Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study

Matthew Cotten, Simon J Watson*, Paul Kellam*, Abdullah A Al-Rabeeah, Hatem Q Makhdoom, Abdullah Assiri, Jaffar A Al-Tawfiq, Rafat F Alhakeem, Hossam Madani, Fahad A AlRabiah, Sami Al Hajjar, Wafa N Al-nassir , Ali Albarrak, Hesham Flemban, Hanan H Balkhy, Sarah Alsubaie, Anne L Palser, Astrid Gall, Rachael Bashford-Rogers, Andrew Rambaut, Alimuddin I Zumla*, Ziad A Memish**

Summary

Background **Since June, 2012, Middle East respiratory syndrome coronavirus (MERS-CoV) has, worldwide, caused 104 infections in people including 49 deaths, with 82 cases and 41 deaths reported from Saudi Arabia. In addition to confirming diagnosis, we generated the MERS-CoV genomic sequences obtained directly from patient samples to provide important information on MERS-CoV transmission, evolution, and origin.**

Methods **Full genome deep sequencing was done on nucleic acid extracted directly from PCR-confirmed clinical samples. Viral genomes were obtained from 21 MERS cases of which 13 had 100%, four 85–95%, and four 30–50% genome coverage. Phylogenetic analysis of the 21 sequences, combined with nine published MERS-CoV genomes, was done.**

Findings **Three distinct MERS-CoV genotypes were identified in Riyadh. Phylogeographic analyses suggest the MERS-CoV zoonotic reservoir is geographically disperse. Selection analysis of the MERS-CoV genomes reveals the expected accumulation of genetic diversity including changes in the S protein. The genetic diversity in the Al-Hasa cluster suggests that the hospital outbreak might have had more than one virus introduction.**

Interpretation **We present the largest number of MERS-CoV genomes (21) described so far. MERS-CoV full genome sequences provide greater detail in tracking transmission. Multiple introductions of MERS-CoV are identified and suggest lower R_0 values. Transmission within Saudi Arabia is consistent with either movement of an animal reservoir, animal products, or movement of infected people. Further definition of the exposures responsible for the sporadic introductions of MERS-CoV into human populations is urgently needed.**

Funding **Saudi Arabian Ministry of Health, Wellcome Trust, European Community, and National Institute of Health Research University College London Hospitals Biomedical Research Centre.**

Introduction

Middle East respiratory syndrome (MERS) is a newly described disease in human beings, first reported from Saudi Arabia in September, 2012, after identification of a novel betacoronavirus (MERS-CoV) from a Saudi Arabian patient who died from a severe respiratory illness.[1] As of Sept 12, 2013, there have been 114 laboratory-confirmed cases of MERS-CoV infections with 54 deaths reported to WHO.[2] All cases have been directly or indirectly linked to one of four countries in the Middle East (Saudi Arabia, Jordan, Qatar, and the United Arab Emirates), with most cases (90 cases and 44 deaths) reported from Saudi Arabia, occurring as sporadic, family, or hospital clusters.[2] Human-to-human transmission of MERS-CoV has been documented in England, France, Tunisia, Italy, and Saudi Arabia.

Coronaviruses are a family of viruses infecting birds and mammals. The animal source and mode of transmission of MERS-CoV to human beings is not known. This information is essential for developing interventions for reducing the risk of transmission and developing effective control measures. During the severe acute respiratory syndrome coronavirus (SARS-CoV) epidemic between Nov, 2002, and July, 2003, molecular analysis of SARS-CoV from patients from various geographical regions was essential for understanding viral evolution and the spread of the disease.[3] Phylogenetic analysis suggested that SARS-CoV probably originated in bats and spread to people. A genetic link between the SARS-CoV in people and in civets revealed cross-host evolution.[4]

There is currently little information about the molecular evolution of MERS-CoV and how this relates to virus transmission. The cellular receptor for MERS-CoV has been identified as dipeptidyl peptidase 4 (DPP4, CD26)[5] and the structure of the receptor binding domain of the virus spike protein complexed with DPP4 has been established.[6] The tissue distribution of this receptor within mammals is consistent with the lung and kidney pathology of the virus.[5] The conservation of the receptor across mammals suggests many possible non-human hosts; however, so far no animal reservoir has been identified for MERS-CoV. The closest relative to MERS-CoV was identified through phylogenetic analysis of a short fragment sequenced from the bat species *Neoromicia zuluensis*.[7] However, in view of the estimated

Lancet 2013; 382: 1993–2002

Published Online September 20, 2013 http://dx.doi.org/10.1016/S0140-6736(13)61887-5

See Comment page 1962

*Contributed equally

Global Centre for Mass Gatherings Medicine, Ministry of Health, Riyadh, Saudi Arabia (Prof Z A Memish FRCP, A A Al-Rabeeah FRCS, Prof A I Zumla FRCP, A Assiri MD, R F Alhakeem MD); **Wellcome Trust Sanger Institute, Hinxton, UK** (M Cotten PhD, S J Watson PhD, A L Palser PhD, A Gall Dr Med Vet, R Bashford-Rogers MChem, Prof P Kellam PhD); **Jeddah Regional Laboratory, Ministry of Health, Jeddah, Saudi Arabia** (H Q Makhdoom PhD, Hossam Madani PhD); **Saudi Aramco Medical Services Organisation, Saudi Aramco, Dhahran, Saudi Arabia** (J A Al-Tawfiq MD); **Institute of Evolutionary Biology, Ashworth Laboratories, Kings Buildings, West Mains Road, Edinburgh, UK** (Prof A Rambaut PhD); **Fogarty International Center, NIH, Bethesda, MD, USA** (A Rambaut); **King Faisal Specialist Hospital, Riyadh, Saudi Arabia** (F A AlRabiah MD, S Al Hajjar MD); **Imam Abdulrahman Bin Mohamed Hospital-National Guard Health Affairs-Dammam, Saudi Arabia** (W N Al-nassir MD); **Prince Sultan Military Medical City, Riyadh, Saudi Arabia** (A Albarrak MD); **Alhada Military Hospital, Riyadh, Saudi Arabia** (H Flemban MD); **King Abdulaziz Medical City, Riyadh, Saudi Arabia** (H H Balkhy MD); **Paediatric Infectious Diseases, King Saud University, Saudi Arabia** (S Alsubaie MD); **Division**

of Infection and Immunity, University College London, London, UK (Prof A I Zumla, Prof P Kellam)**; and UCL Hospitals NHS Foundation Trust, London, UK** (Prof A I Zumla)

Correspondence to: Prof Ziad A Memish, Global Centre for Mass Gatherings Medicine, Ministry of Health, Riyadh 11176, Saudi Arabia **zmemish@yahoo.com**

evolutionary rate of MERS-CoV, the most recent common ancestor between this isolate and MERS-CoV existed in bats more than 44 years ago.[8] A recent serological study of dromedary camels in Oman and the Canary Islands found cross-reactive antibodies to MERS-CoV, but the investigators were unable to amplify any MERS-CoV-like viral sequences from the samples.[9] A small fragment of sequence identical to the EMC/2012 MERS-CoV has been reported from a *Taphozous perforatus* bat captured in Saudi Arabia, suggesting a regionally relevant bat reservoir.[10]

