG, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017; 546:401–405.

Guindon S, Gascuel O, Rannala B. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology. 2003 Oct; 52(5):696–704. https://academic.oup.com/sysbio/article/52/5/696/1681984, doi: 10.1080/10635150390235520.

Hasegawa M, Kishino H, Yano Ta. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985; 22(2):160–174.

Heled J, Bouckaert RR. Looking for trees in the forest: summary tree from posterior samples. BMC Evolutionary Biology. 2013 Oct; 13:221. https://doi.org/10.1186/1471-2148-13-221, doi: 10.1186/1471-2148-13-221.

Herrewegh AAPM, Smeenk I, Horzinek MC, Rottier PJM, Groot RJd. Feline Coronavirus Type II Strains 79-1683 and 79-1146 Originate from a Double Recombination between Feline Coronavirus Type I and Canine Coronavirus. J Virol. 1998; 72(5):4508–4514.

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FCC. 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. J Virol. 2008; 82(4):1819–1826.

Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013 Apr; 30(4):772–780. https://academic.oup.com/mbe/article/30/4/772/1073398/MAFFT-Multiple-Sequence-Alignment-Software-Version, doi: 10.1093/molbev/mst010.

Keck JG, Matsushima GK, Makino S, Fleming JO, Vannier DM, Stohlman SA, Lai MM. In vivo RNA-RNA recombination of coronavirus in mouse brain. J Virol. 1988; 62(5):1810–1813.

Kottier SA, Cavanagh D, Britton P. Experimental Evidence of Recombination in Coronavirus Infectious Bronchitis Virus. Virology. 1995 Nov; 213(2):569–580. http://www.sciencedirect.com/science/article/pii/S0042682285700293, doi: 10.1006/viro.1995.0029.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Molecular Biology and Evolution. 2016 Aug; 33(8):2102–2116. https://academic.oup.com/mbe/article/33/8/2102/2578541, doi: 10.1093/molbev/msw064.

Lai MM, Baric RS, Makino S, Keck JG, Egbert J, Leibowitz JL, Stohlman SA. Recombination between nonsegmented RNA genomes of murine coronaviruses. J Virol. 1985; 56(2):449–456.

675 **Lemey P**, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLOS Computa-
676 tional Biology. 2009 Sep; 5(9):e1000520. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.
677 pcbi.1000520, doi: 10.1371/journal.pcbi.1000520.

678 **Lipsitch M**, Barclay W, Raman R, Russell CJ, Belser JA, Cobey S, Kasson PM, Lloyd-Smith JO, Maurer-Stroh S, Riley
679 S, et al. Viral factors in influenza pandemic risk assessment. eLife. 2016; 5:e18491.

680 **Liu D**, Shi W, Shi Y, Wang D, Xiao H, Li W, Bi Y, Wu Y, Li X, Yan J, et al. Origin and diversity of novel avian influenza
681 A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. Lancet. 2013;
682 381(9881):1926–1932.

683 **Lloyd-Smith JO**, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on
684 disease emergence. Nature. 2005 Nov; 438(7066):355–359. https://www.nature.com/nature/journal/v438/
685 n7066/full/nature04153.html, doi: 10.1038/nature04153.

686 **Lycett S**, Bodewes R, Pohlmann A, Banks J, Banyai K, Boni M, Bouwstra R, Breed A, Brown I, Chen H, et al. Role
687 for migratory wild birds in the global spread of avian influenza H5N8. Science. 2016; 354(6309):213–217.

688 **Maio ND**, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent
689 Approximation. PLOS Genetics. 2015 Aug; 11(8):e1005421. http://journals.plos.org/plosgenetics/article?id=10.
690 1371/journal.pgen.1005421, doi: 10.1371/journal.pgen.1005421.

691 **Makino S**, Keck JG, Stohlman SA, Lai MM. High-frequency RNA recombination of murine coronaviruses. J Virol.
692 1986; 57(3):729–737.

693 **Martinez VP**, Bellomo C, San Juan J, Pinna D, Forlenza R, Elder M, Padula PJ. Person-to-person transmission of
694 Andes virus. Emerg Infect Dis. 2005; 11(12):1848–1853.

695 **Martinez-Bakker M**, Bakker KM, King AA, Rohani P. Human birth seasonality: latitudinal gradient and interplay
696 with childhood disease dynamics. In: *Proc R Soc B*, vol. 281; 2014. p. 20132438.

697 **Memish ZA**, Cotten M, Meyer B, Watson SJ, Alsahafi AJ, Rabeeah AAA, Corman VM, Sieberg A, Makhdoom HQ,
698 Assiri A, Masri MA, Aldabbagh S, Bosch BJ, Beer M, Müller MA, Kellam P, Drosten C. Human Infection with
699 MERS Coronavirus after Exposure to Infected Camels, Saudi Arabia, 2013. Emerg Infect Dis. 2014; 20(6):1012.

700 **Mueller NF**, Rasmussen DA, Stadler T. MASCOT: Parameter and state inference under the marginal structured
701 coalescent approximation. bioRxiv. 2017 Sep; p. 188516. https://www.biorxiv.org/content/early/2017/09/13/
702 188516, doi: 10.1101/188516.

703 **Müller MA**, Corman VM, Jores J, Meyer B, Younan M, Liljander A, Bosch BJ, Lattwein E, Hilali M, Musa BE,
704 Bornstein S, Drosten C. MERS Coronavirus Neutralizing Antibodies in Camels, Eastern Africa, 1983–1997.
705 Emerg Infect Dis. 2014; 20(12).

706 **Müller MA**, Meyer B, Corman VM, Al-Masri M, Turkestani A, Ritz D, Sieberg A, Aldabbagh S, Bosch BJ, Lattwein E,
707 Alhakeem RF, Assiri AM, Albarrak AM, Al-Shangiti AM, Al-Tawfiq JA, Wikramaratna P, Alrabeeah AA, Drosten
708 C, Memish ZA. Presence of Middle East respiratory syndrome coronavirus antibodies in Saudi Arabia: a
709 nationwide, cross-sectional, serological study. Lancet Infect Dis. 2015; 15(5):559–564.

710 **Notohara M**. The coalescent and the genealogical process in geographically structured population. J Math Biol.
711 1990; 29:59–75.

712 **Park M**, Loverdo C, Schreiber SJ, Lloyd-Smith JO. Multiple scales of selection influence the evolutionary emer-
713 gence of novel pathogens. Philos Trans Royal Soc B. 2013; 368(1614):20120333.

714 **Park SS**, Wernery U, Corman VM, Wong EYM, Tsang AKL, Muth D, Lau SKP, Khazanehdari K, Zirkel F, Ali M, Nagy P,
715 Juhasz J, Wernery R, Joseph S, Syriac G, Elizabeth SK, Patteril NAG, Woo PCY, Drosten C. Acute Middle East
716 Respiratory Syndrome Coronavirus Infection in Livestock Dromedaries, Dubai, 2014. Emerg Infect Dis. 2015;
717 21(6):1019.

718 **Pond K**, L S, Posada D, Gravenor MB, Woelk CH, Frost SDW. GARD: a genetic algorithm for recombination
719 detection. Bioinformatics. 2006 Dec; 22(24):3096–3098. https://academic.oup.com/bioinformatics/article/22/
720 24/3096/208339/GARD-a-genetic-algorithm-for-recombination, doi: 10.1093/bioinformatics/btl474.

721 **Quick J**, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, De Pinna E, Robinson
722 E, Struthers K, Webber M, Catto A, Dallman TJ, Hawkey P, Loman NJ. Rapid draft sequencing and real-
723 time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biology. 2015 May; 16:114.
724 https://doi.org/10.1186/s13059-015-0677-2, doi: 10.1186/s13059-015-0677-2.

