Repeated introductions and intensive community transmission fueled a mumps virus outbreak in Washington State

Louise H. Moncla[1#], Allison Black[1,2#], Chas DeBolt[3], Misty Lang[3], Nicholas R. Graff[3], Ailyn C. Pérez-Osorio[3], Nicola F. Müller[1], Dirk Haselow[4], Scott Lindquist[3], Trevor Bedford*[1,2]

1. Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States

2. Department of Epidemiology, University of Washington, Seattle, Washington, United States

3. Office of Communicable Disease Epidemiology, Washington State Department of Health, Shoreline, Washington, United States

4. Arkansas Department of Health, Little Rock, Arkansas, United States

[#] these authors contributed equally to this work

* corresponding author

Abstract

In 2016/2017, Washington State experienced a mumps outbreak despite high childhood vaccination rates, with cases more frequently detected among school-aged children and members of the Marshallese community. Sequencing 166 mumps genomes revealed that mumps was introduced into Washington approximately 13 times, primarily from Arkansas, sparking multiple, co-circulating transmission chains. Neither vaccination status nor age were strong determinants of transmission. Instead, the outbreak in Washington was overwhelmingly sustained by transmission within the Marshallese community. Our findings underscore the utility

of genomic data to clarify epidemiologic factors driving transmission, and pinpoint contact networks as critical determinants of mumps transmission in Washington.

Introduction

In 2016 and 2017, mumps virus swept the United States in the country's largest outbreak since the pre-vaccine era [1]. Washington State was heavily affected, reporting 889 confirmed and probable cases. Longitudinal studies [2], epidemiologic outbreak investigations [3], and epidemic models [4] suggest that mumps vaccine-induced immunity wanes over 13-30 years, consistent with the preponderance of young adult cases in recent outbreaks. Like with other recent mumps outbreaks, most Washington cases in 2016/17 were vaccinated. Unusually though, incidence was highest among children aged 10-18 years, younger than expected given waning immunity. The outbreak was also peculiar in that approximately 52% of the total cases were Marshallese, an ethnic community that comprises ~0.3% of Washington's population. These same phenomena were also observed in Arkansas. Of the 2,954 confirmed and probable Arkansas cases, 57% were Marshallese, and 57% of cases were children aged 5-17 [5]. Amongst school-aged children in Arkansas and Washington, >90% had previously received 2 doses of MMR vaccine [5]. The high proportion of vaccinated cases, younger-than-expected age at infection, disproportionate impact on the Marshallese community, and epidemiologic link to Arkansas suggest that factors beyond waning immunity are necessary to explain mumps transmission during this outbreak in Washington.

Between 1947 and 1986, the United States occupied the Republic of Marshall Islands and detonated the equivalent of >7000 Hiroshima size nuclear bombs as part of its nuclear testing program[6]. The effects were devastating, precipitating widespread environmental destruction,

nuclear contamination, and dire health consequences [7–11]. Marshallese individuals inhabiting the targeted atolls were forcibly moved to other islands, and many were exposed to nuclear fallout [12] that persists on the Islands today [13]. Marshallese individuals living on and off the Islands experience significant health disparities including a higher burden from infectious diseases and chronic health conditions [14–16]. Compounding these disparities, Marshallese individuals are specifically excluded from Medicaid eligibility despite legal residency in the US permitted under the Compact of Free Association (COFA) Treaty. As a result, many US-residing Marshallese are uninsured, with poor access to healthcare [17].

We used a genomic epidemiological approach to investigate patterns of mumps transmission in Washington. We sequenced 110 mumps viral genomes obtained from specimens collected from laboratory-confirmed mumps cases in Washington State and another 56 from other US states collected between 2006 and 2018. We applied phylodynamic methods to these sequence data to investigate mumps introductions into and spread within Washington State with primary focus on age, demographics and vaccination history as explanatory variables for transmission.