Genomic sequencing of MERS-CoV is important, and molecular epidemiology can reveal spatiotemporal patterns that help identify whether all MERS-CoV infections originated from a single zoonotic event, with subsequent human-to-human transmission, or from many zoonotic events at several geographic locations. This information is crucial for an accurate assessment of the epidemic potential of MERS-CoV.[11] In addition to confirming diagnosis, generating sequence data directly from epidemiologically defined cases provides the essential basis for defining the spread, evolution, and origin of

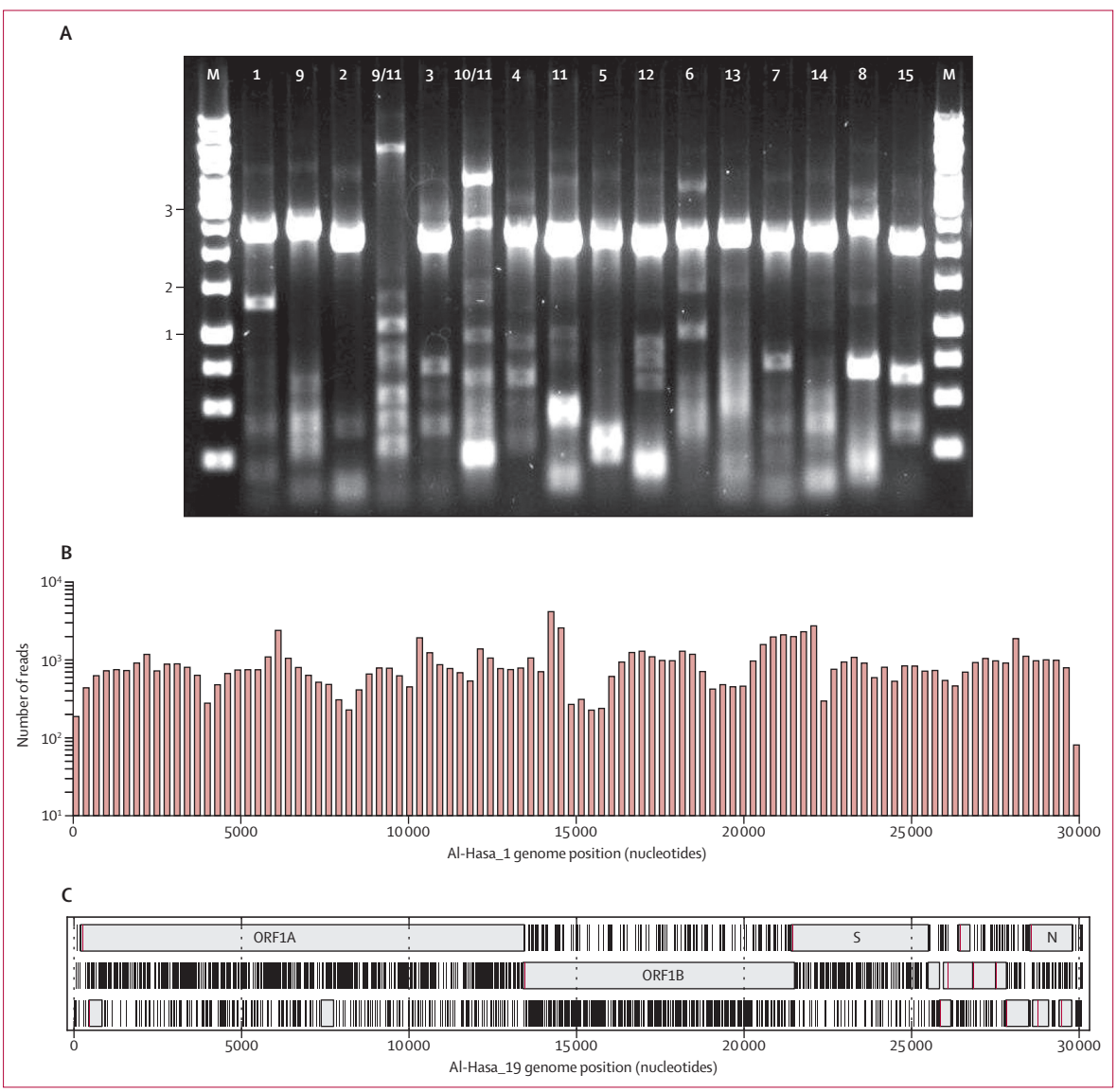

Figure 1: **Deep sequencing process**
Products of RT-PCR amplification of MERS-CoV patient RNA (A). All expected products are 2–3 kb in length, with amplicons 9/11 and 10/11 producing 6 and 4 kb products. DNA marker sizes are shown in kb. (B) 100 000 random reads (a tenth of the entire sample dataset for Al-Hasa_19) were mapped to the Al-Hasa_1_2013 genome. The histogram shows the positions of each read across the genome with 100–1000 reads at each position across the entire genome. The actual coverage is ten times greater, yielding 1000–10 000 coverage. Open reading frame (ORF) map (C) of a successfully assembled Al-Hasa_19 genome. An open box shows open reading frames of 100 codons or greater in length. Stop codons are shown by vertical black lines. The first ATG in each ORF is shown by a vertical red line. MERS-CoV=Middle East respiratory syndrome coronavirus.

MERS-CoV infections. Here we report MERS-CoV genomes obtained directly from 21 patients with MERS from across Saudi Arabia and assess the spatiotemporal distribution of MERS-CoV in Saudi Arabia.

Methods

Screening tests and genome sequencing

Clinical samples were screened with RT-PCR, as described elsewhere,[12,13] with amplification targeting both the upE and ORF1A for confirmation. Deep sequencing was done on nucleic acid extracted from real-time PCR confirmed cases of MERS-CoV in Saudi Arabia. Typically 50 µL of nucleic acid was generated from 200 µL of tracheal aspirate or from a nasopharyngeal or throat swab with automated processing. PCR amplified DNA amplicons covering the entire MERS-CoV genome were prepared as described elsewhere.[14] The PCR amplicons for each sample were pooled for Illumina library (Illumina, San Diego, CA, USA) preparation with each sample processed to include a unique barcode sequence. Standard MiSeq 150 base pair paired-end reads were generated. Sequence data were de-multiplexed into sample-specific readsets, processed to remove primer sequences at the ends of reads, and trimmed from their 3′ end until the median phred-scaled quality score was >35·0, discarding reads smaller than 125 nucleotides. The processed readsets were assembled into large contiguous sequences (contigs) using the de-novo assembler SPAdes.[15] The readset was also mapped against the Al-Hasa_1_2013 genome (GenBank accession number KF186567[16]) using SMALT (version 0.5.0) to generate a reference-based consensus. In cases where the de-novo assembly was split over several contigs, they were aligned against the reference-based consensus before merging. Any differences between de-novo and reference-based methods were resolved by direct examination of the original read set.

Phylogenetic analyses

The 13 complete genomes and eight 32% or greater genomes were aligned with the nine published MERS-CoV genomes (GenBank accession numbers JX869059, KC776174, KC667074, KF192507, KF186564, KF186565, KF186566, KF186567, plus England2_HPA, not yet in GenBank) using MEGA5.[17] The ends of the alignment were trimmed to the longest shared sequence, and the Bayesian inference of the phylogeny done using MRBAYES version 3.2.1[18] under a GTR+Γ_4 substitution model. The hypothetical ancestral sequences were generated with a likelihood-based ancestral reconstruction method implemented in HYPHY version 2.1.2.[19] Custom Python scripts were used to determine non-synonymous changes along each internal branch of the phylogeny.