Rasmussen DA, Volz EM, Koelle K. Phylodynamic Inference for Structured Epidemiological Models. PLOS Computational Biology. 2014 Apr; 10(4):e1003570. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003570, doi: 10.1371/journal.pcbi.1003570.

Reed KD, Melski JW, Graham MB, Regnery RL, Sotir MJ, Wegner MV, Kazmierczak JJ, Stratman EJ, Li Y, Fairley JA, Swain GR, Olson VA, Sargent EK, Kehl SC, Frace MA, Kline R, Foldy SL, Davis JP, Damon IK. The detection of monkeypox in humans in the Western Hemisphere. N Engl J Med. 2004; 350(4):342–350.

Reusken CBEM, Farag EABA, Haagmans BL, Mohran KA, Godeke GJ, V, Raj S, Alhajri F, Al-Marri SA, Al-Romaihi HE, Al-Thani M, Bosch BJ, Eijk AAvd, El-Sayed AM, Ibrahim AK, Al-Molawi N, Müller MA, Pasha SK, Drosten C, AlHajri MM, et al. Occupational exposure to dromedaries and risk for MERS-CoV infection, Qatar, 2013–2014. Emerg Infect Dis. 2015; 21(8):1422.

Reusken CB, Haagmans BL, Müller MA, Gutierrez C, Godeke GJ, Meyer B, Muth D, Raj VS, Vries LSD, Corman VM, Drexler JF, Smits SL, El Tahir YE, De Sousa R, van Beek J, Nowotny N, van Maanen K, Hidalgo-Hermoso E, Bosch BJ, Rottier P, et al. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. Lancet Infect Dis. 2013; 13(10):859–866.

Reusken CBEM, Messadi L, Feyisa A, Ularamu H, Godeke GJ, Danmarwa A, Dawo F, Jemli M, Melaku S, Shamaki D, Woma Y, Wungak Y, Gebremedhin EZ, Zutt I, Bosch BJ, Haagmans BL, Koopmans MPG. Geographic distribution of MERS coronavirus among dromedary camels, Africa. Emerg Infect Dis. 2014; 20(8):1370–1374.

Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, Chen H, Webster RG, Peiris JSM, Guan Y. Dating the emergence of pandemic influenza viruses. Proc Natl Acad Sci USA. 2009; 106(28):11709–11712.

The WHO MERS-CoV Research Group. State of knowledge and data gaps of Middle East respiratory syndrome coronavirus (MERS-CoV) in humans. PLoS Curr. 2013; Edition 1.

Turner TL, Hahn MW, Nuzhdin SV. Genomic islands of speciation in Anopheles gambiae. PLoS Biol. 2005; 3(9):e285.

Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. 2014 Aug; 30(16):2272–2279. https://academic.oup.com/bioinformatics/article/30/16/2272/2748160/Efficient-Bayesian-inference-under-the-structured, doi: 10.1093/bioinformatics/btu201.

Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013; 9(3):e1002947.

Volz EM. Complex Population Dynamics and the Coalescent under Neutrality. Genetics. 2011 Jan; p. genetics.111.134627. http://www.genetics.org/content/early/2011/10/27/genetics.111.134627, doi: 10.1534/genetics.111.134627.

Wernery U. Camelid immunoglobulins and their importance for the new-born – a review. J Vet Med B. 2001; 48(8):561–568.

World Health Organization. Disease outbreak news – 2016 December 19; 2016, available at http://www.who.int/csr/don/19-december-2016-2-mers-saudi-arabia/en/.

World Health Organization. WHO MERS-CoV global summary and assessment of risk; 2017, available at http://www.who.int/emergencies/mers-cov/risk-assessment-july-2017.pdf.

Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J Mol Evol. 1994; 39(3):306–314.

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med. 2012; 367(19):1814–1820.

Zhang Z, Shen L, Gu X. Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. Sci Rep. 2016; 6:25049.

Supplementary File 1. Strain names, accessions (where available), identified host and reported collection dates for MERS-CoV genomes used in this study.

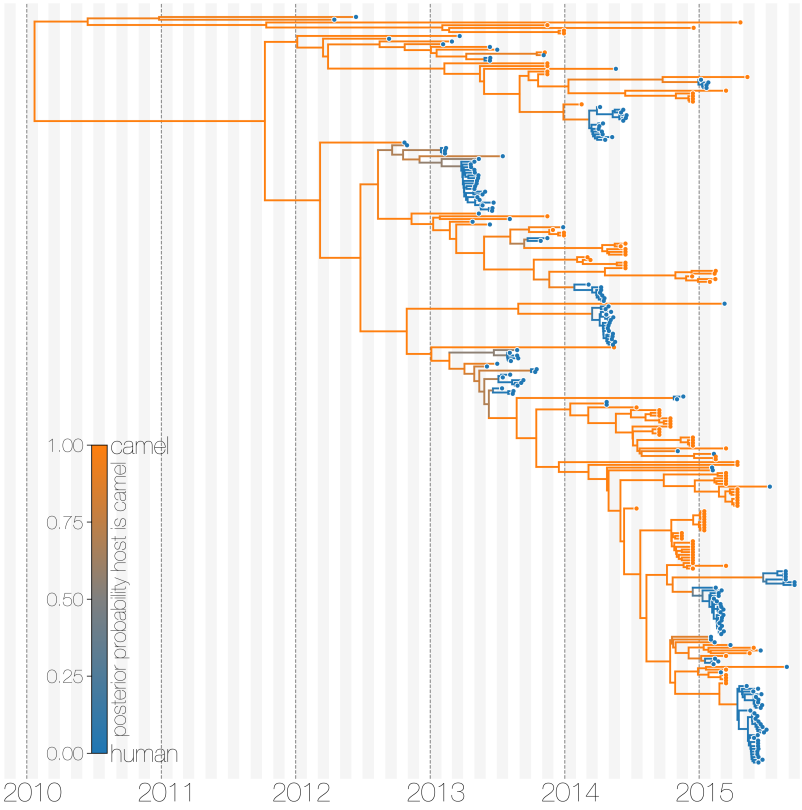


Figure 1. Typed maximum clade credibility tree of MERS-CoV genomes from 174 human viruses and 100 camel viruses. Maximum clade credibility (MCC) tree showing inferred ancestral hosts for MERS-CoV recovered with the structured coalescent. The vast majority of MERS-CoV evolution is inferred to occur in camels (orange) with human outbreaks (blue) representing evolutionary dead-ends for the virus. Confidence in host assignment is depicted as a colour gradient, with increased uncertainty in host assignment (posterior probabilities close to 0.5) shown as grey. While large clusters of human cases are apparent in the tree, significant contributions to human outbreaks are made by singleton sequences, likely representing recent cross-species transmissions that were caught early.

Figure 1–Figure supplement 1. Evolutionary history of MERS-CoV partitioned between camels and humans. This is the same tree as shown in Figure 1, but with contiguous stretches of MERS-CoV evolutionary history split by inferred host: camels (top in orange) and humans (bottom in blue). This visualisation highlights the ephemeral nature of MERS-CoV outbreaks in humans, compared to continuous circulation of the virus in camels.

Figure 1–Figure supplement 2. Posterior backwards migration rate estimates for two choices of prior. Negligible flow of MERS-CoV lineages from humans into camels is recovered regardless of prior choice (note that rates are backwards in time). Plots show the 95% highest posterior density for the estimated migration rate from the human deme into the camel deme looking backwards in time (orange) and *vice versa* (blue). Dotted lines indicate exponential priors specified for migration rates, with mean 1.0 (bottom) or 10.0 (top).

Figure 1–Figure supplement 3. Maximum clade credibility (MCC) tree with ancestral state reconstruction according to a discrete trait model. MCC tree is presented the same as Figure 1 and Figure 1-Figure supplement 4, with colours indicating the most probable state reconstruction at internal nodes. Unlike the structured coalescent summary shown in Figure 1 where camels are reconstructed as the main host where MERS-CoV persists, the discrete trait approach identifies both camels and humans as major hosts with humans being the source of MERS-CoV infection in camels.