Methods

Mumps surveillance in Washington State

Mumps is a notifiable condition in Washington State. Therefore, per the Washington Administrative Code (WAC), as specified in WAC Chapter 246-101, healthcare providers, healthcare facilities, and laboratories must report cases of mumps or possible mumps to the local health jurisdiction (LHJ) of the patient's residence. LHJ staff initiate case investigations and facilitate optimal collection and testing of diagnostic specimens. Buccal swabs and urine are acceptable specimens for real-time reverse transcription polymerase chain reaction (qRT-PCR),

a preferred diagnostic test for mumps. Most mumps rRT-PCR tests for Washington State residents are performed at the Washington State Public Health Laboratories, where all positive specimens are archived.

Individuals testing positive for mumps ribonucleic acid (RNA) by qRT-PCR are classified as confirmed mumps cases if they have a clinically-compatible illness (i.e., an illness involving parotitis or other salivary gland swelling lasting at least 2 days, aseptic meningitis, encephalitis, hearing loss, orchitis, oophoritis, mastitis, or pancreatitis). During case investigations, case-patients or their proxies are interviewed. Information about demographics, illness characteristics, vaccination history, and potential for exposure to and transmission of mumps are solicited from each case-patient. The Washington State Department of Health (DOH) receives, organizes, performs quality control on, and analyzes data from LHJ case reports and supports investigations upon request.

Community feedback

In order to ensure that this study was faithful to the experience of the Marshallese community in Washington State, we sought paid consultation from a local Marshallese community health advocate. We conducted video and telephone interviews to directly address the impacts of mumps transmission on the Marshallese community, community healthcare goals and priorities, and the impacts of the mumps outbreak on stigmatization. This feedback informed what is being presented herein, provided crucial context for understanding mumps transmission, and allowed us to work with the community to determine how best to discuss Marshallese involvement in the outbreak.

Sample collection and IRB approval

This study was approved by the Fred Hutchinson Cancer Research Center (FHCRC) Institutional Review Board (IR File #: 6007-944) and by the Washington State Institutional Review Board, and classified as not involving human subjects. Samples were selected for sequencing to maximize temporal and epidemiologic breadth and to ensure successful sequencing. As such, samples were chosen based on the date of sample collection, the PCR cycle threshold (Ct) (targeting samples with Cts < 30), case vaccination status, and community status (Marshallese or non-Marshallese). All metadata, including case vaccination status were transferred from WA DOH to FHCRC in a de-identified form.

We also sequenced an additional set of 56 samples collected in Wisconsin, Ohio, Missouri, Alabama, and North Carolina provided by the Wisconsin State Laboratory of Hygiene. 10 of these samples were collected in Wisconsin during the 2006/2007 Midwestern college campus outbreaks, 6 samples were collected in 2014, and the rest were collected between 2016 and 2018. For these samples, we received metadata describing sample Ct value and date of collection. All metadata were received by FHCRC in de-identified form.

Viral RNA extraction, cDNA synthesis, and amplicon generation

Viral RNA was extracted from buccal swabs using either the QiAmp Viral RNA Mini Kit (Qiagen, Valencia, CA, USA) or the Roche MagNA Pure 96 DNA and viral NA small volume kit (Roche, Basel, Switzerland). For samples extracted with the QiAmp Viral RNA Mini Kit, 500 µl of buccal swab fluid was spun at 5000 x g for 5 minutes at 4°C to pellet host cells. The supernatant was then removed and centrifuged at 14,000 rpm for 90 minutes at 4°C to pellet virions. Excess fluid was discarded, and the peletted virions were resuspended in 150-200 µl of fluid. Resuspended viral particles were then used as input to the QiAmp Viral RNA Mini Kit (Qiagen, Valencia, CA,

