MERS-CoV spillover at the camel-human interface

Gytis Dudas[1], Luiz Max Carvalho[2], Andrew Rambaut[2,3,4] & Trevor Bedford[1]

[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, [4]Centre for Immunology, Infection and Evolution at the University of Edinburgh, Edinburgh, UK

August 6, 2017

Abstract

Middle East respiratory syndrome coronavirus (MERS-CoV) is a zoonotic virus originating in camels that has been causing significant mortality and morbidity in humans in the Arabian Peninsula. The epidemiology of the virus remains poorly understood, with hospital outbreaks, isolated cases with known exposure to camels and apparent community transmission occurring simultaneously. While traditional and seroepidemiological studies have been employed extensively throughout the epidemic, viral sequence data have not been utilised to their full potential in understanding transmission patterns within the outbreak. Here we use existing MERS-CoV sequence data to explore the phylodynamics of the virus 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 traditional 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. Genomic epidemiology, however, can fill this critical gap and has repeatedly shown the utility of viral sequence data in outbreak scenarios (Liu et al., 2013; Gire et al., 2014; Grubaugh et al., 2017), where data are relatively cheap to produce. These data can stand in for diagnostics and 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 traditional 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 distinct sampling patterns, population dynamics within demes 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. Once human MERS-CoV infections are established, however, we find that MERS-CoV is poor at transmitting between humans. 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 S1). 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 S2). 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 signicantly 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

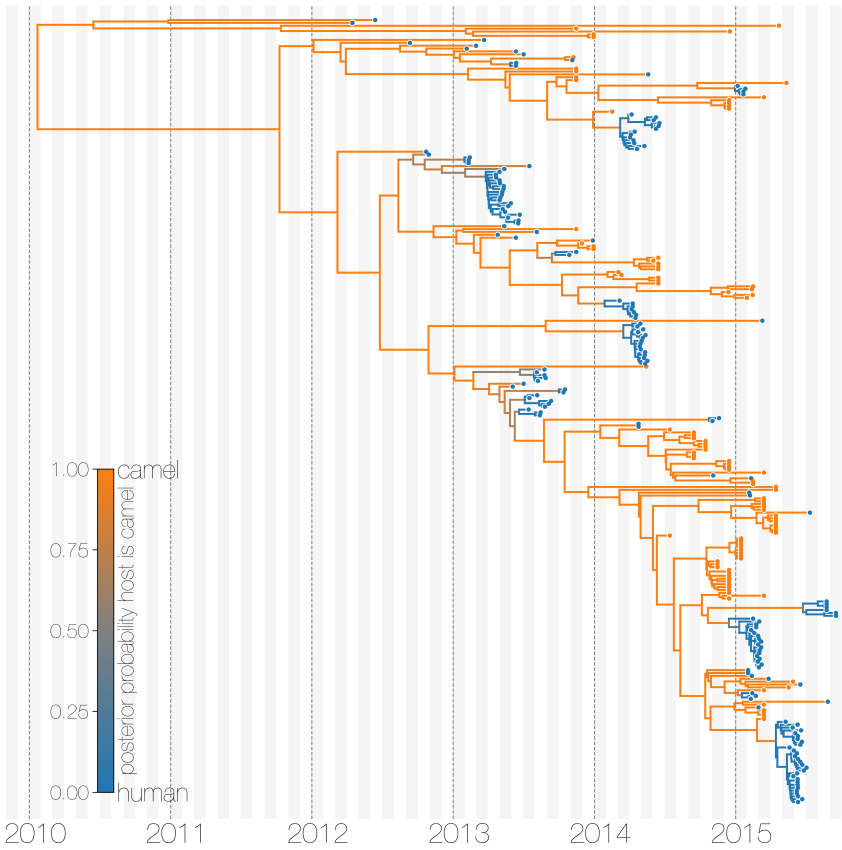

Figure 1. Typed maximum clade credibility tree of MERS-CoV genomes from 174 human viruses and 100 camel viruses. Maximum clade credibility (MCC) tree showing inferred ancestral hosts for MERS-CoV recovered with the structured coalescent. The vast majority of MERS-CoV evolution is inferred to occur in 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.

2013. Our finding of negligible human-to-camel transmission is robust to choice of prior (Figure S2).

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 S1). 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 analyses recover these results despite sequence data heavily skewed towards non-uniformly sampled human cases and are robust to choice of prior. This suggests 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.

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, which is more familiar to epidemiologists and provides insight into epidemic behavior. 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 take place at different levels of bias, wherein bias enriches sequencing of larger case clusters. 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 mean, median and standard deviation of sequence cluster sizes in each simulation (left-hand

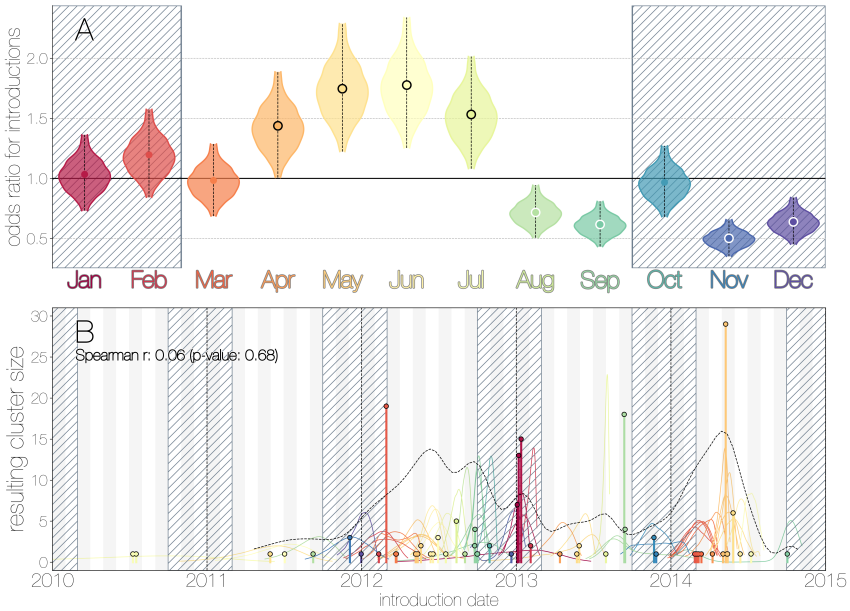

Figure 2. Seasonality of MERS-CoV introduction events. A) Posterior density estimates partitioned by month showing the 95% highest posterior density interval for relative odds ratios of MERS-CoV introductions into humans. Posterior means are indicated with circles. Evidence for increased or decreased 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 $\rho = 0.06$, $p = 0.68$).

panels 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 produre is shown in Figure S3. Generally, higher R_0 results in fewer larger transmission clusters, while lower R_0 results in many smaller transmission clusters (Figure 3).

We find that observed phylogenetic patterns of sequence clustering strongly support R_0 below 1.0 (middle panels in Figure 3). For increasing levels of bias mean R_0 values observed in matching simulations are 0.844, 0.730, and 0.683, respectively. While the 95% percentiles for R_0 values are close to 1.0 (0.720–0.995) 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 (Figure 3). Under unbiased sequencing only 0.6% of simulations fit our phylogenetic observations, while 2.7% and 2.7% of simulations fit for bias levels of 2.0