A subalignment of the 20 sequences comprising the Al-Hasa outbreak was generated to assess the transmission dynamics of the hospital outbreak. The number of nucleotide differences between each pair of sequences

	Type of sample used as source of viral nucleic acid	Proportion of MERS-CoV genome*	Patient†	Sample date	GenBank accession number
Al_Hasa_1_2013	Tracheal aspirate	100%	V	May 9, 2013	KF186567
Al_Hasa_2_2013	Tracheal aspirate	100%	J	April 21, 2013	KF186566
Al_Hasa_3_2013	Tracheal aspirate	99·9%	I	April 22, 2013	KF186565
Al-Hasa_7_2013	Tracheal aspirate	92·8%	H	May 1, 2013	KF600623, KF600655
Al_Hasa_4_2013	Tracheal aspirate	99·9%	K	May 1, 2013	KF186564
Al-Hasa_15_2013	Tracheal aspirate	99·9%	Y	May 11, 2013	KF600645
Al-Hasa_16_2013	Tracheal aspirate	99·9%	T	May 12, 2013	KF600644
Al-Hasa_13_2013	Nasopharyngeal swab	37·5%	Q	May 7, 2013	KF600616, KF600637, KF600640, KF600650, KF600656
Al-Hasa_22_2013	Tracheal aspirate	46·8%	X	May 9, 2013	KF600617, KF600619, KF600621, KF600625, KF600631, KF600633
Al-Hasa_9_2013	Tracheal aspirate	46·1%	F	May 1, 2013	KF600622, KF600639, KF600648, KF600649, KF600654
Al-Hasa_10_2013	Nasopharyngeal swab	32·0%	M	May 2, 2013	KF600614, KF600624, KF600629, KF600636, KF600641, KF600642, KF600646, KF600653
Al-Hasa_11_2013	Tracheal aspirate	90·5%	G	May 3, 2013	KF600629, KF600636, KF600646
Al-Hasa_8_2013	Nasopharyngeal swab	73·8%	O	May 1, 2013	KF600618, KF600626, KF600635, KF600638
Al-Hasa_14_2013	Nasopharyngeal swab	75·2%	S	May 8, 2013	KF600615, KF600643
Al-Hasa_12_2013	Nasopharyngeal swab	99·9%	B	May 7, 2013	KF600627
Riyadh_1_2012	Tracheal aspirate	99·8%	RY1‡	Oct 23, 2012	KF600612
Riyadh_2_2012	Nasopharyngeal swab	99·9%	RY2§	Oct 30, 2012	KF600652
Buraidah_1_2013	Unknown source	99·9%	BR1	May 13, 2013	KF600630
Al-Hasa_17_2013	Tracheal aspirate	100%	AH17	May 15, 2013	KF600647
Al-Hasa_18_2013	Tracheal aspirate	100%	AH18	May 23, 2013	KF600651
Al-Hasa_19_2013	Nasopharyngeal swab	100%	AH19	May 23, 2013	KF600632
Bisha_1_2012	Nasopharyngeal swab	99·8%	BS1¶	June 19, 2012	KF600620
Riyadh_3_2013	Respiratory swab	99·8%	RY3	Feb 5, 2013	KF600613
Al-Hasa_21_2013	Throat swab	99·8%	AH21	May 30, 2013	KF600634
Hafr-Al-Batin_1_2013	Nasopharyngeal swab	100%	HB1	June 4, 2013	KF600628

MERS-CoV=Middle East respiratory syndrome coronavirus. *Proportion of genome obtained compared with a full genome value of 30 119 nucleotides. †Code used in figure 2; single letter codes refer to patients from Assiri and colleagues.[16] ‡Patient described by Albarrak and colleagues.[24] §Patient described by Memish and colleagues.[25] ¶Same patient providing sample for van Boheemen and colleagues.[26]

Table 1: **MERS-CoV genome data**

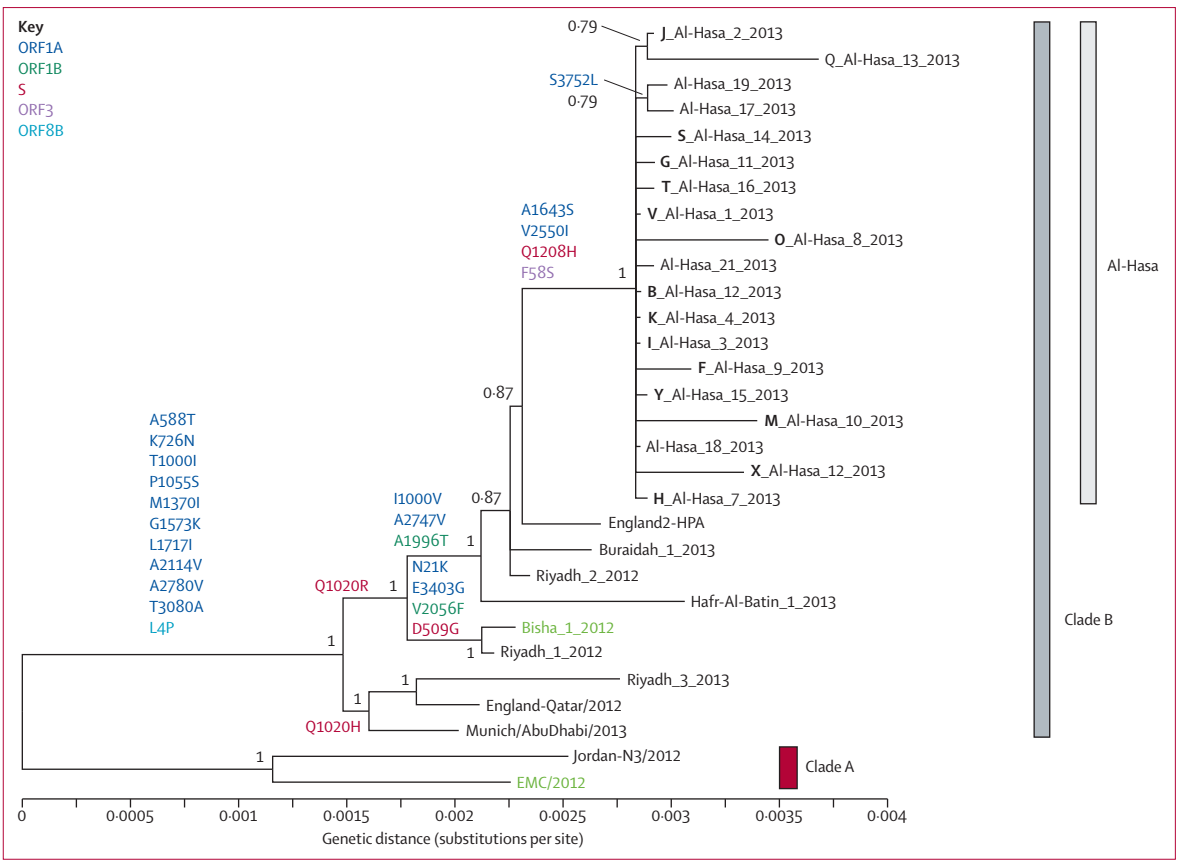

Figure 2: **Bayesian-inferred phylogeny of all 21 new sequences**
Combined with the published genomes (EMC/2012 [GenBank number JX869059], Jordan-N3 [KC776174], Munich/AbuDhabi [KF192507], England-Qatar_2012 [KC667074], Al-Hasa_1_2013 [KF186567], Al-Hasa_2_2013 [KF186566], Al-Hasa_3_2013 [KF186565], Al-Hasa_4_2013 [KF186564], and England2-HPA [no number available]). The single letter patient codes from Assiri and colleagues[16] are given where appropriate. Clade A, clade B, and the Al-Hasa cluster are marked with vertical bars. Aminoacid changes along the internal branches were established though likelihood-based ancestral state reconstruction. These are shown above the branches and colour-coded by ORF. The scale bar below the phylogeny shows the genetic distance, in substitutions per site, from the arbitrary midpoint root. Bayesian posterior probabilities for each clade are given above the relevant node. MERS-CoV=Middle East respiratory syndrome coronavirus. ORF=open reading frame.