122 USA), following manufacturer's instructions, and eluting in 30 µl of buffer AVE. For extraction
123 with the MagNA Pure, we followed manufacturer's instructions.
124
125 cDNA was generated with the Protoscript II First strand synthesis kit (New England Biolabs,
126 Ipswich MD, USA), using 8 µl of vRNA as input and priming with 2 µl of random hexamers.
127 vRNA and primers were incubated at 65°C for 5 minutes. Following this incubation, 10 µl of
128 Protoscript II reaction mix (2x) and 2 µl of Protoscript II enzyme mix (10x) were added to each
129 reaction and incubated at 25°C for 5 minutes, then 42°C for 1 hour, followed by a final
130 inactivation step at 80°C for 5 minutes. To amplify the full mumps genomes, we used Primal
131 Scheme (http://primal.zibraproject.org/) to design overlapping, ~1500 base pair amplicons
132 spanning the entirety of the mumps virus genome. Each tiled set of primes overlapped by ~100
133 base pairs. We used the following primer sequences:
134
135 mumps_1.5kb_1F (Forward): ACCAAGGGGAAAATGAAGATGGG, pool 1
136 mumps_1.5kb_1R (Reverse): TAACGGCTGTGCTCTAAAGTCAT, pool 1
137 mumps_1.5kb_2_F (Forward)TTGTTGACAGGCTTGCAAGAGG, pool 2
138 mumps_1.5kb_2_R (Reverse): TTGTTCAAGATGTTGCAGGCGA, pool 2
139 mumps_1.5kb_3_F (Forward): TGCAACCCCATATGCTCACCTA, pool 1
140 mumps_1.5kb_3_R (Reverse): AGTTTGTTCCTGCCTTTGCACA, pool 1
141 mumps_1.5kb_4_F (Forward): AGTGAGAGCAGTTCAGATGGAAGT, pool 2
142 mumps_1.5kb_4_R (Reverse): CCCTCCATTAGACCGGCACTTA, pool 2
143 mumps_1.5kb_5_F (Forward): AACAACAGTGTTCCAGCCACAA, pool 1
144 mumps_1.5kb_5_R (Reverse): GGTGGCACTGTCCGATATTGTG, pool 1
145 mumps_1.5kb_6_F (Forward): TGCCGTTCAATCATGAGACATAAAGA, pool 2
146 mumps_1.5kb_6_R (Reverse): CGTAGAGGAGTTCATACGGCCA, pool 2

mumps_1.5kb_7_F (Forward): TGTCTGTGCCTGGAATCAGATCT, pool 1

mumps_1.5kb_7_R (Reverse): CGTCCTTCCAACATATCAGTGACC, pool 1

mumps_1.5kb_8_F (Forward): CCAAAAGACAGGTGAGTTAACAGATTT, pool 2

mumps_1.5kb_8_R (Reverse): ACGAGCAAAGGGGATGATGACT, pool 2

mumps_1.5kb_9_F (Forward): TTTGGCACACTCCGGTTCAAAT, pool 1

mumps_1.5kb_9_R (Reverse): TGACAATGGTCTCACCTCCAGT, pool 1

mumps_1.5kb_10_F (Forward): ACTCGCACAGTATCTATTAGATCGTGA, pool 2

mumps_1.5kb_10_R (Reverse): GCCCAGCCAGAGTAAACAAACA, pool 2

mumps_1.5kb_11_F (Forward): GCCAAGCAGATGGTAAACAGCA, pool 1

mumps_1.5kb_11_R (Reverse): GGCTCTCTCCAACATGCTGTTC, pool 1

mumps_1.5kb_12_F (Forward): GCGGGGCCTCTATGTCACTTAT, pool 2

mumps_1.5kb_12_R (Reverse): CCAAGGGGAGAAAGTAAAATCAAT, pool 2

Primers were pooled into 2 pools as follows: the first contained primer pairs 1, 3, 5, 7, 9, and 11, all pooled at 10 uM. The second pool contained primer pairs 2, 4, 6, 8, 10, and 12. All primers in pool 2 were pooled at 10 uM, except for primer pair 4, which was added at a 20 uM concentration. PCR was performed with the Q5 Hotstart DNA polymerase (New England Biolabs, Ipswich, MD, USA), using 11.75 µl of nuclease-free water, 5 µl of Q5 reaction buffer, 0.5 µl of 10 mM dNTPs, 0.25 µl, 2.5 of pooled primers, and 5 µl of cDNA. Amplicons were generated with the following PCR cycling conditions: 98°C for 30 seconds, followed by 30 cycles of: 98°C for 15 seconds, then 67°C for 5 minutes. Cycling was concluded with a 10°C hold. PCR products were run on a 1% agarose gel, and bands were cut out and purified using the QiAquick gel extraction kit (Qiagen, Valencia, CA, USA), following the manufacturer's protocol. All optional steps were performed, and the final product was eluted in 30 μl of buffer EB. For samples extracted on the MagNA Pure, amplicons were cleaned using a 1x bead cleanup with