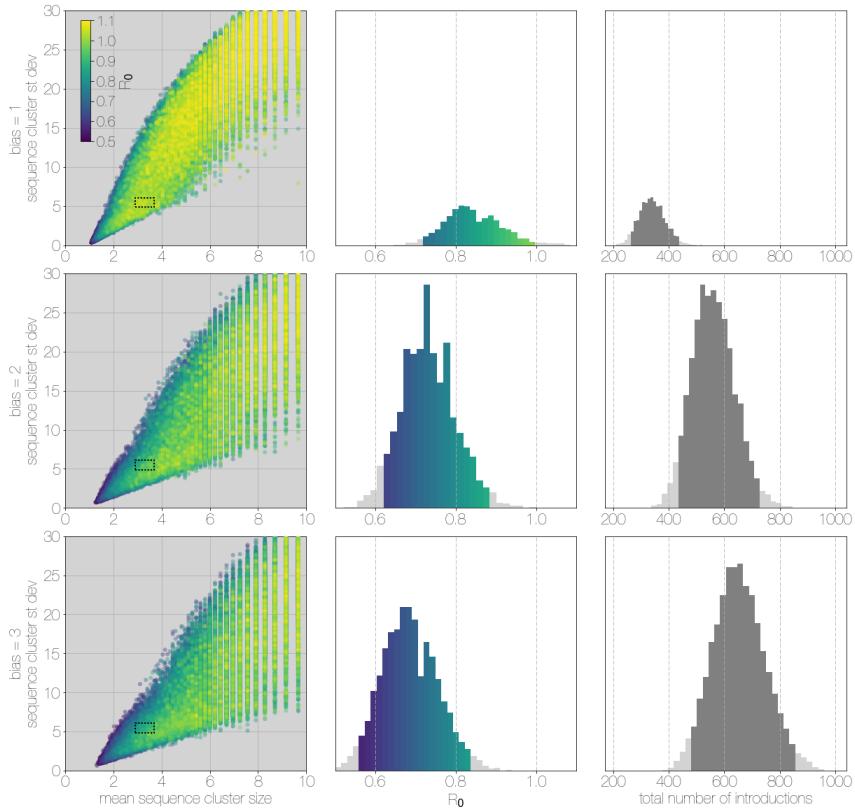

Figure 3. Monte Carlo simulations of human transmission clusters. Each row corresponds to a different bias value used to concentrate the subsampling of sequences from cases, and goes from 1 (no bias) to 2, and 3 (increasing levels of bias which make large case clusters to be more likely to be sequenced). Leftmost scatter plots show results of individual Monte Carlo simulations, 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 standard deviation observed for empirical MERS-CoV data. The distribution of R_0 values found within the dotted rectangle is shown in the middle, on the same y-axis across all levels of bias. 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 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. These Monte Carlo simulations indicate R_0 for MERS-CoV is likely to be below 1.0, with biased sequencing and numbers of zoonotic transmissions numbering in the hundreds.

and 3.0, respectively. Correspondingly, we estimate 10% support for a model with bias level 1.0, 45% support for a model with bias level 2.0, and 45% support for a model with bias level 3.0. Model averaging would suggest plausible R_0 values between 0.57 and 0.91.

Lower values for R_0 in turn suggest relatively large numbers of zoonotic transfers of viruses into humans (right-hand panels in Figure 3). The median number of cross-species introductions observed in simulations matching empirical data without bias are 339 (95% percentiles 255–431). These numbers jump up to 558 (95% percentiles 424–720) for bias = 2 and 650 (95% percentiles 481–850) for bias = 3 simulations, which as mentioned previously match considerably better to MERS phylogenetic data. Model averaging would suggest

plausible numbers of introductions between 299 and 818. 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 S4).

Overall, our analyses indicate that MERS-CoV is poorly suited for human-to-human transmission, with an estimated $R_0 < 0.91$ and sequence sampling likely to be biased towards large hospital outbreaks. 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 explore the role of recombination in shaping 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 spatial clustering of alleles along the genome, due to how template switching, the primary mechanism of recombination in RNA viruses, occurs. 3Seq relies on the spatial structure of nucleotide similarities 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 S5). Neither method can identify where in the tree recombination occurred, but each full asterisk in Figure S5 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 S6) with some evidence of spatial clustering of such alleles. Although the evolutionary centrality of camel viruses (Figure 1) may be sufficient to argue that camels are the host where MERS-CoV recombines, incidence of MERS-CoV is known to be much higher in camels (Müller et al., 2014; Corman et al., 2014;

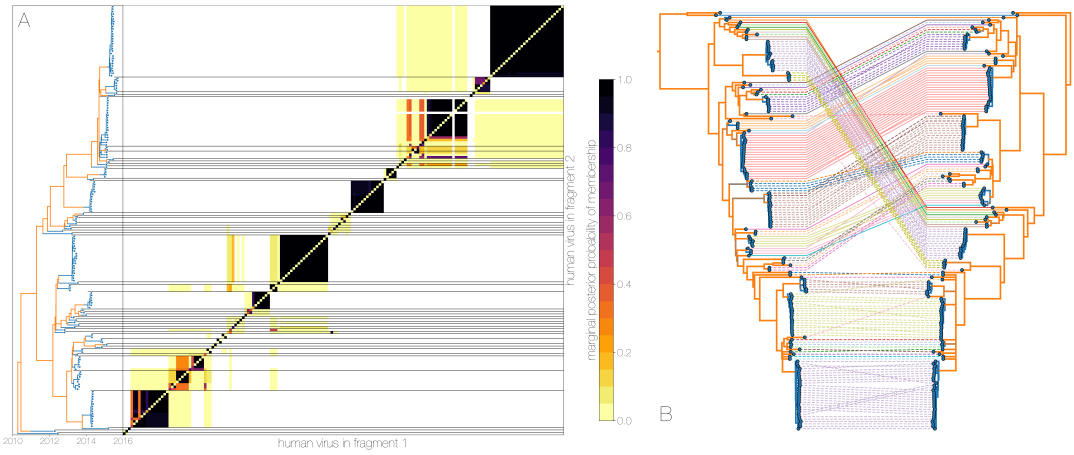

Figure 4. Recombinant features of MERS-CoV phylogenies. A) Heatmap showing the posterior probability that viruses from trees of different genomic fragments belong to the same introduction event. Genomic fragment 1 represents nucleotide positions 1 to 21000 and genomic fragment 2 represents nucleotide positions 21001 to 29364. The same set of viruses are arrayed on the x-axis and on the y-axis; the x-axis shows identity of these viruses along genomic fragment 1 and the y-axis shows identity of these viruses along genomic fragment 2. Viruses are ordered by their appearance in tree of genomic fragment 2 reduced to just the human tips and coloured by inferred host (blue for human, orange for camel) on the left. Human clusters are largely well-supported as monophyletic and consistent across trees of both genomic fragments. B) Phylogenies derived from genomic fragment 1 (left) and genomic fragment 2 (right), reduced to just the human tips. Taxa descended from the same introduction in fragment 1 are connected to their counterparts in fragment 2 with combinations of line dotting and colour unique to each introduction clade. Branches are coloured by inferred ancestral host state (human in blue, camel in orange). Human clusters in blue change phylogenetic positions between the trees together, with little incongruence within clusters. This is evidence for recombinant viruses generated in the camel reservoir entering human populations.

Chu et al., 2014; Reusken et al., 2014; Ali et al., 2017). This provides ideal conditions for co-infection with distinct genotypes, which is a pre-requisite for detectable RNA virus recombination to occur.

Conversely, our results strongly suggest that co-infection of humans with distinct lineages of MERS-CoV should be exceedingly rare. We find little evidence that recombination will interfere with the inference of human outbreak clusters (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 S7). Critically, 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. MERS-CoV evolution on the reservoir side, though complicated by recombination, is nonetheless still amenable to phylogenetic methods, in part through limited diversity of the virus in camels (see next section). In humans MERS-CoV evolution should be far easier to track as the only detectable and problematic recombinants are more likely to arise within the transmission chain, than through human co-infection with distinct MERS-CoV lineages.