was calculated and compared with the expected number of mutations in view of the time elapsed between the sampling dates of each pair. Because the number of mutations between two sequences accords with Poisson distribution, rejection of the epidemiological linkage between the two is considered if the observed number of mutations falls outside of the 95% upper confidence level in the cumulative density function. To account for the many independent comparisons, a Bonferroni correction was applied, adjusting the significance level to $3{\cdot}85\times10^{-3}$.

Temporal dynamics of MERS-CoV were assessed with time-resolved phylogenies using the Bayesian Markov Chain Monte Carlo method in BEAST version 1.7.5.[20] Because of the close epidemiological relation between the nine Al-Hasa isolates, the cluster was collapsed to the earliest isolate (Al-Hasa_2_2013) resulting in a final set of 11 MERS-CoV isolates before generating molecular clock phylogenies. A second alignment of only-coding regions (ORF1ab, S, ORF3, ORF4a, ORF4b, ORF5, E, M, and N), was used for selecting codon-position substitution models. The most appropriate evolutionary model was established by comparing the marginal likelihood of different models, estimated using the path-sampling approach implemented in BEAST. The HKY+Γ_4 substitution model was selected as most appropriate, with separate rates for the three codon positions, under an uncorrelated lognormal molecular clock[21] and a flexible Gaussian Markov random field Bayesian skyride coalescent.[22]

The geographical locations of ancestral nodes in the time-resolved phylogeny were inferred with a discrete phylogeographical diffusion model; Bayesian stochastic search variable selection identified the statistically significant transition rates between locations. Both symmetrical and asymmetrical substitution models were tested in combination with either a geographical-distance-informed model, in which the transition rate of a virus between locations is inversely proportional to the distance between them, or a strict model, with equal rates between all locations. Marginal likelihood values for each model were estimated using path-sampling, and

employed in the Bayes factor comparison to establish the better supported model. There was no difference between the four models (Bayes factors <3), so in view of the limited number of sequences and locations, the symmetrical substitution model under equal rates was chosen to reduce the risk of over-paramaterisation.

Positively-selected sites in the MERS-CoV genomes were detected with the mixed effects model of evolution implemented in HyPhy[23] to calculate the ratio of synonymous to non-synonymous substitutions (ω) for each codon, allowing this ratio to vary between the branches of the alignment's phylogenetic tree to detect occurrences of episodic selection, where the site under selection only occurs on a subset of samples on a specific lineage in the phylogeny.

Role of the funding source

The sponsor of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

Sequencing of the MERS-CoV needs conversion of the 30 000 nucleotide RNA genome to DNA using reverse transcription and subsequent PCR amplification. Figure 1 shows a typical set of successful reverse transcription PCR amplicons. The amplicons for each sample were pooled for Illumina sequencing as described in the methods section. Coverage of the resulting short-reads across the MERS-CoV genome was assessed by mapping a random 100 000 read subset to the Al-Hasa_1_2013 genome (figure 1), where the average coverage for all samples was shown to be greater than 3×10^3. After de-novo assembly, genomes were validated by ensuring that the expected open reading frames (ORFs) were intact (figure 1). This method was used to generate complete or partial genomes from 21 MERS cases with sample collection dates from June, 2012, to June, 2013 (table 1).

The phylogenetic relation of the Saudi Arabian MERS-CoV sequences was assessed in combination with the nine published MERS-CoV genomes. The earliest reported MERS-CoV genomes are from June, 2012, from a patient in Bisha (EMC/2012),[26] and a retrospectively identified Jordan patient from April, 2012 (Jordan-N3). These two genomes seem distinct and are provisionally grouped here in a separate clade A (figure 2). The EMC/2012 sequence was obtained after extensive cell culture passage to establish a virus isolate;[26] there are 76 single nucleotide differences between EMC/2012 and the root of the clade B lineage. EMC/2012 and Jordan-N3 share 44 nucleotide changes not present in any other known MERS-CoV genome. Because of the lack of reported technical details for the Jordan sequence, the tissue culture adaptation of EMC/2012, and the shared polymorphisms despite large difference in geographical origins of clade A, we focused our analysis on sequences obtained directly from patient material, namely those in clade B. Importantly, genome Bisha_1_2012 (figure 2, in light green) was obtained with direct sequencing of nasopharyngeal swab material from the same patient reported as the source of the EMC/2012 virus (figure 2, in light green).[26]

A time-resolved phylogeny was generated from all epidemiologically unlinked viruses with genome coverage greater than 70% (figure 3). The tightly-linked Al-Hasa cluster was collapsed to a representative isolate (Al-Hasa_2_2013) along with the Al-Hasa_8_2013 isolate that was determined to be epidemiologically unrelated (figure 4, table 2). Geographical locations of the ancestral viruses were co-estimated along with the phylogeny, to assess the spatial evolution of the virus. These results suggest the circulating virus in Saudi Arabia is centred around Riyadh, with sporadic excursions to other centres; the most probable geographical location for all of the internal nodes, except for the Al-Hasa outbreak, is in Riyadh (figure 3). However, with the exception of the ancestor to the Al-Hasa outbreak, the posterior probability on these ancestral geographical locations is low, highlighting the degree of uncertainty on these inferred locations. The number of isolates for each location is small, with only Riyadh and Al-Hasa having more than one isolate. Because the two Al-Hasa isolates cluster together, the positioning of Riyadh along the internal branches could be due to it being the only location present at many distinct positions in the topology.

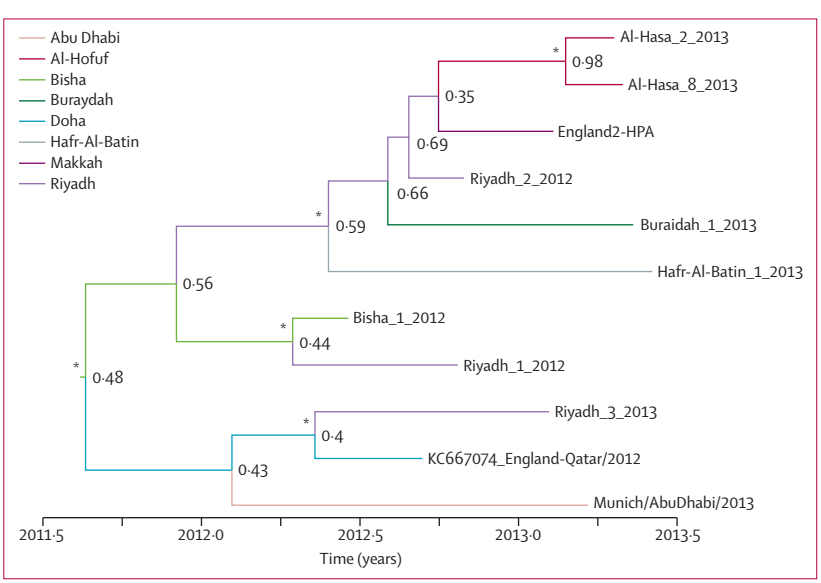

Figure 3: **Time-resolved phylogenetic tree**
Based on the concatenated coding regions of the MERS-CoV genome. Branch colours show the most probable geographical location for that branch, established with a discrete traits model implemented in BEAST version 1.7.5.[20] Change in branch colour shows a change in geographical location during its evolutionary history. Node labels show the posterior probability for the inferred geographical location at that node. Asterisks show nodes with >0·95 posterior probability support for that clade. The posterior probabilities on the geographical location of the root are Al-Hofuf 0·03, Riyadh 0·48, Buraidah 0·04, Bisha 0·18, Abu Dhabi 0·05, Doha 0·13, Hafr-Al-Batin 0·04, and Makkah 0·04. MERS-CoV=Middle East respiratory syndrome coronavirus