Figure 1–Figure supplement 4. Maximum clade credibility (MCC) tree of structured coalescent model with enforced equal coalescence rates. MCC tree is presented the same as Figures 1 and 1-Figure supplement 3, with colours indicating the most probable state reconstruction at internal nodes. Similar to Figure 1-Figure supplement 3 enforcing equal coalescence rates between demes in a structured coalescent model identifies humans as a major MERS-CoV host and the source of viruses in camels.

Figure 1–Figure supplement 5. Maximum likelihood (ML) tree of MERS-CoV genomes coloured by origin of sequence. Maximum likelihood tree shows genetic divergence between MERS-CoV genomes collected from camels (orange tips) and humans (blue tips).

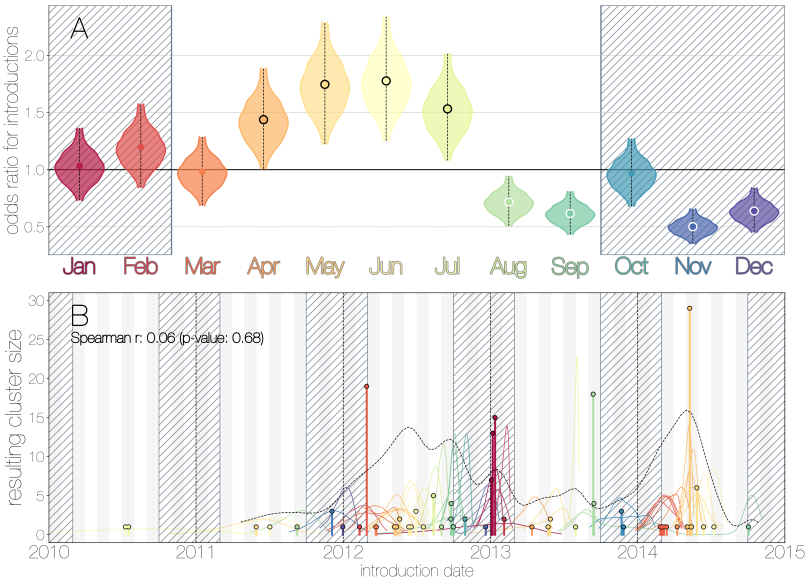


Figure 2. Seasonality of MERS-CoV introduction events. A) Posterior density estimates partitioned by month showing the 95% highest posterior density interval for relative odds ratios of MERS-CoV introductions into humans. Posterior means are indicated with circles. Evidence for increased or decreased risk (95% HPD excludes 1.0) for introductions are indicated by black or white circles, respectively. Hatched area spanning October to February indicates the camel calving season. B) Sequence cluster sizes and inferred dates of introduction events. Each introduction event is shown as a vertical line positioned based on the median introduction time, as recovered by structured coalescent analyses and coloured by time of year with height indicating number of descendant sequences recovered from human cases. 95% highest posterior density intervals for introductions of MERS-CoV into humans are indicated with coloured lines, coloured by median estimated introduction time. The black dotted line indicates the joint probability density for introductions. We find little correlation between date and size of introduction (Spearman $\rho = 0.06$, $p = 0.68$).

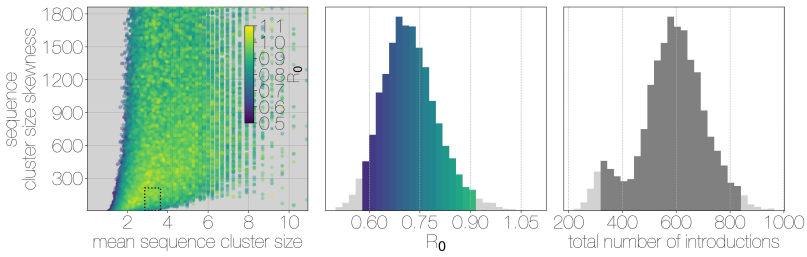


Figure 3. Monte Carlo simulations of human transmission clusters. Leftmost scatter plot shows the distribution of individual Monte Carlo simulation sequence cluster size statistics (mean and skewness) coloured by the R_0 value used for the simulation. The dotted rectangle identifies the 95% highest posterior density bounds for sequence cluster size mean and skewness observed for empirical MERS-CoV data. The distribution of R_0 values that fall within 95% HPDs for sequence cluster size mean, standard deviation, skewness and number of introductions, is shown in the middle, on the same y-axis. Bins falling inside the 95% percentiles are coloured by R_0, as in the leftmost scatter plot. The distribution of total number of introductions associated with simulations matching MERS-CoV sequence clusters is shown on the right. Darker shade of grey indicates bins falling within the 95% percentiles. Monte Carlo simulations indicate R_0 for MERS-CoV in humans is likely to be below 1.0, with numbers of zoonotic transmissions numbering in the hundreds.

Figure 3–Figure supplement 1. Monte Carlo simulations of human transmission clusters. From top to bottom each row corresponds to departures from completely random sequencing efforts with respect to case cluster size (bias parameter=1.0) to sequencing increasingly biased towards capturing large case clusters (bias=2.0, bias=3.0). Leftmost scatter plots show the distribution of individual Monte Carlo simulation sequence cluster size statistics (mean and skewness) coloured by the R_0 value used for the simulation. The dotted rectangle identifies the 95% highest posterior density bounds for sequence cluster size mean and skewness observed for empirical MERS-CoV data. The distribution of R_0 values matching empirical data are shown in the middle, on the same y-axis across all levels of the bias parameter. Under unbiased sequencing (bias=1.0) only 0.45% of simulations fit our phylogenetic observations, while 1.79% and 1.67% of simulations fit for bias levels of 2.0 and 3.0, respectively. Correspondingly, we estimate 11.6% support for a model with bias level 1.0, 45.7% support for a model with bias level 2.0, and 42.7% support for a model with bias level 3.0. Bins falling inside the 95% percentiles are coloured by R_0, as in the leftmost scatter plot. While the 95% percentiles for R_0 values are close to 1.0 (0.71–0.98) for the unbiased sequencing simulation (*i.e.* uniform sequencing efforts, in which every case is equally likely to be sequenced), we also note that increasing levels of bias are considerably more to likely to generate MERS-CoV-like sequence clusters. The distribution of total number of introductions associated with simulations matching MERS-CoV sequence clusters is shown in the plots on the right, on the same y-axis across all levels of bias. Darker shade of grey indicates bins falling within the 95% percentiles. The median number of cross-species introductions observed in simulations matching empirical data without bias are 346 (95% percentiles 262–439). These numbers jump up to 568 (95% percentiles 430–727) for bias = 2.0 and 656 (95% percentiles 488–853) for bias = 3.0 simulations. Model averaging would suggest plausible numbers of introductions between 311 and 811.

Figure 3–Figure supplement 2. Monte Carlo simulation schematic. Case clusters are simulated according to Equation 1 until an outbreak size of 2000 cases is reached. We sample 174 cases from each simulation to represent sequencing of human MERS cases. 'Sequencing' is carried out by using multivariate hypergeometric sampling, representing sampling cases without replacement to be sequenced. Sequencing simulations take place at three levels of bias: 1.0, where every case is equally likely to be sequenced, and 2.0 and 3.0, where cases from larger clusters are increasingly more likely to be sequenced. The distribution of simulated sequence clusters is summarised by its mean, median and standard deviation. A simulation is considered to match if the mean, median and standard deviation of its sequence cluster sizes falls within the 95% highest posterior density interval of observed MERS-CoV sequence clusters. R_0 values that ultimately generate data matching empirical observations, as well as associated numbers of 'introductions' are retained as estimates. These estimates are summarised in Figure 3.