Ampure XP beads. Final cleaned amplicons were quantified using the Qubit dsDNA HS Assay kit (Thermo Fisher, Waltham, MA, USA).

Library preparation and sequencing

For each sample, pool 1 and pool 2 amplicons were combined in equimolar concentrations to a total of 0.5 ng in 2.5 μl. Libraries were prepared using the Nextera XT DNA Library Prep Kit (Illumina, San Diego, CA, USA), following manufacturer's instructions, but with reagent volumes halved for each step, for the majority of samples in our dataset. For samples processed in our last sequencing run, several samples had higher Ct values. We therefore chose to process these samples using the standard 1x reagent volumes for the library preparation step. All libraries were purified using Ampure XP beads (Beckman Coulter, Brea, CA, USA), using a 0.6x cleanup, a 1x cleanup, and a final 0.7x cleanup. At each step, beads were washed twice with 200-400 μl of 70% ethanol. The final product was eluted off the beads with 10 μl of buffer EB. Tagmentation products were quantified with the Qubit dsDNA HS Assay kit (Thermo Fisher, Waltham, MA, USA), and run on a Tapestation with the TapeStation HighSense D5K assay (Agilent, Santa Clara, CA, USA) to determine the average fragment length. All but 8 samples and negatives were pooled together in 6 nM libraries and run on 300 bp x 300 bp v3 kits on the Illumina MiSeq, with a 1% spike-in of PhiX. The remaining 8 samples (MuVs/Washington.USA/1.17/FH77[G], MuVs/Washington.USA/12.17/FH78[G], MuVs/Washington.USA/16.17/FH79[G], MuVs/Washington.USA/19.17/FH80[G], MuVs/Washington.USA/20.17/FH81[G], MuVs/Washington.USA/20.17/FH82[G], MuVs/Washington.USA/29.17/FH83[G], and MuVs/Washington.USA/2.17/FH84[G]) were pooled to a 1.2 nM library, and run as a 50 pM library with 2% PhiX on the Illumina iSeq, with a 151 bp x 151 bp v3 kit.

Negative controls

A negative control (nuclease-free water) was run for each viral RNA extraction, reverse transcription reaction, and for each pool for each PCR reaction. These negative controls were carried through the library preparation process and sequenced alongside actual samples. Any samples whose negative controls from any step in the process resulted in >10x mumps genome coverage were re-extracted and sequenced.

Bioinformatic processing of sequencing reads

Human reads were removed from raw FASTQ files by mapping to the human reference genome GRCH38 with bowtie2 [18] version 2.3.2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Reads that did not map to the human genome were output to separate FASTQ files and used for all subsequent analyses. Illumina data was analyzed using the pipeline described in detail at https://github.com/lmoncla/illumina_pipeline. Briefly, raw FASTQ files were trimmed using Trimmomatic [19] (http://www.usadellab.org/cms/?page=trimmomatic), trimming in sliding windows of 5 base pairs and requiring a minimum Q-score of 30. Reads that were trimmed to a length of <100 base pairs were discarded. Trimming was performed with the following command: java -jar Trimmomatic-0.36/trimmomatic-0.36.jar SE input.fastq output.fastq SLIDINGWINDOW:5:30 MINLEN:100. Trimmed reads were mapped to a consensus sequence from Massachusetts (Genbank accession: MF965301) using bowtie2[18] version 2.3.2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml), using the following command: bowtie2 -x reference_sequence.fasta -U read1.trimmed.fastq,read2.trimmed.fastq -S output.sam --local. Mapped reads were imported into Geneious (https://www.geneious.com/) for visual inspection and consensus calling. To avoid issues with mapping to an improper reference sequence, we then remapped each sample's trimmed FASTQ files to its own consensus sequence. These bam files were again manually inspected in Geneious, and a final consensus sequence was

called, with nucleotide sites with <20x coverage output as an ambiguous nucleotide ("N"). All genomes with >50% Ns were discarded. All genomes with >50% coverage are available at https://github.com/blab/mumps-seq/tree/master/data.