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, S8). However, we do note that estimates of relative genetic diversity are relatively low overall with a mean estimate of $N_e\tau$ of 3.49 years (95% HPD: 2.71–4.38), and consequently MERS-CoV phylogeny resembles a ladder often seen in human influenza A virus phylogenies (Bedford et al., 2011). This estimate of $N_e\tau$ can be translated into an estimate of effective prevalence N_e given an estimate of generation time or serial intervals. Given a 20 day duration of infection in

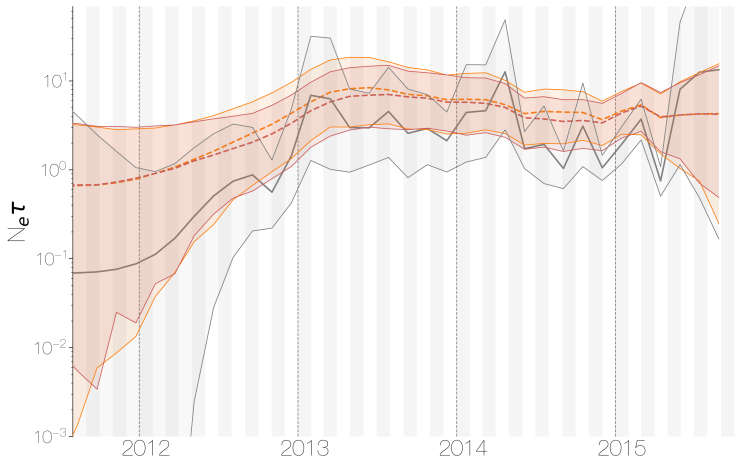

Figure 5. Demographic history of MERS-CoV in Arabian peninsula camels. Demographic history of MERS-CoV in camels, as inferred via a skygrid coalescent tree prior (Gill et al., 2013). Three skygrid reconstructions are shown, red and orange for each of the stationary distributions reached by MCMC with 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 $N_e\tau$.

camels (Adney et al., 2014), we expect a 10 day serial interval or τ of 0.027 years per generation. Dividing by τ yields an estimate of N_e of 129 (100–162) individuals.

Here, we compare this estimate of effective prevalence to a rough estimate of census prevalence. 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 10 day serial intervals gives an average of $N = 5042$ infections.

We thus arrive a ratio of $N/N_e = 39.1$ (31.1–50.4). 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, S1). This inference succeeds under different choices of priors for unknown demographic parameters (Figure S2) and in the presence of strong biases in sequence sampling schemes (Figure 3). 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). 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).

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 a serious global threat, we would like to highlight that its epidemiology is nonetheless concerning.

MERS-CoV may join the list of pathogens able to jump species barriers but not spread efficiently in the new host, but every system is distinct. Pathogens such as *Bacillus anthracis*, Andes hantavirus (Martinez et al., 2005), monkeypox (Reed et al., 2004), triple reassortant and H3N2v influenza A viruses (Shinde et al., 2009; Epperson et al., 2013) belong to this list and yet only triple reassortant viruses eventually contributed to a pandemic, due to epidemiology and phylodynamics confined to influenza A virus. When information about a system is lacking either because it is recent or entirely novel or if the system is inherently unpredictable, sequence data can and should be used jointly with whatever data sources are available to provide the unique pathogen perspective on an infectious disease outbreak. Although MERS-CoV seems to cause self-limiting outbreaks in humans with no evidence of worsening outbreaks over time, sustained evolution of the virus in the reservoir and flow of viral lineages into humans provides the substrate for a more transmissible variant of MERS-CoV to possibly emerge. We encourage continued genomic surveillance of MERS-CoV in the camel reservoir and from sporadic human cases.

Methods

Sequence data

All MERS-CoV sequences were downloaded from GenBank and accession numbers are given in Table S1. 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. Sequences were annotated with available collection dates and hosts, designated as camel or human and 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. The final dataset consisted of 174 genomes from human infections and 100 genomes from camel infections (Table S1).

Structured coalescent analyses

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. This model 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.

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.

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.

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 variable $Y_{ij} = \sum_{k=1}^{L} Z_{ijk}$ with the hierarchical model:

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

Odds ratios of introductions can then be computed for each month i as $\text{OR}_i = \exp(\beta_i)$.

Epidemiological analyses

As of 10 May 2017, the World Health Organization has been notified of 1952 cases of MERS-CoV. We thus simulate final transmission chain sizes using equation 1 (Lloyd-Smith et al., 2005; Blumberg and Lloyd-Smith, 2013) until we reach an epidemic comprised of 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. Each simulation results in a vector of outbreak sizes C, where C_i is the size of the ith transmission cluster and $\sum_{i=1}^{K} C_i = 2000$, where K is the number of clusters generated. We sample from the case cluster size vector according to a multivariate hypergeometric distribution to simulate sequencing (algorithm 1). The resulting sequence cluster size vector 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 MERS-CoV sequences used in this study. Since the sampling scheme operates under equi-probable 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. This makes clusters likely to be 'sequenced' even

more likely to be sequenced and *vice versa*. 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.

Let R_0 be the basic reproductive number ($0 < R_0 < 1$), and $\omega > 0$ be the a dispersion parameter, then the probability of observing a stuttering chain (cluster) r of size j is (Blumberg and Lloyd-Smith, 2013)

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

Although case clusters and their sizes are difficult to infer directly and require detailed epidemiological follow up, sequence data have fewer limitations.Our structured coalescent analyses indicate that every MERS outbreak is a contained cross-species transmission 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 distribution of case clusters. We employ a Monte Carlo simulation approach to identify simulations where the recovered distribution of sequence cluster sizes fall within the 95% highest posterior density intervals for three summary statistics of MERS-CoV sequence cluster sizes recovered via structured coalescent analyses: mean, median and standard deviation. These values are 2.90–3.70 for mean sequence cluster size, 4.87–6.07 for standard deviation of sequence cluster sizes and a median sequence cluster size of 1.

Data: Array of case cluster sizes in outbreak $\boldsymbol{C} = [C_1, C_2, \ldots, C_K]$, sequences available M, total outbreak size $\sum_{i=1}^{K} C_i$.
Result: Array of sequence cluster sizes sampled: $\boldsymbol{S} = [S_1, S_2, \ldots, S_K]$.
Draw S_i from a hypergeometric distribution with C_i successes, $\sum_{i=1}^{K} C_i - C_i$ 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,
  $\boldsymbol{p} = [p(0)^{\text{bias}}, p(1)^{\text{bias}}, \ldots, p(C_i)^{\text{bias}}] \times (\sum_i p_i^{\text{bias}})^{-1}$, where $p(\cdot)$ is the pmf for a hypergeometric distribution with C_i successes, $\sum_{i=1}^{K} C_i - C_i$ 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 used to represent a random draw from the diversity of MERS-CoV circulating in camels. These sequences 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 the Arabian Peninsula and Egypt are distinct epidemics with relatively poor sampling in the latter. 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. 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.

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. 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. 2013. Typology of camel farming system in Saudi Arabia. *Emirates Journal of Food and Agriculture*. 25:250.

Adney DR, van Doremalen N, Brown VR, Bushmaker T, Scott D, de Wit E, Bowen RA, Munster VJ. 2014. Replication and Shedding of MERS-CoV in Upper Respiratory Tract of Inoculated Dromedary Camels. *Emerging Infectious Diseases*. 20:1999–2005.

Ali MA, Shehata MM, Gomaa MR, et al. (18 co-authors). 2017. Systematic, active surveillance for Middle East respiratory syndrome coronavirus in camels in Egypt. *Emerg Microbes Infect*. 6:e1.

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

Assiri A, Al-Tawfiq JA, Al-Rabeeah AA, et al. (13 co-authors). 2013a. Epidemiological, demographic, and clinical characteristics of 47 cases of Middle East respiratory syndrome coronavirus disease from Saudi Arabia: a descriptive study. *Lancet Infect Dis*. 13:752–761.

Assiri A, McGeer A, Perl TM, et al. (18 co-authors). 2013b. Hospital outbreak of Middle East respiratory syndrome coronavirus. *N Engl J Med*. 369:407–416.

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

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

Boheemen Sv, Graaf Md, Lauber C, et al. (11 co-authors). 2012. Genomic characterization

of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *mBio*. 3:e00473–12.

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

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

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

Cauchemez S, Nouvellet P, Cori A, et al. (25 co-authors). 2016. Unraveling the drivers of MERS-CoV transmission. *Proc Natl Acad Sci USA*. 113:9081–9086.

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

Chu DK, Poon LL, Gomaa MM, et al. (13 co-authors). 2014. MERS coronaviruses in dromedary camels, Egypt. *Emerg Infect Dis*. 20:1049–1053.

Corman VM, Jores J, Meyer B, et al. (13 co-authors). 2014. Antibodies against MERS coronavirus in dromedary camels, Kenya, 1992–2013. *Emerg Infect Dis*. 20:1319–1322.

Cotten M, Watson SJ, Kellam P, et al. (22 co-authors). 2013. Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet*. 382:1993–2002.

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

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

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

Dudas G, Carvalho LM, Bedford T, et al. (96 co-authors). 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. 544:309–315.