Figure 3–Figure supplement 3. Results of Monte Carlo simulations with vast underestimation of cases. The plot is identical to Figure 3-Figure supplement 1, but instead of 2000 cases, simulations were run with 4000 cases. With more unobserved cases the R_0 values matching observed MERS-CoV sequence clusters can only be smaller, with a corresponding increase in numbers of zoonotic transmissions. However, the numbers of simulations that match MERS-CoV data go down as well.

Figure 3–Figure supplement 4. Boxplots of matching simulated case and sequence cluster distributions. Boxplots indicate frequency of case (blue, top) and sequence (red, bottom) cluster sizes across simulations at different bias levels, marginalised across R_0 values. Outliers are shown with transparency, medians are indicated with thick black lines. Case clusters exhibit a strong skew with large numbers of singleton introductions and a substantial tail at higher levels of bias.

Figure 3–Figure supplement 5. Quantile-quantile (Q-Q) plot of empirical and simulated sequence cluster sizes. Density of sequence cluster size percentiles (1st–99th, calculated across a grid of 50 values) calculated for random states from the posterior distribution (x-axis) and matching simulations (y-axis). Most values fall on the one-to-one line, with a heavier tail in mid-sized sequence clusters in empirical data, manifesting as a greater density of points below the one-to-one line in the middle.

Figure 3–Figure supplement 6. Numbers of epidemiological simulations conforming to empirical observations. Numbers indicate the total number of epidemiological simulations under each combination of

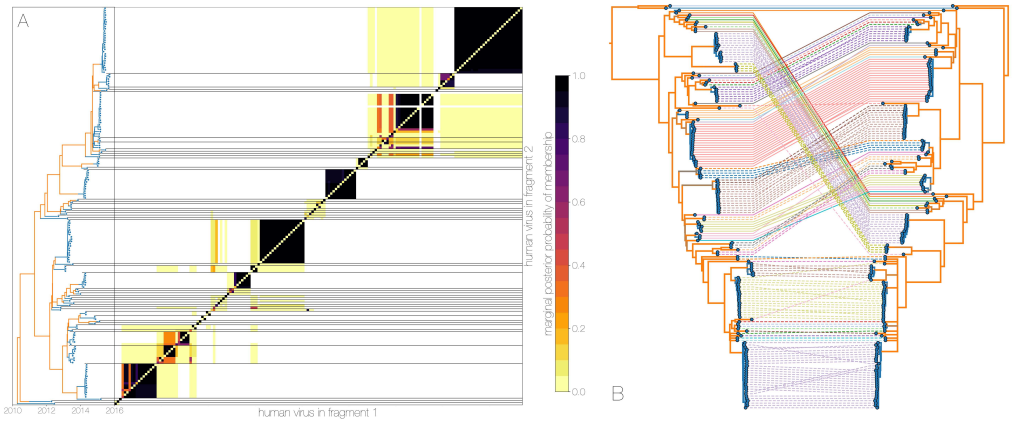


Figure 4. Recombinant features of MERS-CoV phylogenies. A) Marginal posterior probabilities of taxa collected from humans belonging to the same clade in phylogenies derived from different parts of the genome. Taxa are ordered according to phylogeny of fragment 2 (genome positions 21001 to 29364) reduced to just the human tips and is displayed on the left. Human clusters are largely well-supported as monophyletic and consistent across trees of both genomic fragments. B) Tanglegram connecting the same taxa between a phylogeny derived from fragment 1 (left, genome positions 1 to 21000) and fragment 2 (right, genome positions 21001 to 29364), reduced to just the human tips and branches with posterior probability < 0.1 collapsed. Human clusters exhibit limited diversity and corresponding low levels of incongruence within an introduction cluster.

Figure 4–Figure supplement 1. Tests of recombination across MERS-CoV clades. Maximum clade credibility tree of MERS-CoV genomes annotated with results of two recombination detection tests (PHI and 3Seq) applied to descendent sequences of each clade. Both tests identify large portions of existing sequence data as containing signals of recombination. Note that markings do not indicate where recombinations have occurred on the tree, merely the minimum distance in sequence/time space between recombining lineages.

Figure 4–Figure supplement 2. MERS-CoV genomes exhibit high numbers of non-clonal loci. Ancestral state reconstruction (right) identifies a large number of sites in which mutations have occurred more than once in the tree (homoplasies, orange) or are reversions (red) from a state arising in an ancestor. Mutations that apparently only occur once in the tree (synapomorphies) are shown in grey. The maximum likelihood phylogeny on the left is coloured by whether sequences were sampled in humans (blue) or camels (orange).

Figure 4–Figure supplement 3. Human clade sharing between genomic fragments 1 and 2. Central scatter plot shows the posterior probability of human clades shared between genomic fragments 1 and 2, in their respective trees. Left and bottom scatter plots track the posterior probability of human clades only observed in fragment 2 (left) or fragment 1 (bottom). The cumulative probability of human clades present in either tree are tracked by plots on the right (fragment 2) and top (fragment 1). Most of the probability mass is concentrated within human clades that are present in trees of both genomic fragment 1 and 2 (0.9701 and 0.9474 of all human clades across posteriors, respectively).

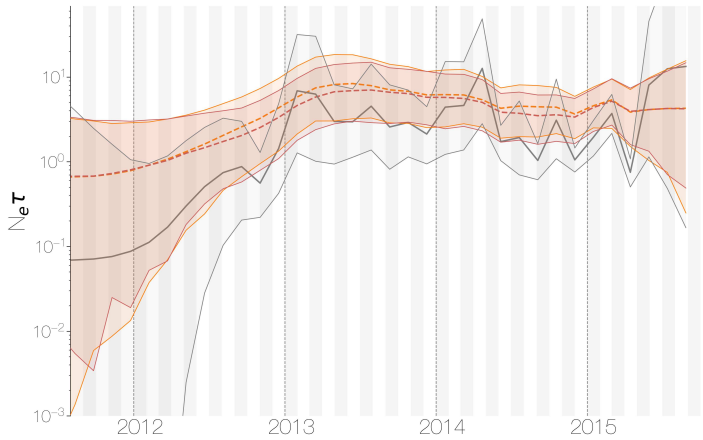


Figure 5. Demographic history of MERS-CoV in Arabian peninsula camels. Demographic history of MERS-CoV in camels, as inferred via a skygrid coalescent tree prior (***Gill et al., 2013***). Three skygrid reconstructions are shown, red and orange for each of the stationary distributions reached by MCMC with the whole genome and a black one where the genome was split into ten partitions. Shaded interval indicates the 95% highest posterior density interval for the product of generation time and effective population size, $N_e\tau$. Midline tracks the inferred median of $N_e\tau$.

Figure 5–Figure supplement 1. Skygrid comparison between whole and fragmented genomes. Inferred median $N_e\tau$ recovered using a skygrid tree prior on whole genome (bottom) and ten genomic fragments with independent trees (left), coloured by time. Dotted line indicates the one-to-one line.