General availability of analysis software and data

All code used to analyze data, input files for BEAST, and all code used to generate figures for this manuscript are publicly available at https://github.com/blab/mumps-wa-phylodynamics. Raw FASTQ files with human reads removed are available under SRA project number PRJNA641715. All protocols for generating sequence data as well as the consensus genomes are available at https://github.com/blab/mumps-seq. Consensus genomes have also been deposited to Genbank under accessions MT859507-MT859672.

Dataset curation and divergence tree generation

We downloaded all currently available, complete mumps genomes from North America from the NIAID Virus Pathogen Database and Analysis Resource (ViPR) [20] through http://www.viprbrc.org/. We also obtained mumps genomes from British Columbia, Ontario, and Arkansas from the Wisconsin State Laboratory of Hygiene. We obtained written permission from sequence authors for any sequence that had not previously been published on. Sequences and metadata were cleaned and organized using fauna, a database system that is part of the Nextstrain platform. Sequences were processed using Nextstrain's augur software [21], and filtered to include only those with at least 8,000 bases and were sampled in North America in 2006 or later. Genomes were aligned with MAFFT [22], and trimmed to the reference sequence (MuV/Gabon/13/2[G], GenBank accession: KM597072). We inferred a maximum likelihood phylogeny using IQTREE [23], and inferred a molecular clock and temporally-resolved phylogeny using TreeTime [24]. Sequences with an estimated clock rate that deviated from the other

sequences by >4 times the interquartile distance were removed from subsequent analysis. We inferred the root-to-tip distance with TempEst version 1.5.1 [25] with the best fitting root by the heuristic residual mean squared function.

Phylogenetic analysis of full North American mumps genomes

Using the same set of genome sequences used for divergence tree estimation, we aligned sequences with MAFFT and inferred time-resolved phylogenies in BEAST version 1.8.4 [26]. We used a skygrid population size prior with 100 bins, and a skygrid cutoff of 25 years, allowing us to estimate 4 population sizes each year. We used an HKY nucleotide substitution model with 4 gamma rate categories, and a strict clock with an exponential prior with a mean of 4.31×10^{-4}. This mean clock rate was set based on the inferred substitutions per site per year from all North American mumps genomes on nextstrain.org/mumps/na. We used a discrete trait model [27] and estimated migration rates using BSSVS and ancestral states with 27 geographic locations. Here, “state” refers to the inferred ancestral identity of an internal node, where the inferred identity could be any of the 27 geographic locations (US states and Canadian provinces) in the dataset. For the prior on non-zero rates for BSSVS, we specified a Poisson distribution with mean 0.69 with an offset of 26. As a prior on each pairwise migration rate, we used an exponential distribution with mean 1. All other priors were left at default values. We ran this analysis for 100 million steps, sampling every 10,000, and removed the first 10% of sampled states as burnin. A maximum clade credibility tree was summarized with TreeAnnotator, using the mean heights option. All tree plotting was performed with baltic (https://github.com/evogytis/baltic). Input XML files and output results are available at https://github.com/blab/mumps-wa-phylodynamics/tree/master/phylogeography.