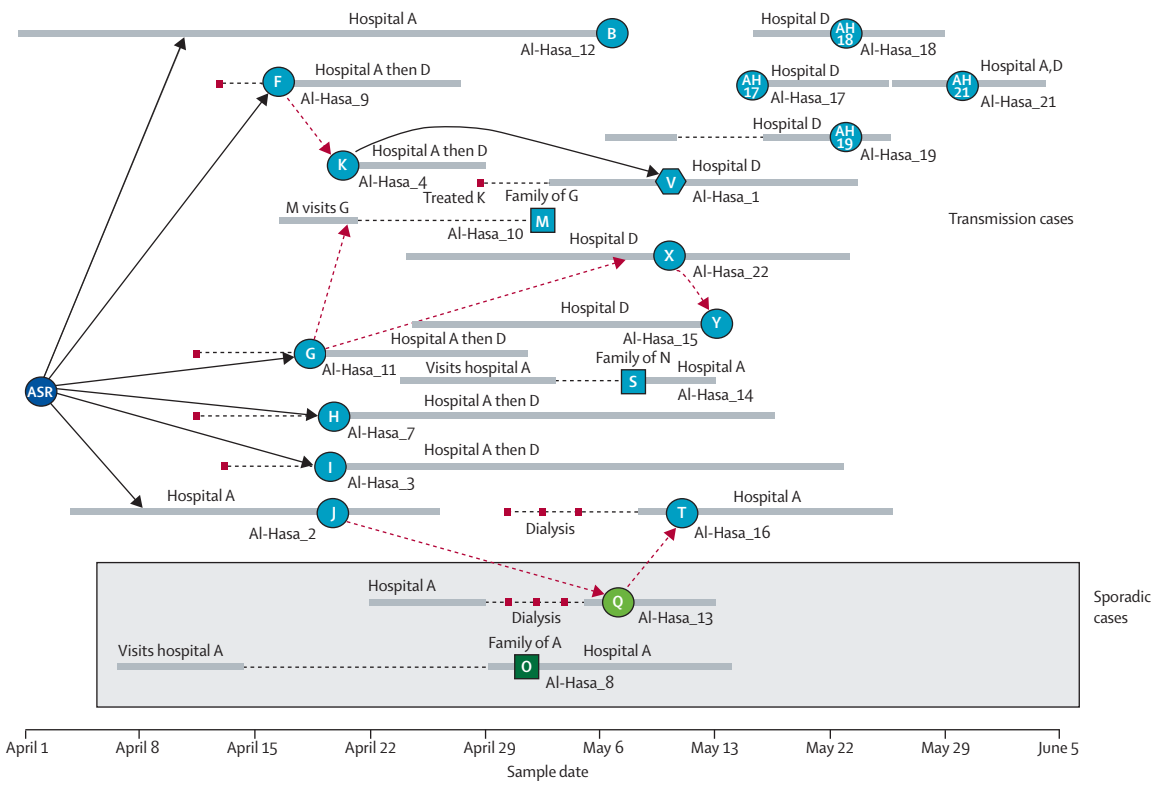

Figure 4: **Transmission analysis**

Proposed transmission network for all available Al-Hasa MERS-CoV genomes. The light blue nodes show genomes with at least one statistically possible transmission linkage to a known MERS-CoV genome (transmission cases). The coloured nodes in the bottom panel (sporadic cases) are MERS-CoV genomes that cannot be linked by direct transmission to any known MERS-CoV case. Genomes are organised by date of sample collection, and stays in hospital A or D are shown by grey horizontal bars. Red squares show potential transmission, either by shared dialysis session or by direct exposure to a known MERS case. Black arrows show proposed transmission pairs supported by the sequence data, dashed red arrows show proposed transmission pairs not supported by the sequence data. The expected number of differences between each pair of sequences was calculated as the product of the interval between sampling, the evolutionary rate of the virus, and the maximum common length between the two virus genomes. Assuming that the number of differences between two sequences over a length of time is Poisson distributed, with λ equal to the expected number of mutations, the probability of getting the observed number of mutations between the two sequences by chance can be calculated from the cumulative density function of the Poisson distribution. A transmission pair was rejected if the cumulative probability value was less than 0·05. A Bonferroni correction was applied to account for multiple comparisons, resulting in an adjusted significance level of $3{\cdot}85\times10^{-3}$. ASR=ancestral sequence reconstruction as defined in table 2. MERS-CoV=Middle East respiratory syndrome coronavirus.

	Δ time (days)	Recorded substitutions	Expected substitutions	p value* (X≥observation)
ASR *vs* B	34	0	1·869	1
ASR *vs* F	28	3	0·714	$3{\cdot}58\times10^{-2}$
ASR *vs* G	30	2	1·494	$4{\cdot}40\times10^{-1}$
ASR *vs* H	28	1	1·436	$7{\cdot}62\times10^{-1}$
ASR *vs* I	19	0	1·045	1
ASR *vs* J	18	1	0·985	$6{\cdot}26\times10^{-1}$
F *vs* K	0	3	0	NA†
G *vs* X	6	7	0·139	$1{\cdot}76\times10^{-10}$‡
G *vs* M	1	4	0·018	$4{\cdot}55\times10^{-9}$‡
J *vs* Q	16	8	0·332	$2{\cdot}74\times10^{-9}$‡
K *vs* V	8	0	0·422	1
Q *vs* T	5	9	0·106	$4{\cdot}00\times10^{-15}$‡
X *vs* Y	2	7	0·057	$3{\cdot}50\times10^{-13}$‡

MERS-CoV=Middle East respiratory syndrome coronavirus. ASR=likelihood-based ancestral state reconstruction of the root of the Al-Hasa clade. *p values for accepting or rejecting the null hypothesis are significant at the 0·05 level, after using the Bonferroni correction to adjust the p value for each hypothesis to $3{\cdot}85\times10^{-3}$. †The samples from F and K were collected on the same day resulting in an expected substitution value of 0. Therefore no conclusions could be made for this pair. ‡Transmission pair not statistically supported.