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

Epperson S, Jhung M, Richards S, et al. (21 co-authors). 2013. Human infections with influenza A(H3N2) variant virus in the United States, 2011–2012. *Clin Infect Dis*. 57:S4–S11.

Fagbo SF, Skakni L, Chu DKW, Garbati MA, Joseph M, Hakawi AM. 2015. Molecular epidemiology of hospital outbreak of Middle East respiratory syndrome, Riyadh, Saudi Arabia, 2014. *Emerg Infect Dis*. 21:1981.

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

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. 2013. Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Mol Biol Evol*. 30:713.

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

Grubaugh ND, Ladner JT, Kraemer MU, et al. (67 co-authors). 2017. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*. 546:401–405.

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

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

Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FCC. 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. *J Virol*. 82:1819–1826.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. 30:772–780.

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

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

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

Liu D, Shi W, Shi Y, et al. (11 co-authors). 2013. Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. *Lancet*. 381:1926–1932.

Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature*. 438:355–359.

Lycett S, Bodewes R, Pohlmann A, et al. (11 co-authors). 2016. Role for migratory wild birds in the global spread of avian influenza H5N8. *Science*. 354:213–217.

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

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

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

Memish ZA, Cotten M, Meyer B, et al. (17 co-authors). 2014. Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. *Emerg Infect Dis*. 20:1012.

Müller MA, Corman VM, Jores J, et al. (12 co-authors). 2014. MERS coronavirus neutralizing antibodies in camels, Eastern Africa, 1983–1997. *Emerg Infect Dis*. 20.

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

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

Park SS, Wernery U, Corman VM, et al. (19 co-authors). 2015. Acute Middle East respiratory syndrome coronavirus infection in livestock dromedaries, Dubai, 2014. *Emerg Infect Dis*. 21:1019.

Pond K, L S, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics*. 22:3096–3098.

Quick J, Ashton P, Calus S, et al. (18 co-authors). 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biology*. 16:114.

Reed KD, Melski JW, Graham MB, et al. (19 co-authors). 2004. The detection of monkeypox in humans in the Western Hemisphere. *N Engl J Med*. 350:342–350.

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

Reusken CB, Messadi L, Feyisa A, et al. (17 co-authors). 2014. Geographic distribution of MERS coronavirus among dromedary camels, Africa. *Emerg Infect Dis*. 20:1370–1374.

Reusken CBEM, Farag EABA, Haagmans BL, et al. (21 co-authors). 2015. Occupational exposure to dromedaries and risk for MERS-CoV infection, Qatar, 2013–2014. *Emerg Infect Dis*. 21:1422.

Shinde V, Bridges CB, Uyeki TM, et al. (23 co-authors). 2009. Triple-reassortant swine influenza A (H1) in humans in the United States, 2005–2009. *N Engl J Med*. 360:2616–2625.

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

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

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

Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. 2014. Efficient Bayesian inference under the structured coalescent. *Bioinformatics*. 30:2272–2279.

Volz EM, Koelle K, Bedford T. 2013. Viral phylodynamics. *PLoS Comput Biol*. 9:e1002947.

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

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

World Health Organization. 2017. Middle East respiratory syndrome coronavirus (MERS-CoV). Available at http://www.who.int/mediacentre/factsheets/mers-cov/en/.

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

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med*. 367:1814–1820.

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

Table S1. Strain names, accessions (where available), identified host and reported collection dates for MERS-CoV genomes used in this study.

	strain	accession	host	collection date
1	KSA-378	KJ713296	camel	2013-11
2	KSA-363	KJ713298	camel	2013-11
3	KSA-503	KJ713297	camel	2013-11
4	KSA-376	KJ713299	camel	2013-11
5	KSA-505	KJ713295	camel	2013-11
6	Jeddah-1	KF917527	camel	2013-11-08
7	NRCE-HKU205	KJ477102	camel	2013-11-15
8	KFU-HKU1	KJ650297	camel	2013-11-30
9	KFU-HKU13	KJ650295	camel	2013-12-30
10	Camel_Egypt_NRCE-HKU271		camel	2013-12-30
11	Camel_Egypt_NRCE-HKU270		camel	2013-12-30
12	KFU-HKU19Dam	KJ650296	camel	2013-12-30
13	Qatar_2_2014	KJ650098	camel	2014-02-16
14	UAE/D469-14	KU242424	camel	2014-03-04
15	UAE/D511-14	KU242423	camel	2014-03-12
16	Jeddah/F13A/2014	KT368824	camel	2014-05
17	UAE/D1164.10/2014	KP719928	camel	2014-06
18	UAE/D1339.2/2014	KP719931	camel	2014-06
19	UAE/D1164.11/2014	KP719929	camel	2014-06
20	UAE/D1164.9/2014	KP719927	camel	2014-06
21	UAE/D1209/2014	KP719933	camel	2014-06
22	UAE/D1164.14/2014	KP719930	camel	2014-06
23	UAE/D1243.12/2014	KP719932	camel	2014-06
24	D1164.1/14	KX108937	camel	2014-06-02
25	Riyadh/Ry23N/2014	KT368825	camel	2014-07
26	Riyadh/Ry84N/2014	KT368826	camel	2014-07
27	Jeddah/S93/2014	KT368855	camel	2014-09
28	Jeddah/401/2014	KT368827	camel	2014-09
29	Jeddah/S100/2014	KT368853	camel	2014-09
30	Jeddah/S99/2014	KT368857	camel	2014-09
31	Jeddah/S94/2014	KT368856	camel	2014-09
32	Jeddah/S73/2014	KT368854	camel	2014-09
33	Jeddah/O47b/2014	KT368852	camel	2014-10
34	Jeddah/O23b/2014	KT368849	camel	2014-10
35	Jeddah/O24/2014	KT368850	camel	2014-10
36	Jeddah/O30/2014	KT368851	camel	2014-10
37	Jeddah/N51/2014	KT368846	camel	2014-11
38	Jeddah/N68b/2014	KT368848	camel	2014-11
39	Jeddah/N62b/2014	KT368847	camel	2014-11
40	Jeddah/D40/2014	KT368834	camel	2014-12
41	Jeddah/D90/2014	KT368844	camel	2014-12

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
42	Jeddah/D88/2014	KT368843	camel	2014-12
43	Jeddah/D36/2014	KT368832	camel	2014-12
44	Jeddah/D35/2014	KT368831	camel	2014-12
45	Jeddah/D92/2014	KT368845	camel	2014-12
46	Jeddah/D49/2014	KT368841	camel	2014-12
47	Jeddah/D34/2014	KT368830	camel	2014-12
48	Jeddah/D33b/2014	KT368829	camel	2014-12
49	Jeddah/D42/2014	KT368835	camel	2014-12
50	Jeddah/D50b/2014	KT368842	camel	2014-12
51	Jeddah/D45/2014	KT368837	camel	2014-12
52	Jeddah/D46b/2014	KT368838	camel	2014-12
53	Jeddah/D43b/2014	KT368836	camel	2014-12
54	Jeddah/D100/2014	KT368828	camel	2014-12
55	Jeddah/D47/2014	KT368839	camel	2014-12
56	Jeddah/D38b/2014	KT368833	camel	2014-12
57	Jeddah/D48/2014	KT368840	camel	2014-12
58	D2597.2/14	KX108938	camel	2014-12-13
59	Egypt_NRCE-NC163/2014	KU740200	camel	2014-12-17
60	Jeddah/Jd7/2015	KT368861	camel	2015-01
61	Jeddah/Jd86/2015	KT368863	camel	2015-01
62	Jeddah/Jd90/2015	KT368865	camel	2015-01
63	Jeddah/Jd1b/2015	KT368858	camel	2015-01
64	Jeddah/Jd4/2015	KT368859	camel	2015-01
65	Jeddah/Jd85/2015	KT368862	camel	2015-01
66	Jeddah/Jd6b/2015	KT368860	camel	2015-01
67	Jeddah/Jd87/2015	KT368864	camel	2015-01
68	D252/15	KX108939	camel	2015-01-30
69	Jeddah/Jd199/2015	KT368867	camel	2015-02
70	Jeddah/Jd175/2015	KT368866	camel	2015-02
71	D374/15	KX108940	camel	2015-02-12
72	D383/15	KX108941	camel	2015-02-14
73	D389/15	KX108942	camel	2015-02-15
74	Riyadh/Ry63/2015	KT368876	camel	2015-03
75	Riyadh/Ry136/2015	KT368868	camel	2015-03
76	Riyadh/Ry178/2015	KT368874	camel	2015-03
77	Riyadh/Ry162/2015	KT368871	camel	2015-03
78	Riyadh/Ry86/2015	KT368879	camel	2015-03
79	Taif/T150/2015	KT368889	camel	2015-03
80	Riyadh/Ry137/2015	KT368869	camel	2015-03
81	Riyadh/Ry179/2015	KT368875	camel	2015-03
82	Riyadh/Ry177/2015	KT368873	camel	2015-03
83	Riyadh/Ry79/2015	KT368878	camel	2015-03
84	Riyadh/Ry173/2015	KT368872	camel	2015-03