Testing for descendants in divergence trees

To determine whether specific groups were more likely to be found at the beginnings of transmission chains than other groups, we developed a statistic to quantify transmission in the tree. Using the tree JSON output from the Nextstrain pipeline [21], we traversed the tree from root to tip. For each tip that lay on an internal node, i.e., had a branch length of nearly 0 ($< 1 \times 10^{-16}$), we counted the number of descendants. We collapsed very small branches (those with branch lengths less 1×10^{-16}) to obtain polytomies. We then classified tips as either having descendants (i.e., the number of descendents was > 0) or not having descendants. Here, we define a "descendant" as a tip that occurs in any downstream portion of the tree, i.e., it falls along the same lineage but to the right of the parent tip. A diagram of what we classify as "descendant tips" is shown in Figure 4a. The probability of having descendants was evaluated as a function of community status, age, and vaccination status with logistic regression as described below.

For each Washington tip in the tree, we classified it as either having descendants (coded as a 1) or not having descendants (coded as 0). For each tip, we coded it's corresponding age, vaccination status, and community membership as a predictor variable input into a logistic regression model. We coded these attributes as follows: For community membership, non-Marshallese tips were coded as 0 and Marshallese tips were coded as 1. Age was coded as a single, continuous variable. In our dataset, there were 3 classifications for vaccination status: up-to-date, not up-to-date, and unknown vaccination status. According to the Advisory Committee on Immunization Practices (ACIP) [28], individuals aged 5-18 had to have received both recommended doses of mumps-containing vaccine, children aged 15 months to 5 years required 1 dose of mumps-containing vaccine, and adults over 18 had to have received at least 1 dose of mumps-containing vaccine to be classified as up-to-date for mumps vaccination. Individuals under 15 months are considered up-to-date without any doses of mumps-containing vaccine. Not up-to-date individuals are those with a known vaccination status who did not

qualify under criteria to be classified as up-to-date. Individuals who could not provide documentation regarding their MMR vaccination history were considered to have "unknown" vaccination status. Individuals with "known" vaccination status could either be fully up-to-date, undervaccinated, or unvaccinated. To ensure that we measured the effect of vaccination among individuals who knew their vaccination status, we coded vaccination information using two dummy variables in our logistic regression, one signifying whether vaccination status was known or not, and one indicating whether vaccination was up-to-date or not. We then fit a logistic regression model to this data using the glm package in R (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm), specifying a binomial model as

$\Pr(\text{having descendants}) \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, where x_1 represents 0 or 1 value for member of Marshallese community (Not Marshallese coded as a 0, Marshallese coded as a 1), x_2 represents the numeric value of age, x_3 represents 0 or 1 value for whether vaccination status is unknown (having a known vaccination status coded as a 0, having an unknown vaccination status coded as a 1) and x_4 represents 0 or 1 value for whether vaccination status is up-to-date (up-to-date coded as a 0 and not up-to-date coded as a 1). Under this formulation, an individual with unknown vaccination status would be coded as x_3=1, x_4=0, an individual who is up-to-date would be coded as x_3=0, x_4=0, and an individual who is not up-to-date is coded as x_3=0, x_4=1. This encoding allows us to evaluate the effects of having an unknown vaccination status and a vaccination status that is not up-to-date. Age was normalized such that values fall between 0 and 1 by: (x_2 - minimum age in dataset) / (maximum age in dataset - minimum age in dataset).

P-values were assigned via a Wald test, and inferred coefficients were exponentiated to return odds ratios. All code used to parse the divergence tree and formulate and fit the regression

model are available at https://github.com/blab/mumps-wa-phylodynamics/blob/master/divergence-tree-analyses/Regression-analysis-on-descendants-in-divergence-tree.ipynb.

Rarefaction analysis to estimate transmission clusters

Using the full set of North American mumps sequences, we designated all non-Washington North American sequences as "background" sequences. We then separated Washington sequences into Marshallese tips (57 total sequences) and non-Marshallese tips (52 total sequences). For this analysis, we excluded the genotype K sequence in our dataset due to its extreme divergence from other viruses sampled in Washington, which were all genotype G. For each group (Marshallese vs. non-Marshallese), we then generated subsampled datasets comprised of a random sample of 1 to *n* sequences, where *n* is the number of total sequences available for that group. For each number of sequences, we performed 10 independent subsampling trials. Subsampling was performed without replacement. So, for community members, we generated 10 datasets in which 1 community member sequence was sampled, then 10 datasets in which 2 community members sequences were sampled, etc. up to 10 datasets in which all 57 community members sequences were sampled. For each subsampled dataset, we then combined these subsampled datasets with the background North American sequences, and reran the Nextstrain pipeline. For each subsample and trial, we infer geographic transmission history across the tree and enumerate the number of introductions into Washington. For each number of sequences tested, *n*, we report the number of trials resulting in that number of inferred introductions, and the mean number of inferred introductions across the 10 trials.