Table 2: **MERS-CoV genome variation in support of human transmission pairs**

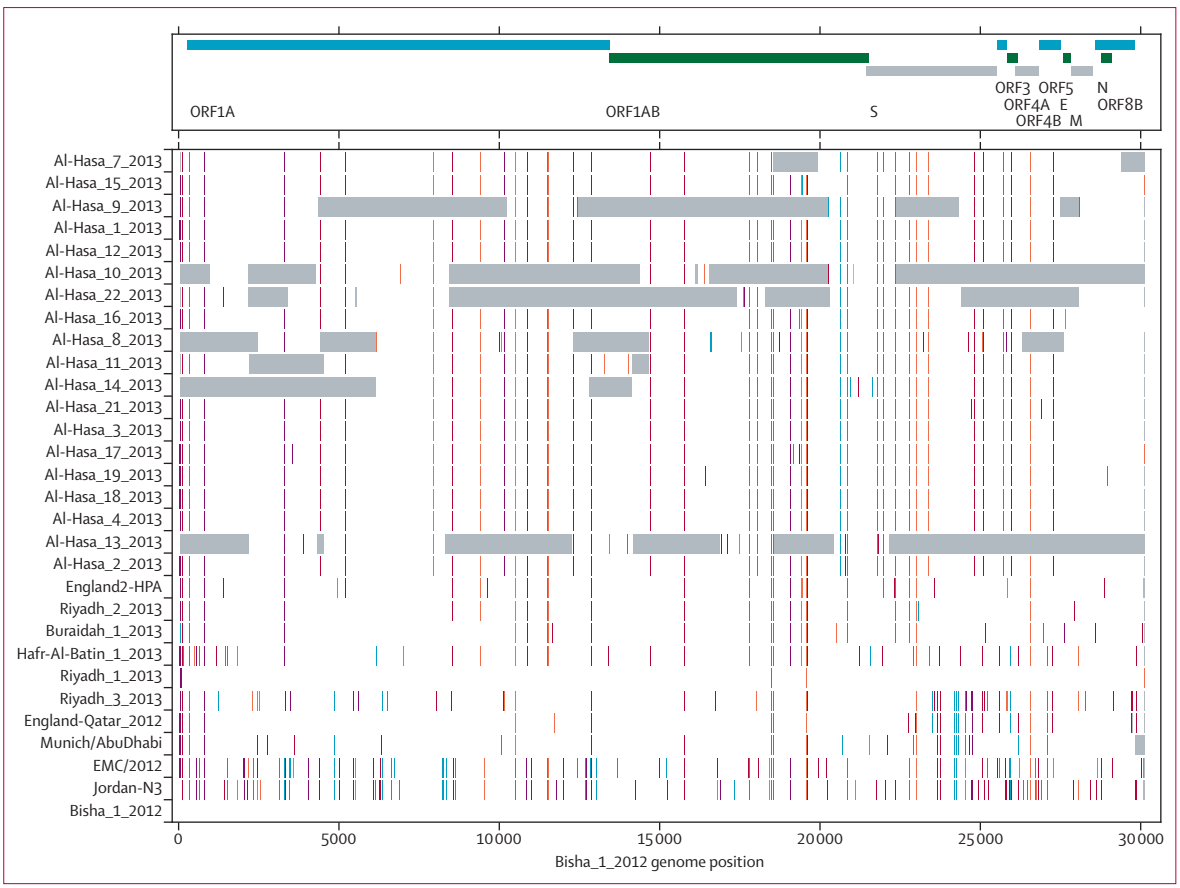

Figure 5: **Single nucleotide differences between the Bisha_1_2012 genome and all other genomes**
Difference shown by vertical coloured bars. Grey=gap in the query sequence. Orange=change to A. Red=change to T. Blue=change to G. Purple=change to C. The ORF map of the MERS-CoV genome with the major ORFs marked is provided above the figure for reference. MERS-CoV=Middle East respiratory syndrome coronavirus. ORF=open reading frame.

However, although the posterior probability on the root node being Riyadh is low (0·48), it is still more than double that of any other location, suggesting that the low probabilities are due to the small number of genomes (the legend of figure 3 lists the posterior probabilities). Additional genomes will be needed to improve confidence in these inferences.

Estimation of the evolutionary rate for this expanded number sequences shows MERS-CoV evolving at $6{\cdot}3\times10^{-4}$ substitutions/site per year (95% highest posterior density [HPD] $1{\cdot}4\times10^{-4}$ to $1{\cdot}1\times10^{-3}$) establishing the time to most recent common ancestor for clade B (excluding EMC/2102 and Jordan-N3) as July, 2011 (95% HPD July, 2007, to June, 2012).

Single nucleotide differences between all MERS-CoV genomes were assessed (figure 5). In this analysis, all available MERS-CoV genomes were aligned and the nucleotide at each position was compared with the nucleotide at the same position in the earliest genome, Bisha_1_2102. Many nucleotide changes were in the last third of the genome; if these changes result in aminoacid changes they might alter important accessory proteins as well as the receptor-binding spike protein encoded at nucleotide positions 21000–25 500. Selection analysis of the 11 epidemiologically unlinked MERS-CoV genomes was done to detect evidence of natural selection in the MERS-CoV genome. Using the mixed effects model of evolution method, codon 1020 in the spike gene was identified as being under episodic selection (p=0·021) along the branch between the main Saudi Arabian lineage and the Munich/AbuDhabi/2013 cluster. The isolates in the Munich/AbuDhabi/2013 cluster have a histidine at this position, whereas those in the Saudi Arabian lineage have an arginine (figure 2). The presence of an arginine at this position in the Saudi Arabian lineage suggests a potential cleavage site for the endosomal protease furin[27] or for trypsin-like proteases.[28,29] Because of the possible role of spike protein cleavage in potentiating membrane fusion and coronavirus entry,[30] and previous observations on SARS coronavirus spike evolution,[31,32] the phenotypic effect of this change should be established.

Evidence was sought for multiple, distinct zoonotic introductions into human beings of MERS-CoV compared with a single introduction of clade B with

sustained human-to-human transmission. The genome Riyadh_1_2012 was obtained from an early MERS patient in Riyadh on Oct 23, 2012.[24] Genome Riyadh_2_2012 was from patient 2 in the Riyadh family cluster obtained 7 days later on Oct 30, 2012.[25] The two genomes are phylogenetically distinct and in view of the incubation periods of MERS-CoV[16] suggest at least two distinct lineages were circulating in Riyadh in October, 2012.

15 of the new virus genomes from the Al-Hasa region form a polytomy with the previously described Al-Hasa_1_2013 through Al-Hasa_4_2013 cluster[16] (marked Al-Hasa in figure 2). These genomes were used to assess MERS-CoV transmission patterns in detail. In view of the rate of evolution for MERS-CoV, we tested the likelihood of the transmission network for the Al-Hasa cluster.[16] Although patient A was suggested to be the index case in Al-Hasa, transmitting to patient C and then onward, neither patient A nor C were molecularly confirmed MERS-CoV cases. The similarity of viruses from patients B, F, G, H, and I suggests that a closely related virus could be the important early virus in the outbreak.

We reconstructed the hypothetical ancestral sequence (ASR) of the Al-Hasa cluster as a surrogate for this original virus. Pairwise statistical assessment of epidemiologically defined transmission events, in view of the observed genetic divergence of the virus genomes and the time intervals, supports eight of 13 proposed transmissions in the Al-Hasa outbreak[16] (table 2); however, transmissions between the Al-Hasa patients G:M, G:X, X:Y, J:Q, and Q:T are not statistically supported (table 2, figure 4), suggesting independent sources of infections within the Al-Hasa outbreak (figure 4).