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
85	Taif/T157b/2015	KT368890	camel	2015-03
86	Riyadh/Ry159b/2015	KT368870	camel	2015-03
87	Riyadh/Ry64/2015	KT368877	camel	2015-03
88	Taif/T3/2015	KT368880	camel	2015-04
89	Taif/T16/2015	KT368882	camel	2015-04
90	Taif/T22/2015	KT368883	camel	2015-04
91	Taif/T92/2015	KT368887	camel	2015-04
92	Taif/T7/2015	KT368881	camel	2015-04
93	Taif/T91b/2015	KT368886	camel	2015-04
94	Taif/T68/2015	KT368884	camel	2015-04
95	Taif/T89/2015	KT368885	camel	2015-04
96	Taif/T98/2015	KT368888	camel	2015-04
97	D998/15	KX108943	camel	2015-04-23
98	D1157/15	KX108944	camel	2015-05-12
99	D1189.1/15	KX108946	camel	2015-05-18
100	D1271/15	KX108945	camel	2015-05-29
101	Jordan-N3/2012	KC776174	human	2012-04-15
102	EMC/2012	JX869059	human	2012-06-13
103	England/1/2012	KC164505	human	2012-09-11
104	Riyadh_1_2012	KF600612	human	2012-10-23
105	Riyadh_2_2012	KF600652	human	2012-10-30
106	Riyadh_3_2013	KF600613	human	2013-02-05
107	England/3/2013	KM210278	human	2013-02-10
108	England/2/2013	KM015348	human	2013-02-10
109	England/4/2013	KM210277	human	2013-02-13
110	Riyadh_4_2013	KJ156952	human	2013-03-01
111	Munich/AbuDhabi/2013	KF192507	human	2013-03-22
112	Al-Hasa_2_2013	KF186566	human	2013-04-21
113	Al-Hasa_3_2013	KF186565	human	2013-04-22
114	UAE-FRA1_1627-2013_BAL	KJ361500	human	2013-04-26
115	Al-Hasa_4_2013	KF186564	human	2013-05-01
116	Al-Hasa_7_2013	KF600623, KF600655	human	2013-05-01
117	Al-Hasa_8_2013	KF600618, KF600626, KF600635, KF600638	human	2013-05-01
118	Al-Hasa_25_2013	KJ156866	human	2013-05-02
119	Al-Hasa_11_2013	KF600629, KF600636, KF600646	human	2013-05-03
120	Al-Hasa_12_2013	KF600627	human	2013-05-07

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
121	Al-Hasa_14_2013	KF600615, KF600643	human	2013-05-08
122	Al-Hasa_1_2013	KF186567	human	2013-05-09
123	Al-Hasa_15_2013	KF600645	human	2013-05-11
124	Al-Hasa_16_2013	KF600644	human	2013-05-12
125	Buraidah_1_2013	KF600630	human	2013-05-13
126	Al-Hasa_23_2013	KJ156860, KJ156894, KJ156929, KJ156923, KJ156862	human	2013-05-13
127	Al-Hasa_17_2013	KF600647	human	2013-05-15
128	Al-Hasa_19_2013	KF600632	human	2013-05-23
129	Al-Hasa_18_2013	KF600651	human	2013-05-23
130	Al-Hasa_21_2013	KF600634	human	2013-05-30
131	Hafr-Al-Batin_1_2013	KF600628	human	2013-06-04
132	Wadi-Ad-Dawasir_1_2013	KJ156881	human	2013-06-12
133	Taif_1_2013	KJ156949	human	2013-06-12
134	Taif_2_2013	KJ156896, KJ156876	human	2013-06-12
135	Taif_3_2013	KJ156938, KJ156897, KJ156922, KJ156868, KJ156921, KJ156915, KJ156906	human	2013-06-13
136	Al-Hasa_26_2013	KJ156882, KJ156941, KJ156872	human	2013-06-18
137	Al-Hasa_27_2013	KJ156943, KJ156939	human	2013-06-19
138	Al-Hasa_28_2013	KJ156887, KJ156940, KJ156889, KJ156893, KJ156884, KJ156930, KJ156928, KJ156909	human	2013-06-22

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
139	Riyadh_6_2013	KJ156879, KJ156947, KJ156890, KJ156908, KJ156927	human	2013-07-02
140	Riyadh_5_2013	KJ156944	human	2013-07-02
141	Riyadh_7_2013	KJ156937, KJ156905	human	2013-07-15
142	Riyadh_8_2013	KJ156880, KJ156942	human	2013-07-17
143	Riyadh_9_2013	KJ156869	human	2013-07-17
144	Hafr-Al-Batin_2_2013	KJ156910	human	2013-08-05
145	Asir_2_2013	KJ156863, KJ156899, KJ156912, KJ156900, KJ156898, KJ156945, KJ156932	human	2013-08-05
146	Riyadh_11_2013	KJ156946, KJ156911	human	2013-08-06
147	Riyadh_12_2013	KJ156926, KJ156901	human	2013-08-08
148	Riyadh_13_2013	KJ156888, KJ156873	human	2013-08-13
149	Riyadh_14_2013	KJ156934	human	2013-08-15
150	Hafr-Al-Batin_4_2013	KJ156931, KJ156895, KJ156864, KJ156861	human	2013-08-25
151	Hafr-Al-Batin_5_2013	KJ156951, KJ156924, KJ156954, KJ156913	human	2013-08-25
152	Riyadh_17_2013	KJ156918, KJ156920, KJ156865	human	2013-08-26
153	Hafr-Al-Batin_6_2013	KJ156874	human	2013-08-28
154	Riyadh_10_2013	KJ156891, KJ156936, KJ156907	human	2013-09-05
155	Madinah_3b_2013	KJ156950, KJ156916	human	2013-09-11