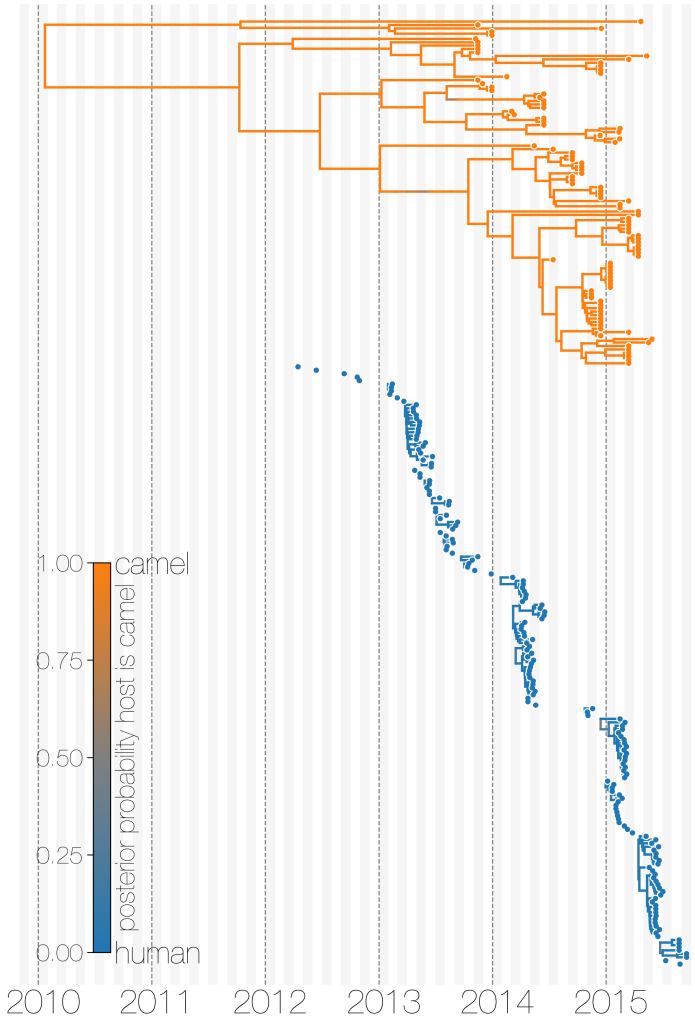


Figure 1–Figure supplement 1. Evolutionary history of MERS-CoV partitioned between camels and humans. This is the same tree as shown in Figure 1, but with contiguous stretches of MERS-CoV evolutionary history split by inferred host: camels (top in orange) and humans (bottom in blue). This visualisation highlights the ephemeral nature of MERS-CoV outbreaks in humans, compared to continuous circulation of the virus in camels.

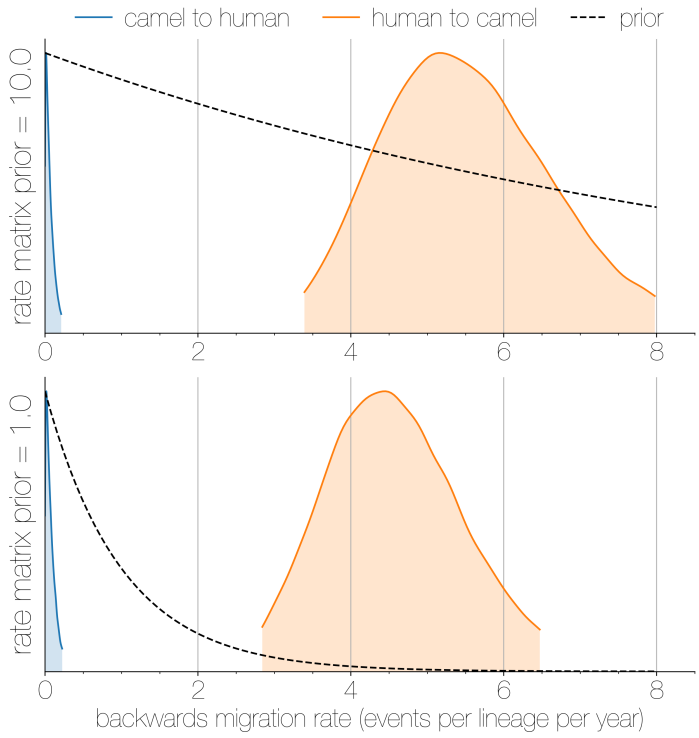


Figure 1–Figure supplement 2. Posterior backwards migration rate estimates for two choices of prior. Negligible flow of MERS-CoV lineages from humans into camels is recovered regardless of prior choice (note that rates are backwards in time). Plots show the 95% highest posterior density for the estimated migration rate from the human deme into the camel deme looking backwards in time (orange) and *vice versa* (blue). Dotted lines indicate exponential priors specified for migration rates, with mean 1.0 (bottom) or 10.0 (top).

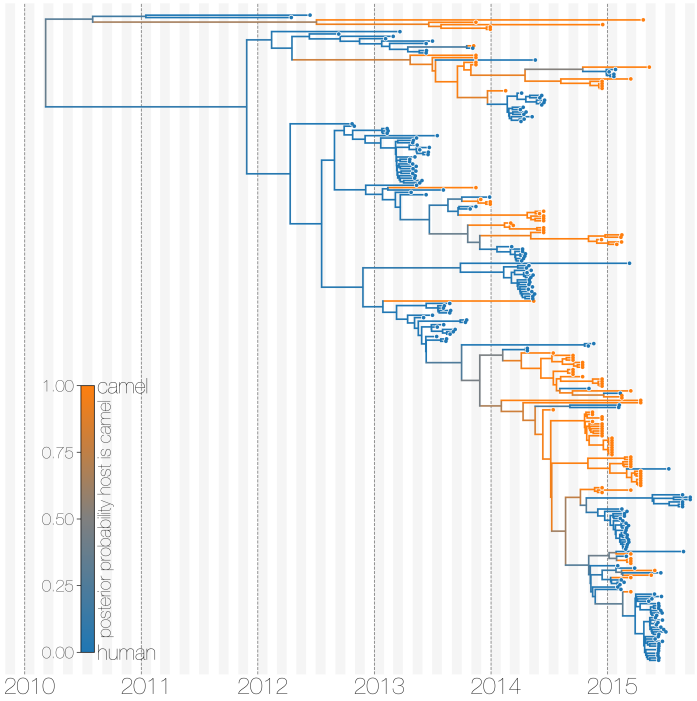


Figure 1–Figure supplement 3. Maximum clade credibility (MCC) tree with ancestral state reconstruction according to a discrete trait model. MCC tree is presented the same as Figure 1 and Figure 1-Figure supplement 4, with colours indicating the most probable state reconstruction at internal nodes. Unlike the structured coalescent summary shown in Figure 1 where camels are reconstructed as the main host where MERS-CoV persists, the discrete trait approach identifies both camels and humans as major hosts with humans being the source of MERS-CoV infection in camels.

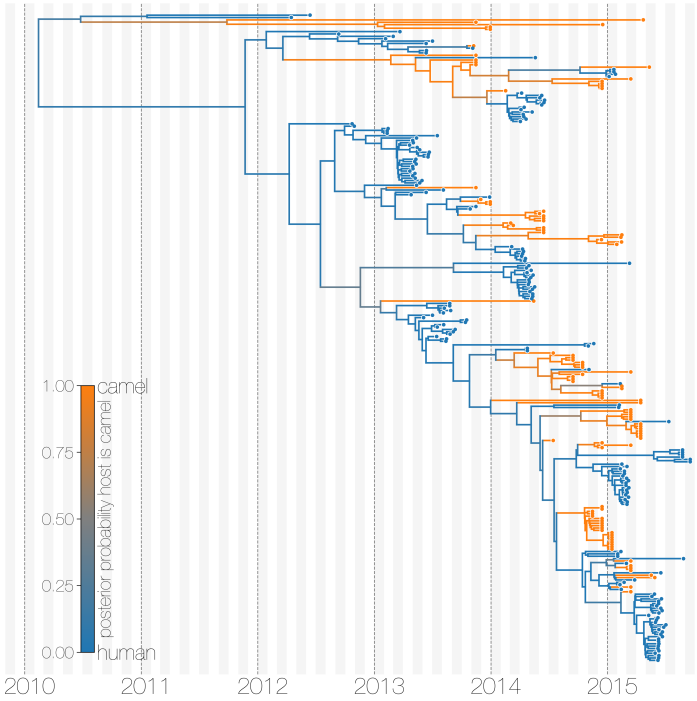


Figure 1–Figure supplement 4. Maximum clade credibility (MCC) tree of structured coalescent model with enforced equal coalescence rates. MCC tree is presented the same as Figures 1 and 1-Figure supplement 3, with colours indicating the most probable state reconstruction at internal nodes. Similar to Figure 1-Figure supplement 3 enforcing equal coalescence rates between demes in a structured coalescent model identifies humans as a major MERS-CoV host and the source of viruses in camels.