Inference of community transmission dynamics using a structured coalescent model

To infer the rates of migration between community and non-community members and to infer ancestral states of Washington internal nodes, we employed a structured coalescent model. Here, "state" refers to the inferred ancestral identity of an internal node, where the identity could be inferred as "Marshallese" or "not Marshallese". The multitype tree model [29] in BEAST 2 v2.6.2 [30] infers the effective population sizes of each deme and the migration rates between them. Because the multitype tree model requires that all partitions contain all demes, we could only analyze 4 clades that circulated in Washington State and included both Marshallese and non-Marshallese tips. We generated an XML in BEAUti v2.6.2 with 4 partitions, and linked the clock, site, and migration models. We used a strict, fixed clock, set to 4.17×10^{-4} substitutions per site year, and used an HKY substitution model with 4 gamma-distributed rate categories. We chose this substitution rate based on the clock rate estimated from the North American tree inferred with IQTree and Treetime. Migration rates were inferred with the prior specified as a truncated exponential distribution with a mean of 1 and a maximum of 50. Effective population sizes were inferred with the prior specified as a truncated exponential distribution with a mean of 1, a minimum value of 0.001, and a maximum value of 10,000. All other priors were left at default values. In order to improve convergence, we employed 3 heated chains using the package CoupledMCMC [31], where proposals for chains to swap were performed every 100 states. The analysis was run for 100 million steps, with states sampled every 1 million steps. We ran this analysis 3 independent times, and combined log and tree file output from those independent runs using LogCombiner, with the first 10% (1000 states) of each run discarded as burnin. We then summarized these combined output log and tree files. A maximum clade credibility tree was inferred using TreeAnnotator with the mean heights option. To ensure that results were not appreciably altered by the migration rate prior, we also repeated these analyses with migration rates inferred with the prior specified as a truncated exponential distribution with a mean of 10 and a maximum of 50.

Although our complete dataset contains approximately equal numbers of sequences from Marshallese and non-Marshallese cases, the 4 clusters analyzed above are enriched among Marshallese tips. To assess the impact of uneven sampling within these clusters on ancestral state inference, we performed a subsampling analysis. For each cluster, we subsampled down the number of Marshallese tips to be equal to the number of non-Marshallese tips, and reran the analysis as above. While the original analysis used 4 subclades containing both Marshallese and non-Marshallese tips, one of these subclades only has 5 tips. Subsampling this particular subtree would have resulted in a subtree with only 2 tips, thus we excluded this clade from the subsampling analysis. For this sensitivity analysis, the 3 subsampled datasets had the following tip composition: primary outbreak clade: 26 Marshallese and 26 non-Marshallese tips; 10-tip introduction: 3 Marshallese and 3 non-Marshallese tips; 8-tip introduction: 4 Marshallese and 4 non-Marshallese tips. We generated 3 randomly subsampled datasets, and for each one ran 3 independent chains, with each chain run for 50 million steps, sampling every 500,000. For one of the subsampled datasets, none of the chains converged after 20 days. In each of the remaining 2 subsampled datasets, 2 out of 3 chains converged. We combined these converged chains using LogCombiner, with the first 10% of each run discarded as burn-in. We then summarized these combined output log and tree files, and inferred a maximum clade credibility tree using TreeAnnotator with the mean heights option.