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
156	Qatar3	KF961221	human	2013-10-13
157	Qatar4	KF961222	human	2013-10-17
158	Oman_2285_2013	KT156560	human	2013-10-28
159	Jeddah-1	KF958702	human	2013-11-05
160	AbuDhabi_UAE_9_2013	KP209312	human	2013-11-15
161	Oman_2874_2013	KT156561	human	2013-12-28
162	AbuDhabi/Gayathi_UAE_2_2014	KP209310	human	2014-03-07
163	Jeddah_C7569/KSA	KM027256	human	2014-04-03
164	Jeddah_C7149/KSA	KM027255	human	2014-04-05
165	Jeddah_C7770/KSA	KM027257	human	2014-04-07
166	AbuDhabi_UAE_8_2014	KP209306	human	2014-04-07
167	AbuDhabi_UAE_16_2014	KP209308	human	2014-04-10
168	AbuDhabi_UAE_18_2014	KP209307	human	2014-04-10
169	Jeddah_C8826/KSA	KM027258	human	2014-04-12
170	AbuDhabi_UAE_26_2014	KP209313	human	2014-04-13
171	Jeddah_C9055/KSA	KM027259	human	2014-04-14
172	Makkah_C9355/KSA/Makkah	KM027261	human	2014-04-15
173	AbuDhabi_UAE_33_2014	KP209311	human	2014-04-17
174	AbuDhabi_UAE_30_2014	KP209309	human	2014-04-19
175	Jeddah_C10306/KSA	KM027260	human	2014-04-21
176	Riyadh_2014KSA_683/KSA/2014	KM027262	human	2014-04-22
177	Riyadh-KKUH-90b		human	2014-04-24
178	Riyadh-KKUH-105		human	2014-04-25
179	Riyadh-KKUH-104		human	2014-04-25
180	KFMC-1	KT121580	human	2014-04-28
181	KFMC-8	KT121579	human	2014-04-30
182	Indiana/USA-1_SaudiArabia_2014	KJ813439	human	2014-04-30
183	KFMC-10	KT121578	human	2014-05-01
184	KFMC-7	KT121581	human	2014-05-03
185	Riyadh-KKUH-291		human	2014-05-06
186	KFMC-9	KT121574	human	2014-05-07
187	KFMC-3	KT121573	human	2014-05-09
188	Florida/USA-2_SaudiArabia_2014	KJ829365	human	2014-05-10
189	KFMC-2	KT121577	human	2014-05-11
190	KFMC-4	KT121575	human	2014-05-12
191	KFMC-5	KT121572	human	2014-05-12
192	Riyadh-KKUH-368		human	2014-05-13
193	KFMC-6	KT121576	human	2014-05-18
194	Riyadh_2014KSA_158/KSA/2014	KM027281	human	2014-05-20
195	Jeddah-KFH-285TA		human	2014-06-03
196	Jeddah-KFH-605TD		human	2014-06-09
197	Jeddah-KFH-668TD		human	2014-06-09
198	Jeddah-KFH-899NF		human	2014-06-16

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
199	Jeddah-KFH-949NSG1		human	2014-06-18
200	Riyadh-KKUH-643		human	2014-11-02
201	Taif/KSA-7032/2014	KU710264	human	2014-11-04
202	Riyadh-KKUH-665		human	2014-11-19
203	Riyadh-KSA-2049/2015	KR011266	human	2015-01-06
204	Riyadh-KSA-2343/2015	KR011264	human	2015-01-21
205	Riyadh-KSA-2345/2015	KR011263	human	2015-01-21
206	Riyadh-KSA-2466/2015	KR011265	human	2015-01-26
207	Kharj-KSA-2599/2015	KT806052	human	2015-02-02
208	Kharj-KSA-2598/2015	KT806053	human	2015-02-02
209	Riyadh-KSA-2716/2015	KT806051	human	2015-02-05
210	Khobar-KSA-6736/2015	KT806048	human	2015-02-07
211	Jeddah-KSA-C20843/2015	KT806044	human	2015-02-09
212	Jeddah-KSA-C20860/2015	KT806055	human	2015-02-10
213	Riyadh_KSA_2959_2015	KT026453	human	2015-02-10
214	Riyadh-KSA-3065/2015	KT806050	human	2015-02-12
215	Najran-KSA-C20915/2015	KT806054	human	2015-02-13
216	Riyadh-KSA-3181/2015	KT806049	human	2015-02-15
217	Riyadh_KKUH_0734		human	2015-02-18
218	Jeddah-KSA-C21271/2015	KT806045	human	2015-02-22
219	Riyadh_KKUH_0755		human	2015-02-23
220	Riyadh_KKUH_0756		human	2015-02-23
221	Riyadh_KKUH_0780		human	2015-02-25
222	Riyadh_KKUH_0801		human	2015-02-27
223	Riyadh_KKUH_0826		human	2015-02-28
224	Riyadh_KKUH_0818		human	2015-02-28
225	Riyadh_KSA_4050_2015	KT026454	human	2015-03-01
226	Riyadh_KKUH_0944		human	2015-03-02
227	Riyadh_KKUH_0939		human	2015-03-02
228	Riyadh_KKUH_1080		human	2015-03-03
229	Riyadh_KKUH_1066		human	2015-03-03
230	Riyadh_KKUH_1145		human	2015-03-04
231	Riyadh_KKUH_1217		human	2015-03-04
232	Riyadh_KKUH_1461		human	2015-03-08
233	Riyadh_KKUH_1470		human	2015-03-08
234	Riyadh_KKUH_1522		human	2015-03-09
235	Germany3/UAE-Dubai/Abu-Dhabi		human	2015-03-11
236	Hufuf-KSA-9158/2015	KT806047	human	2015-03-27
237	Hufuf-KSA-11002/2015	KT806046	human	2015-05-10
238	KOR/KNIH/002_05_2015	KT029139	human	2015-05-20
239	ChinaGD01	KT036372	human	2015-05-28
240	KOR/Seoul/014-2015	KX034093	human	2015-05-30
241	KOREA/Seoul/014-1-2015	KT374052	human	2015-05-31

Continued on next page

Table S1 – continued from previous page

	strain	accession	host	collection date
242	KOREA/Seoul/035-1-2015	KT374054	human	2015-06-03
243	KOR/Seoul/066-2015	KX034095	human	2015-06-04
244	Korea/Seoul/SNU1-035/2015	KU308549	human	2015-06-08
245	KOR/CNUH_SNU/030_06_2015	KT868868	human	2015-06-08
246	KOR/CNUH_SNU/024_06_2015	KT868867	human	2015-06-08
247	KOR/CNUH_SNU/054_06_2015	KT868871	human	2015-06-09
248	KOR/CNUH_SNU/038_06_2015	KT868870	human	2015-06-10
249	KOR/CNUH_SNU/148_06_2015	KT868876	human	2015-06-10
250	KOR/CNUH_SNU/122_06_2015	KT868875	human	2015-06-10
251	KOR/CNUH_SNU/082_06_2015	KT868872	human	2015-06-10
252	KOR/CNUH_SNU/085_06_2015	KT868873	human	2015-06-10
253	KOR/CNUH_SNU/016_06_2015	KT868865	human	2015-06-11
254	KOR/CNUH_SNU/023_06_2015	KT868866	human	2015-06-11
255	KOR/CNUH_SNU/031_06_2015	KT868869	human	2015-06-11
256	KOR/CNUH_SNU/110_06_2015	KT868874	human	2015-06-11
257	KOR/Seoul/050-1-2015	KX034094	human	2015-06-11
258	THA/CU/17_06_2015	KT225476	human	2015-06-17
259	KOR/Seoul/077-2-2015	KX034096	human	2015-06-17
260	KOR/Seoul/080-3-2015	KX034097	human	2015-06-17
261	KOREA/Seoul/163-1-2015	KT374051	human	2015-06-19
262	KOREA/Seoul/168-1-2015	KT374056	human	2015-06-21
263	KOR/Seoul/162-1-2015	KX034098	human	2015-06-22
264	KOR/CNUH_SNU/172_06_2015	KT868877	human	2015-06-22
265	KOR/Seoul/169-2015	KX034099	human	2015-06-26
266	KOR/Seoul/177-3-2015	KX034100	human	2015-07-03
267	Jeddah-KSA-3RS2702/2015	KU851859	human	2015-07-12
268	Riyadh-KSA-16120/2015	KU851861	human	2015-08-24
269	Riyadh-KSA-16117/2015	KU851862	human	2015-08-24
270	Riyadh-KSA-16121/2015	KU851860	human	2015-08-24
271	Riyadh-KSA-16098/2015	KU851864	human	2015-08-24
272	Riyadh-KSA-16077/2015	KU851863	human	2015-08-27
273	Jordan_1_2015	KU233363	human	2015-09-17
274	Jordan_10_2015	KU233362	human	2015-09-17