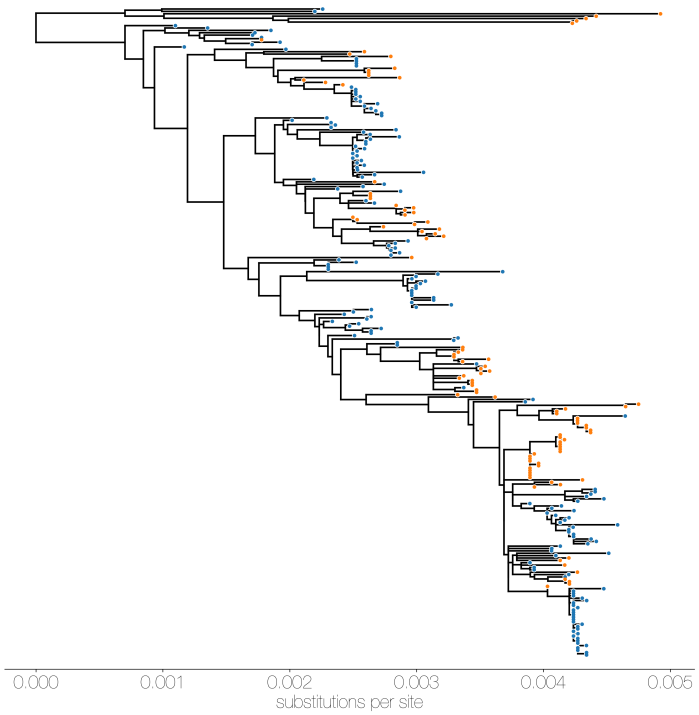


Figure 1–Figure supplement 5. Maximum likelihood (ML) tree of MERS-CoV genomes coloured by origin of sequence. Maximum likelihood tree shows genetic divergence between MERS-CoV genomes collected from camels (orange tips) and humans (blue tips).

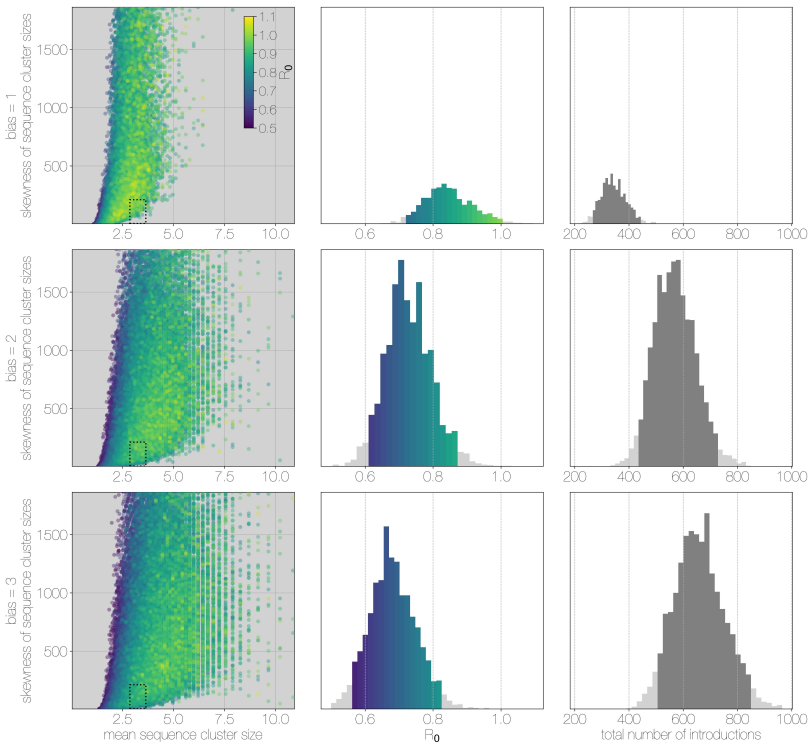
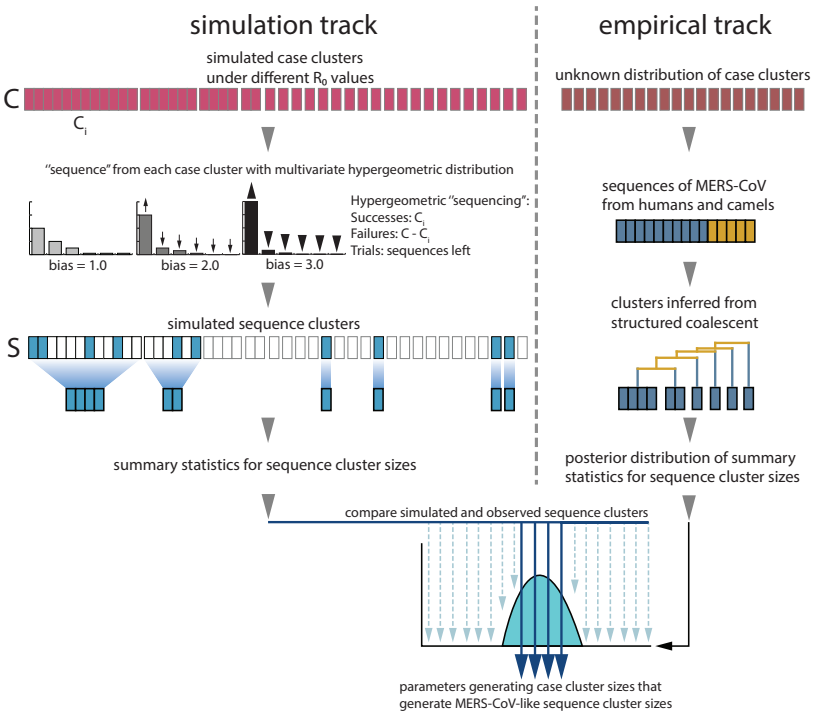


Figure 3–Figure supplement 1. Monte Carlo simulations of human transmission clusters. From top to bottom each row corresponds to departures from completely random sequencing efforts with respect to case cluster size (bias parameter=1.0) to sequencing increasingly biased towards capturing large case clusters (bias=2.0, bias=3.0). Leftmost scatter plots show the distribution of individual Monte Carlo simulation sequence cluster size statistics (mean and skewness) coloured by the R_0 value used for the simulation. The dotted rectangle identifies the 95% highest posterior density bounds for sequence cluster size mean and skewness observed for empirical MERS-CoV data. The distribution of R_0 values matching empirical data are shown in the middle, on the same y-axis across all levels of the bias parameter. Under unbiased sequencing (bias=1.0) only 0.45% of simulations fit our phylogenetic observations, while 1.79% and 1.67% of simulations fit for bias levels of 2.0 and 3.0, respectively. Correspondingly, we estimate 11.6% support for a model with bias level 1.0, 45.7% support for a model with bias level 2.0, and 42.7% support for a model with bias level 3.0. Bins falling inside the 95% percentiles are coloured by R_0, as in the leftmost scatter plot. While the 95% percentiles for R_0 values are close to 1.0 (0.71–0.98) for the unbiased sequencing simulation (*i.e.* uniform sequencing efforts, in which every case is equally likely to be sequenced), we also note that increasing levels of bias are considerably more to likely to generate MERS-CoV-like sequence clusters. The distribution of total number of introductions associated with simulations matching MERS-CoV sequence clusters is shown in the plots on the right, on the same y-axis across all levels of bias. Darker shade of grey indicates bins falling within the 95% percentiles. The median number of cross-species introductions observed in simulations matching empirical data without bias are 346 (95% percentiles 262–439). These numbers jump up to 568 (95% percentiles 430–727) for bias = 2.0 and 656 (95% percentiles 488–853) for bias = 3.0 simulations. Model averaging would suggest plausible numbers of introductions between 311 and 811.