The geographical distribution and phylogenetic relation of MERS-CoV across time in Saudi Arabia was investigated. The close phylogenetic clustering of the MERS-CoV isolates from the Al-Hasa region is consistent with spread via human-to-human transmission. In addition to the Al-Hasa cluster, many well supported phylogenetic clades and singletons (hereafter referred to as genotypes) exist, possibly each from a separate zoonotic event (figure 2). These distinct genotypes, which include the Munich/AbuDhabi, England-Qatar, Riyadh_3 cluster; the Bisha-1, Riyadh_1 cluster; and the singletons Hafr-Al-Batin_1, Riyadh_2, Buraidah_1, England2-HPA, each have strong statistical support, with Bayesian posterior probabilities of 0·87 or greater—often reaching 1·0.

***Panel:* Research in context**

Systematic review

We searched PubMed, ProMed, and Genbank on Aug 15, 2013, for all relevant English language publications with the terms "Middle East respiratory syndrome", "MERS-CoV", and "HCoV-EMC" individually and in combination with the terms "sequence", "genome", and "phylogeny". Articles relevant to our report are cited in the text. The phylogenetic relation of our 21 Saudi Arabian MERS-CoV sequences was assessed in combination with the nine published MERS-CoV genomes. The geographical distribution and phylogenetic relation of MERS-CoV across time in Saudi Arabia was reviewed. Geographical locations of the ancestral viruses were co-estimated along with the phylogeny, to assess the spatial evolution of the virus.

Interpretation

MERS-CoV full genome sequences provide greater detail in tracking transmission. There is little information about the molecular evolution of MERS-CoV and how this relates to virus transmission. MERS-CoV genomes obtained directly from 21 MERS patients from across Saudi Arabia show three distinct MERS-CoV genotypes in Riyadh. Phylogeographic analyses suggest the MERS-CoV zoonotic reservoir is geographically disperse. Selection analysis of the MERS-CoV genomes reveals the expected accumulation of genetic diversity including changes in the S protein. Multiple introductions of MERS-CoV were identified and suggest lower R_0 values. MERS-CoV is evolving within the largely human-to-human Al-Hasa cluster but the phenotypic consequences of these variants needs assessment before adaption can be inferred. Transmission within Saudi Arabia seems consistent with either movement of an animal reservoir, animal products, or movement of infected people. Our results provide important and substantial genomic information on MERS-CoV and a direction for further investigation. That there were three genetically distinct lineages of MERS-CoV in Riyadh suggests it is unlikely that the Riyadh infections are the result of one, continuous human-to-human transmission chain. Further definition of the exposures responsible for the sporadic introductions of MERS-CoV into human populations is urgently needed to provide the necessary information to interrupt transmission and contain the virus.

Discussion

Our study presents the genetic analyses of the largest number of MERS-CoV genomes described so far (panel). The most important findings from this study are that at least two distinct lineages were circulating in Riyadh in October, 2012, and transmission patterns in the epidemic are consistent with both human-to-human transmission and sporadic zoonotic events. An updated evolutionary rate supports the circulation of MERS-CoV since the middle of 2011.

The 21 MERS-CoV full genome sequences provide greater detail in tracking MERS-CoV transmission. We also recreated the virus transmission pathways, with particular reference to the Al-Hasa cluster,[16] to establish whether all MERS-CoV infections in people result from a single zoonotic transfer with subsequent human-to-human transmission or if they are from many zoonotic events. Our data suggest that local spread of virus in the Al-Hasa outbreak might be more complex than previously thought, with additional sources of the virus contributing to hitherto human-to-human transmission chains. This is consistent with the multiple tree clusters described by Breban and colleagues[11] and support their optimistic R_0 scenario for MERS-CoV.

MERS-CoV from the Al-Hasa region is consistent with spread via human-to-human transmission, with the initial case derived from a zoonotic event. In addition to the Al-Hasa cluster (figure 2), multiple distinct MERS-CoV genotypes exist, possibly each from a separate zoonotic event. These include the Munich/AbuDhabi, England-Qatar, Riyadh_3 cluster, the Bisha-1, Riyadh_1 cluster, and the singletons Hafr-Al-Batin_1, Riyadh_2, Buraidah_1, and England2-HPA (Makkah). The Riyadh region shows three distinct MERS-CoV genotypes. In view of the evolutionary rate of the virus it is not expected

that a single zoonotic event was the source of the epidemic. Such high local diversity could arise from zoonotic events in Riyadh if the MERS-CoV diversity in the animal reservoir is being continuously imported from other regions. Alternately, Riyadh is the largest population centre in Saudi Arabia and therefore the largest target for human-to-human transmission. The many circulating genotypes might be a consequence of the movement of infected people from other regions, this is supported by Riyadh_1_2012 clustering with Bisha_1_2012 and Riyadh_3-2013 clustering with viruses from Doha (England_Qatar_2012) and Abu Dhabi (Munich/AbuDhabi), the two largest urban centres in neighbouring countries (figure 6).

MERS-CoV is evolving within the largely human-to-human Al-Hasa cluster but the phenotypic consequences of these variants need assessment before adaption can be inferred. All recent MERS-CoV encoded S proteins differ from the well studied EMC/2012 S protein at codon 1020; therefore, the consequences of this change should be carefully monitored.

Calculations with this larger set of MERS-CoV sequences provide an estimate of the emergence of MERS-CoV in July, 2011, with a broad credible interval (95% HPD July, 2007, to June, 2012). Analyses using short sequences suggest that the MERS-CoV virus might have an ancestor in bats.[7,10,14,26] Using the evolutionary rate from our study we show a substantial period since these viruses shared a common ancestor, suggesting that there might be an intermediary host as the source of human infections. MERS-CoV has not yet been identified in any animal sources. Therefore field studies of all probable reservoir species, including camels, bats, goats, sheep, dogs, cats, rodents, and others in Saudi Arabia and other Middle Eastern countries are ongoing. A history of contact of Saudi Arabian MERS-CoV patients with camels and goats has been reported in some cases. Serological testing for MERS-CoV has detected antibodies to the virus in camels in Oman and the Canary Islands[9] suggesting that a virus that stimulates antibody responses cross-reactive with MERS-CoV has been recently circulating among camels; however, the only way to establish if the virus eliciting the antibody response is the same as human MERS-CoV is to isolate the MERS-CoV itself in a camel. Further, the recent reports of mild and asymptomatic cases discovered through monitoring and testing of contacts of confirmed cases suggest that the focus on severe disease as a surveillance strategy might be missing substantial numbers of milder or asymptomatic cases.

Our results provide important and substantial genomic information on MERS-CoV and a direction for further investigation. That there were three genetically distinct lineages of MERS-CoV in Riyadh suggests it is unlikely that the Riyadh infections are the result of one, continuous human-to-human transmission chain. Further definition of the exposures responsible for the sporadic introductions of MERS-CoV into human populations is urgently needed to provide the necessary information to interrupt transmission and contain the virus.