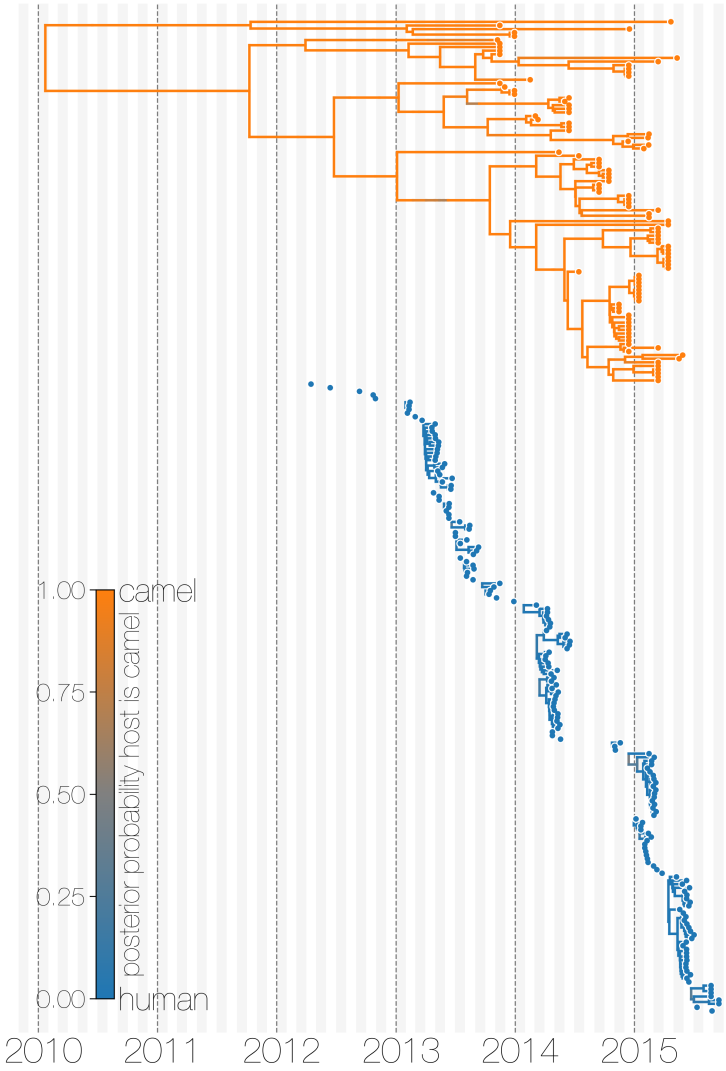

Figure S1. Evolutionary history of MERS-CoV partitioned between camels and humans. This is the same tree as shown in Figure 1, but with contiguous stretches of MERS-CoV evolutionary history split by inferred host: camels (top in orange) and humans (bottom in blue). This visualisation highlights the ephemeral nature of MERS-CoV outbreaks in humans, compared to continuous circulation of the virus in camels.

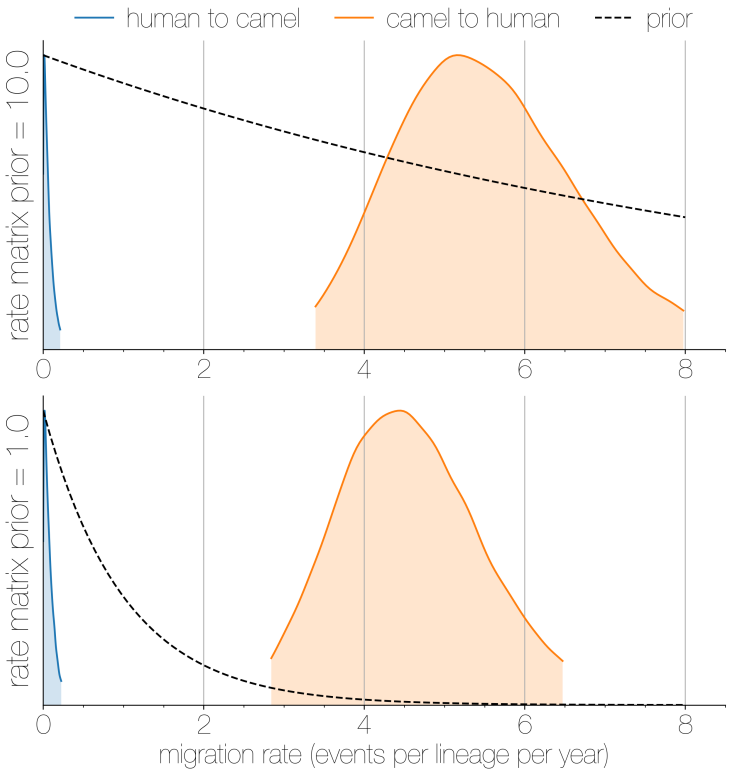

Figure S2. Posterior migration rate estimates for two choices of prior. Negligible flow of MERS-CoV lineages from humans into camels is recovered regardless of prior choice. Plots show the 95% highest posterior density for the estimated migration rate from the human deme into the camel deme (blue) and *vice versa* (orange). Dotted lines indicate exponential priors specified for migration rates, with mean 1.0 (bottom) or 10.0 (top).

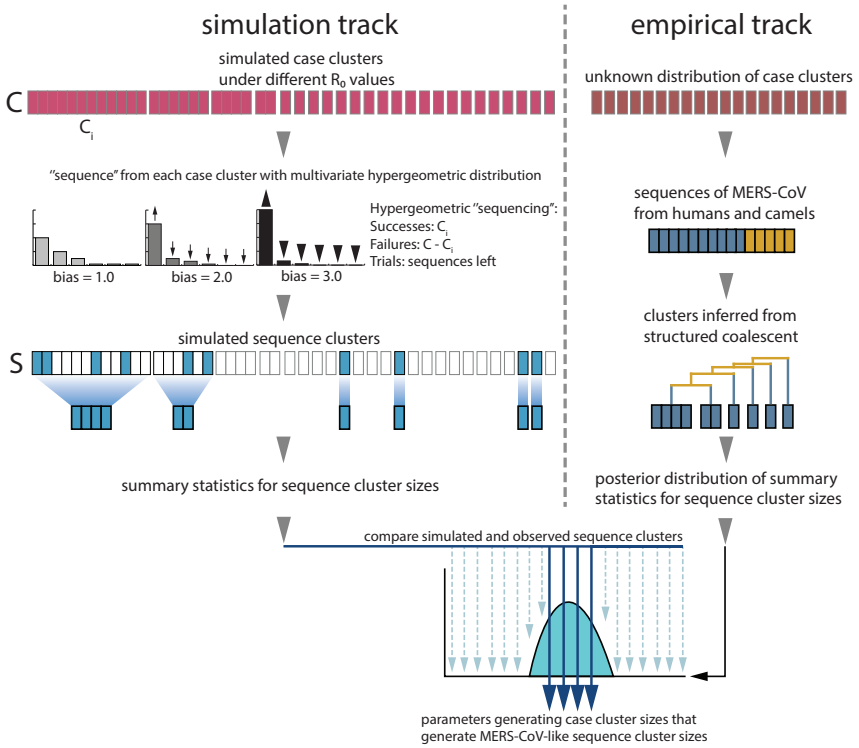

Figure S3. Monte Carlo simulation schematic. Case clusters are simulated according to Equation 1 until an outbreak size of 2000 cases is reached. We sample 174 cases from each simulation to represent sequencing of human MERS cases. 'Sequencing' is carried out by using multivariate hypergeometric sampling, 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.

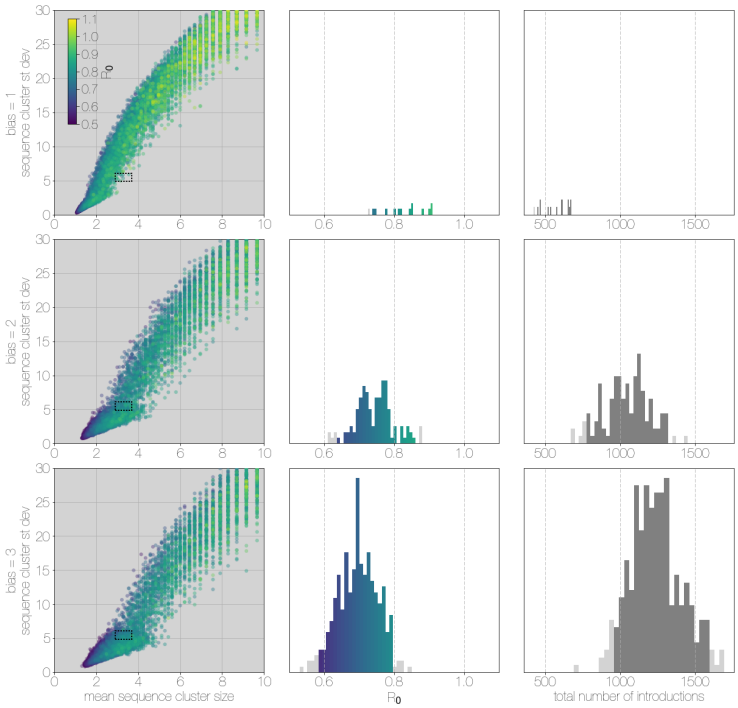

Figure S4. Results of Monte Carlo simulations with vast underestimation of cases. The plot is identical to Figure 3, but instead of 2000 cases, simulations were run with 4000 cases. With more unobserved cases the R_0 values matching observed MERS-CoV sequence clusters can only be smaller, with a corresponding increase in numbers of zoonotic transmissions. However, the numbers of simulations that match MERS-CoV data go down as well.