775

Figure 3–Figure supplement 2. Monte Carlo simulation schematic. Case clusters are simulated according to Equation 1 until an outbreak size of 2000 cases is reached. We sample 174 cases from each simulation to represent sequencing of human MERS cases. 'Sequencing' is carried out by using multivariate hypergeometric sampling, representing sampling cases without replacement to be sequenced. Sequencing simulations take place at three levels of bias: 1.0, where every case is equally likely to be sequenced, and 2.0 and 3.0, where cases from larger clusters are increasingly more likely to be sequenced. The distribution of simulated sequence clusters is summarised by its mean, median and standard deviation. A simulation is considered to match if the mean, median and standard deviation of its sequence cluster sizes falls within the 95% highest posterior density interval of observed MERS-CoV sequence clusters. R_0 values that ultimately generate data matching empirical observations, as well as associated numbers of 'introductions' are retained as estimates. These estimates are summarised in Figure 3.

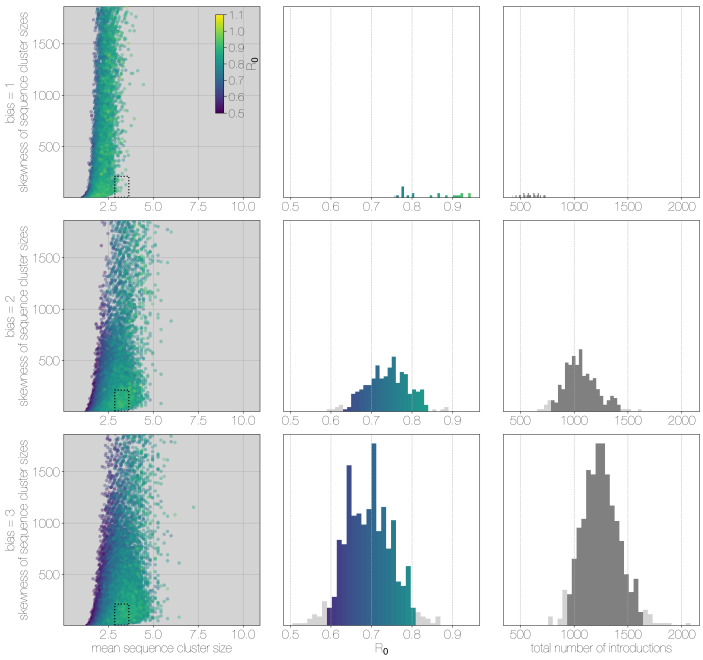


Figure 3–Figure supplement 3. Results of Monte Carlo simulations with vast underestimation of cases. The plot is identical to Figure 3-Figure supplement 1, but instead of 2000 cases, simulations were run with 4000 cases. With more unobserved cases the R_0 values matching observed MERS-CoV sequence clusters can only be smaller, with a corresponding increase in numbers of zoonotic transmissions. However, the numbers of simulations that match MERS-CoV data go down as well.

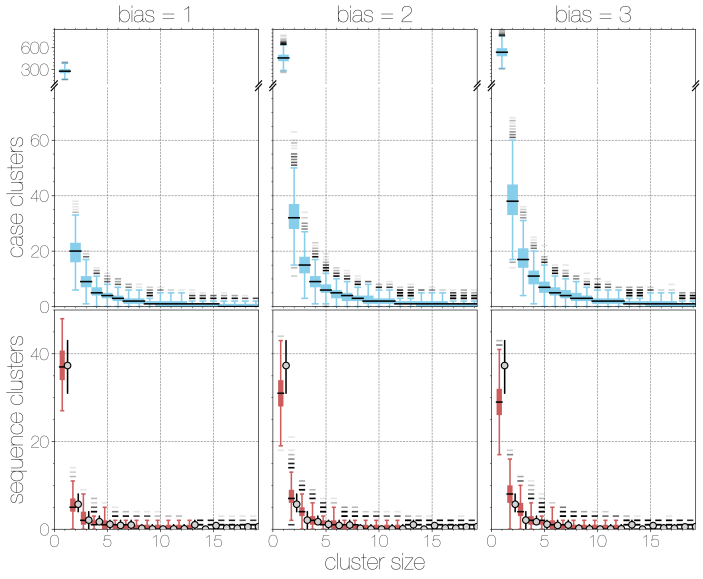


Figure 3–Figure supplement 4. Boxplots of matching simulated case and sequence cluster distributions. Boxplots indicate frequency of case (blue, top) and sequence (red, bottom) cluster sizes across simulations at different bias levels, marginalised across R_0 values. Outliers are shown with transparency, medians are indicated with thick black lines. Case clusters exhibit a strong skew with large numbers of singleton introductions and a substantial tail at higher levels of bias.

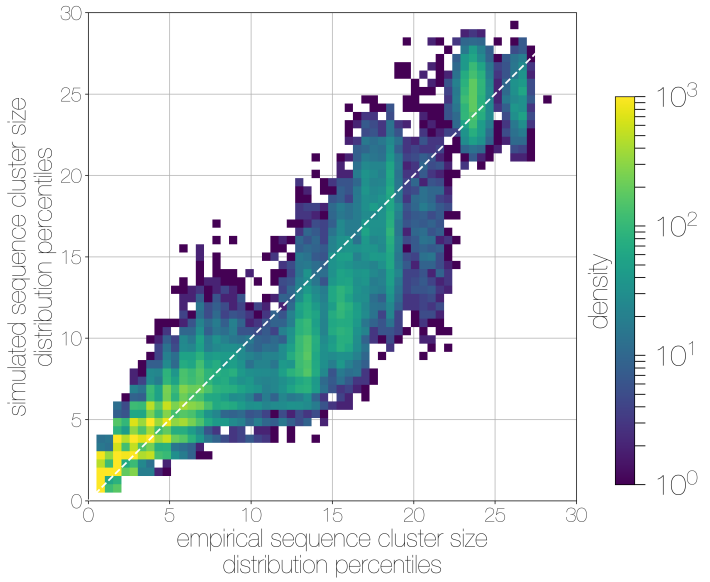


Figure 3–Figure supplement 5. Quantile-quantile (Q-Q) plot of empirical and simulated sequence cluster sizes. Density of sequence cluster size percentiles (1st–99th, calculated across a grid of 50 values) calculated for random states from the posterior distribution (x-axis) and matching simulations (y-axis). Most values fall on the one-to-one line, with a heavier tail in mid-sized sequence clusters in empirical data, manifesting as a greater density of points below the one-to-one line in the middle.

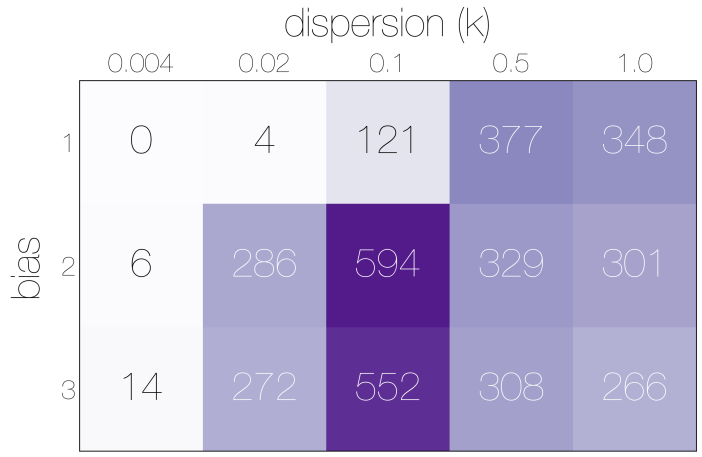


Figure 3–Figure supplement 6. Numbers of epidemiological simulations conforming to empirical observations. Numbers indicate the total number of epidemiological simulations under each combination of bias and dispersion parameter ω that result in MERS-CoV-like sequence cluster sizes. More simulations match observations with bias> 1 and $\omega \approx 0.1$.