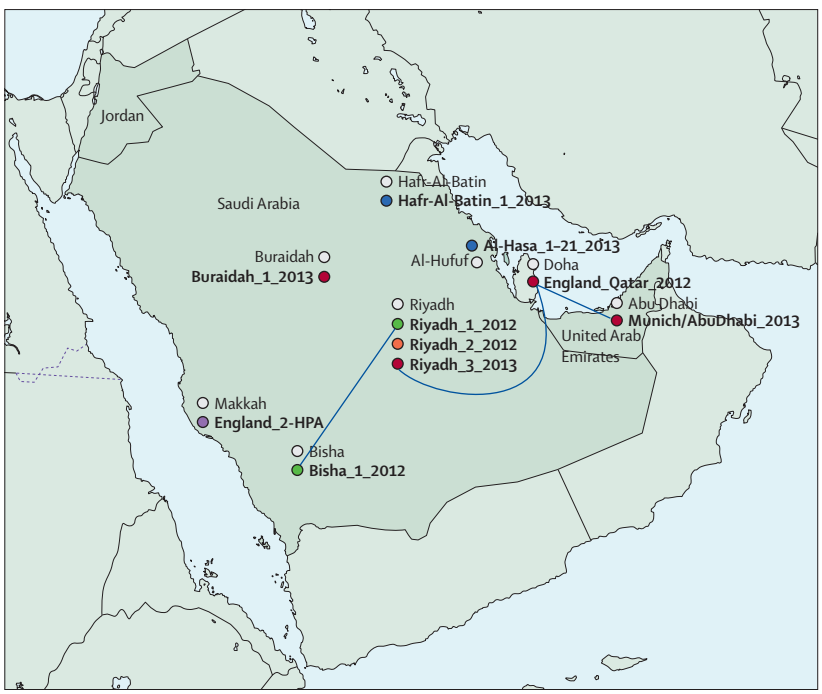

Figure 6: **Geographical distribution of genotypes**
MERS-CoV genotypes (coloured circles with genome names) are shown near the site of probable infection (white-filled circles). The 19 Al-Hasa sequences are shown by a single blue-filled circle. The genetically related genotypes from distinct locations (Bisha_1_2012, Riyadh_1_2012 and Riyadh_3_2013, England_Qatar_2012, Munich_AbuDhabi_2013) are linked with blue lines. MERS-CoV=Middle East respiratory syndrome coronavirus.

Contributors

AIZ, ZAM, and AAA-R conceived the project as part of the Global Centre for Mass Gatherings Medicine R&D initiative on MERS-CoV. ZAM, AIZ, MC, and PK designed the research. MC, SJW, ALP, AG, RB-R, and AR did the research. MC, AIZ, and PK wrote the first draft of the manuscript. MC, SJW, AIZ, AR, PK, and ZAM finalised the paper with contributions from all authors.

Conflicts of interest

We declare that we have no conflicts of interests.

Acknowledgments

The support of all staff at the Saudi Arabian Ministry of Health and the Jeddah regional laboratory is gratefully acknowledged. The sequencing work was supported by the Wellcome Trust Sanger Institute and the European Community's Seventh Framework Programme (FP7/2007–2013) under the project EMPERIE, European Community grant agreement number 223498, and under the project PREDEMICS, grant agreement number 278433. AIZ acknowledges support from the National Institute of Health Research Biomedical Research Centre, University College London Hospitals, the European and Developing Countries Clinical Trials Partnership, and the EC-FW7.

References

1 Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* 2012; **367:** 1814–20.

2 WHO. Middle East respiratory syndrome coronavirus (MERS-CoV)—update, Sept 12, 2013. http://www.who.int/csr/don/2013_08_29/en/index.html (accessed Sept 13, 2013).

3 Guan Y, Peiris JS, Zheng B, et al. Molecular epidemiology of the novel coronavirus that causes severe acute respiratory syndrome. *Lancet* 2004; **363:** 99–104.

4 Song HD, Tu CC, Zhang GW, et al. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proc Natl Acad Sci USA* 2005; **102:** 2430–35.

5 Raj VS, Mou H, Smits SL, et al. Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* 2013; **495:** 251–54.

6 Lu G, Hu Y, Wang Q, et al. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* 2013; **500:** 227–31.

7 Ithete NL, Stoffberg S, Corman VM, et al. Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. *Emerg Infect Dis* 2013; published online July 24. DOI:10.3201/eid1910.130946.

8 Rambaut A. Novel human betacoronavirus molecular evolutionary analyses. http://epidemic.bio.ed.ac.uk/coronaviruses (accessed Sept 3, 2013).

9 Reusken CB, Haagmans BL, Müller MA, et al. Middle East respiratory syndrome coronavirus neutralising serum antibodies in dromedary camels: a comparative serological study. *Lancet Infect Dis* 2013; published online Aug 8. http://dx.doi.org/10.1016/S1473-3099(13)70164-6.

10 Memish Z, Mishra N, Olival K, et al. Middle East respiratory syndrome coronavirus in bats, Saudi Arabia. *Emerg Infect Dis* 2013; published online Aug 23. DOI:10.3201/eid1911.131172.

11 Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. *Lancet* 2013; **382:** 694–99.

12 Corman V, Eckerle I, Bleicker T, et al. Detection of a novel human coronavirus by real-time reverse-transcription polymerase chain reaction. *Euro Surveill* 2012; **17:** 20285.

13 Corman V, Muller M, Costabel U, et al. Assays for laboratory confirmation of novel human coronavirus (hCoV-EMC) infections. *Euro Surveill* 2012; **17:** 20334.

14 Cotten M, Lam TT, Watson SJ, et al. Full-genome deep sequencing and phylogenetic analysis of novel human betacoronavirus. *Emerg Infect Dis* 2013; **19:** 736–42B.

15 Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19:** 455–77.

16 Assiri A, McGeer A, Perl TM, et al. Hospital outbreak of Middle East respiratory syndrome coronavirus. *N Engl J Med* 2013; **369:** 407–16.

17 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011; **28:** 2731–39.

18 Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001; **17:** 754–5.

19 Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005; **21:** 676–79.

20 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012; **29**: 1969–73.

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

22 Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 2008; **25:** 1459–71.

23 Delport W, Poon AF, Frost SD, Kosakovsky Pond SL. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* 2010; **26:** 2455–57.

24 Albarrak AM, Stephens GM, Hewson R, Memish ZA. Recovery from severe novel coronavirus infection. *Saudi Med J* 2012; **33:** 1265–69.

25 Memish ZA, Zumla AI, Al-Hakeem RF, Al-Rabeeah AA, Stephens GM. Family cluster of Middle East respiratory syndrome coronavirus infections. *N Engl J Med* 2013; **368:** 2487–94.

26 van Boheemen S, de Graaf M, Lauber C, et al. Genomic characterization of a newly discovered coronavirus associated with acute respiratory distress syndrome in humans. *MBio* 2012; **3:** e00473–12.

27 Duckert P, Brunak S, Blom N. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* 2004; **17:** 107–12.

28 Bottcher E, Matrosovich T, Beyerle M, Klenk HD, Garten W, Matrosovich M. Proteolytic activation of influenza viruses by serine proteases TMPRSS2 and HAT from human airway epithelium. *J Virol* 2006; **80:** 9896–98.

29 Matsuyama S, Nagata N, Shirato K, Kawase M, Takeda M, Taguchi F. Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. *J Virol* 2010; **84:** 12658–64.

30 Belouzard S, Chu VC, Whittaker GR. Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites. *Proc Natl Acad Sci USA* 2009; **106:** 5871–76.

31 Li W, Zhang C, Sui J, et al. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *EMBO J* 2005; **24:** 1634–43.

32 Sheahan T, Rockx B, Donaldson E, et al. Mechanisms of zoonotic severe acute respiratory syndrome coronavirus host range expansion in human airway epithelium. *J Virol* 2008; **82:** 2274–85.