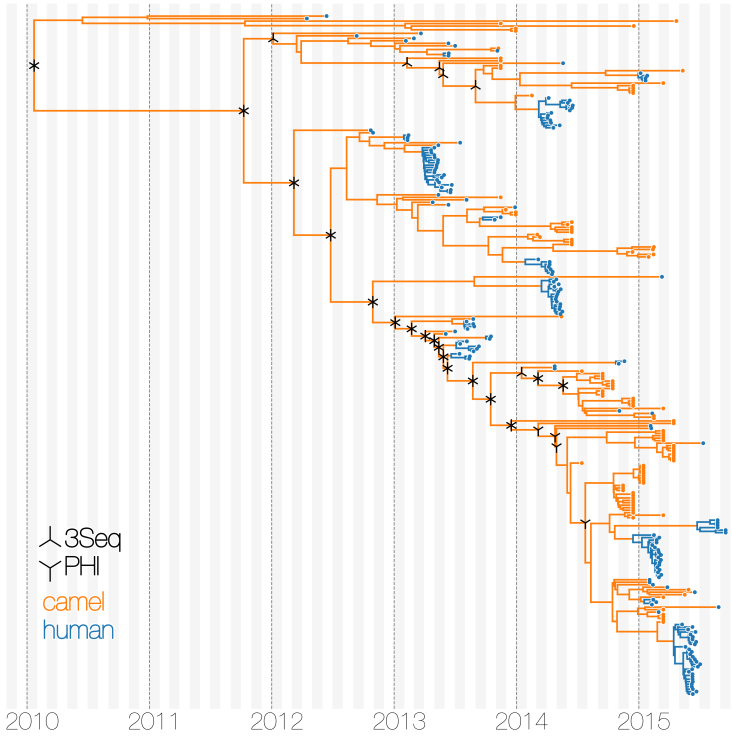

Figure S5. Tests of recombination across MERS-CoV clades. Maximum clade credibility tree of MERS-CoV genomes annotated with results of two recombination detection tests (PHI and 3Seq) applied to descendent sequences of each clade. Both tests identify large portions of existing sequence data as containing 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.

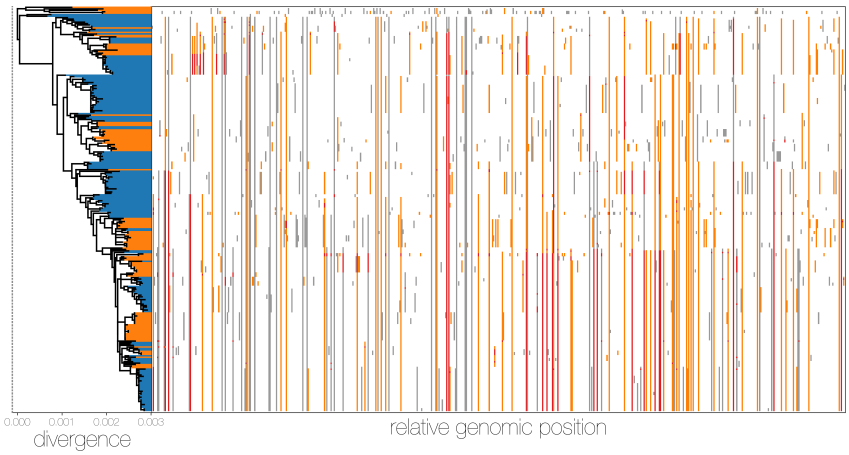

Figure S6. MERS-CoV genomes exhibit high numbers of non-clonal loci. Ancestral state reconstruction (right) identifies a large number of sites in which mutations have occurred more than once in the tree (homoplasies, orange) or are reversions (red) from a state arising in an ancestor. Mutations that 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).

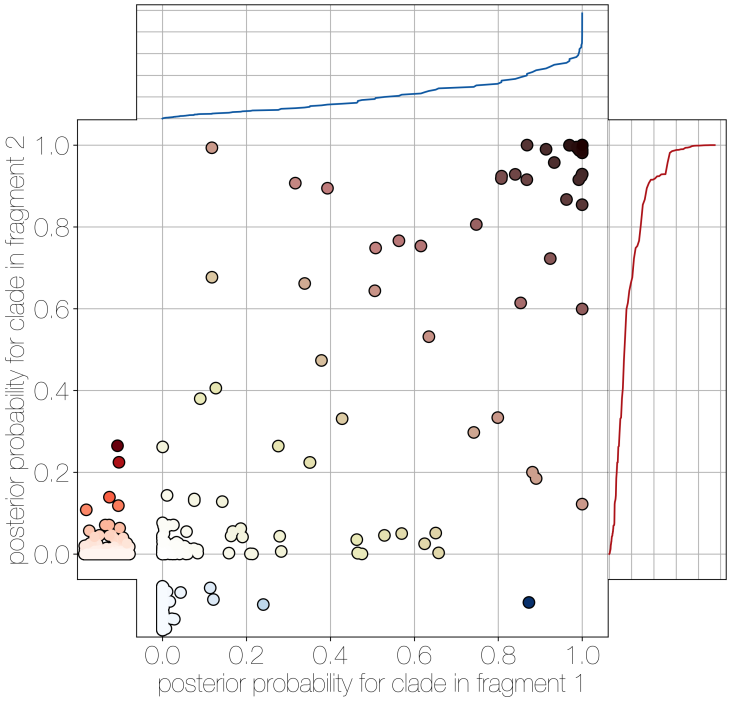

Figure S7. Human clade sharing between genomic fragments 1 and 2. Central scatter plot shows the posterior probability of human clades shared between genomic fragments 1 and 2, in their respective trees. Left and bottom scatter plots track the posterior probability of human clades only observed in fragment 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).

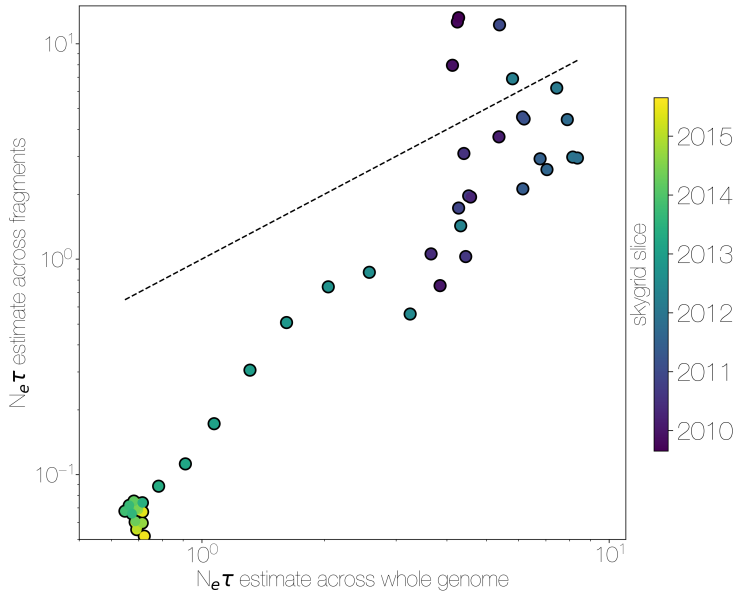

Figure S8. Skygrid comparison between whole and fragmented genomes. Inferred median $N_e\tau$ recovered using a skygrid tree prior on whole genome (bottom) and ten genomic fragments with independent trees (left), coloured by time. Dotted line indicates the one-to-one line.