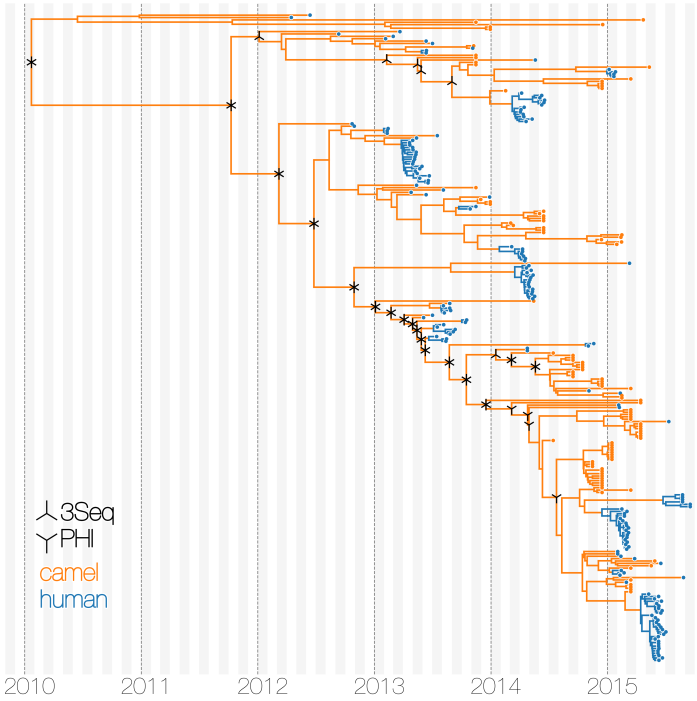


Figure 4–Figure supplement 1. Tests of recombination across MERS-CoV clades. Maximum clade credibility tree of MERS-CoV genomes annotated with results of two recombination detection tests (PHI and 3Seq) applied to descendent sequences of each clade. Both tests identify large portions of existing sequence data as containing signals of recombination. Note that markings do not indicate where recombinations have occurred on the tree, merely the minimum distance in sequence/time space between recombining lineages.

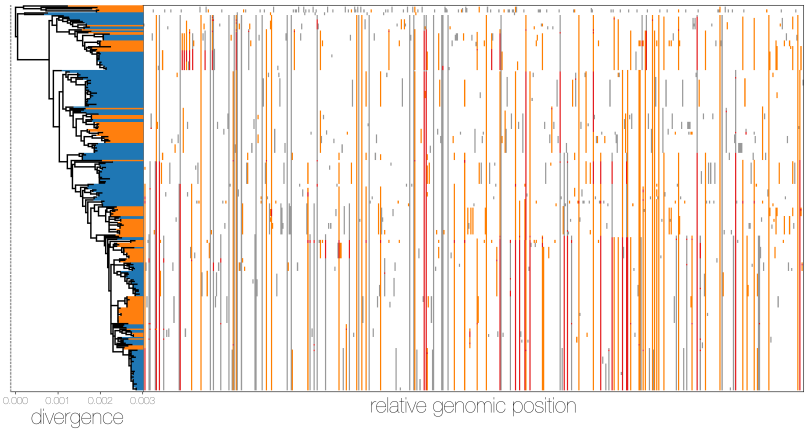


Figure 4–Figure supplement 2. MERS-CoV genomes exhibit high numbers of non-clonal loci. Ancestral state reconstruction (right) identifies a large number of sites in which mutations have occurred more than once in the tree (homoplasies, orange) or are reversions (red) from a state arising in an ancestor. Mutations that apparently only occur once in the tree (synapomorphies) are shown in grey. The maximum likelihood phylogeny on the left is coloured by whether sequences were sampled in humans (blue) or camels (orange).

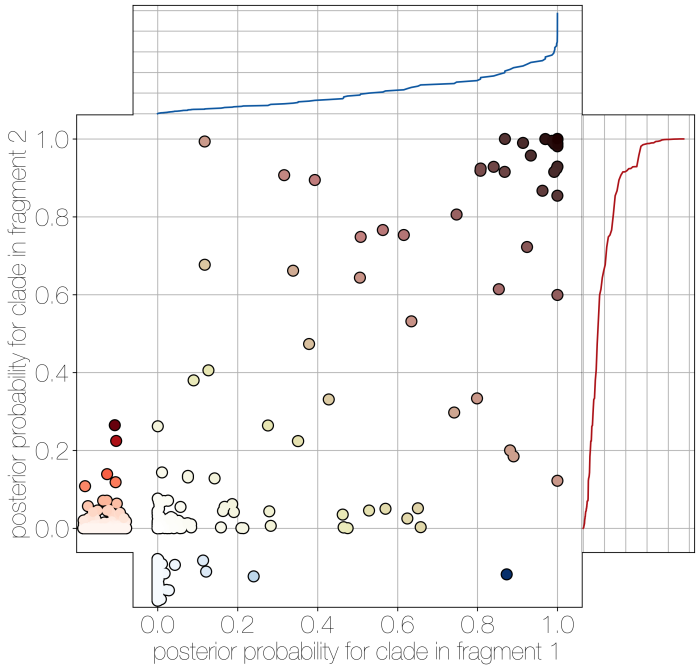


Figure 4–Figure supplement 3. Human clade sharing between genomic fragments 1 and 2. Central scatter plot shows the posterior probability of human clades shared between genomic fragments 1 and 2, in their respective trees. Left and bottom scatter plots track the posterior probability of human clades only observed in fragment 2 (left) or fragment 1 (bottom). The cumulative probability of human clades present in either tree are tracked by plots on the right (fragment 2) and top (fragment 1). Most of the probability mass is concentrated within human clades that are present in trees of both genomic fragment 1 and 2 (0.9701 and 0.9474 of all human clades across posteriors, respectively).

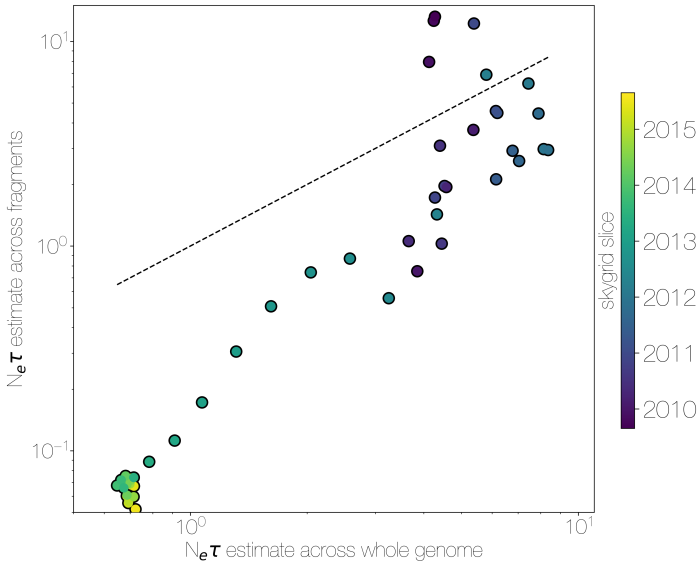


Figure 5–Figure supplement 1. Skygrid comparison between whole and fragmented genomes. Inferred median $N_e\tau$ recovered using a skygrid tree prior on whole genome (bottom) and ten genomic fragments with independent trees (left), coloured by time. Dotted line indicates the one-to-one line.