The analysis as described above assumes that each introduction into Washington State is an independent observation of the same structured coalescent process, and that the dataset represents a random sample of the underlying population. Additionally, this approach requires *a priori* definition of which sequences are part of the same Washington State transmission chain. Finally, the above analysis could only make use of the 4 Washington introductions with both

Marshallese and non-Marshallese tips, and excludes other transmission chains. Because of these issues, we supplemented the above approach with an additional analysis using the approximate structured coalescent [32] in MASCOT [33]. Using all of the Washington sequences, we specified three demes: Marshallese in Washington, non-Marshallese in Washington, and outside of Washington. To account for any transmission that happened outside of Washington State, the "outside of Washington" deme acted as a "ghost deme" from which we did not use any samples. The effective population size of this "outside of Washington" deme then describes the rate at which lineages between any location outside of Washington share a common ancestor. Including specific samples from outside of Washington would bias the inferred effective population size towards the coalescent rates of the sampled locations, by incorporating local transmission dynamics of other locations. We then estimated migration rates and effective population sizes for all 3 demes, but fixed the migration rates such that the unsampled deme ("outside of Washington") could only act as a source population. This is motivated by not having observed obvious migration out of Washington State in our previous analysis here. We ran this analysis for 10 million steps, sampling every 5000, and discarded the first 10% of states as burnin.

Results

Outbreak characteristics and dataset composition

We generated near-complete genome sequences for 110 PCR-positive mumps samples collected throughout Washington State during 2016/2017, and 56 samples collected in Wisconsin, Ohio, Missouri, Alabama, and North Carolina between 2006 and 2018. The Washington State outbreak began in October 2016, and peaked in winter of 2017, culminating in 889 confirmed and probable cases across Washington (**Fig. 1**). Individuals aged <1 to 64

419 years were affected, but the highest rate of infection occurred in children aged 10-14 (44.9
420 cases per 100,000) and 15-19 (47.0 per 100,000) (**Supplemental Table 1**). Among individuals
421 aged 5-19 years old, 91% were considered up-to-date on mumps vaccine. Adults in the age
422 group most likely to be parents of school aged children 20-39 were infected at a rate of only
423 12.9 cases per 100,000, but comprised a significant proportion (29%) of total cases
424 (**Supplemental Table 1**). While Marshallese individuals comprise only ~0.3% of Washington's
425 total population, they accounted for 52% of reported mumps cases (**Supplemental Table 2**).
426 Among Marshallese individuals aged 5-19, 93% were up-to-date on vaccination, suggesting that
427 this over-representation is not attributable to poor vaccine coverage.
428
429 **Outbreaks across North America are related**
430 We combined our sequence data with publicly available full genome sequences sampled from
431 North America between 2006 and 2018, and built a time-resolved phylogeny, inferring migration
432 history among 26 US states and Canadian provinces (**Fig. 2, Supplemental Figure 1**).
433 Sequences from samples collected between 2006 and 2014 clustered with other North
434 American mumps viruses sampled from the same times. Except for 2 sequences (one from
435 Wisconsin in 2006, genotype A, and one from Washington in 2017, genotype K, both excluded
436 from **Fig. 2**), all samples in our dataset were genotype G viruses. Ten Washington sequences
437 were highly divergent from other North American genotype G viruses, with a time to the most
438 recent common ancestor (TMRCA) of ~22 years (**Fig. 2**). The remaining Washington sequences
439 nest within the diversity of other North American viruses, and descend from the same mumps
440 lineage that has circulated in North America since 2006 (**Fig. 2**). We observe substantial
441 geographic mixing along the tree. While viruses from Massachusetts (dark green tips and
442 branches) seeded outbreaks in the Northeast and Midwest, we also infer transmission from
443 Massachusetts to Texas, Louisiana, Alabama, and British Columbia. Despite the close

444 geographic proximity between British Columbia and Washington, most British Columbia
445 sequences form a distinct cluster on a long branch (**Fig. 2**), suggesting seeding from an
446 unsampled location. Although viruses from Washington are scattered throughout the phylogeny,
447 most cluster within a clade of viruses sampled in Arkansas (**Fig. 2**).
448
449 **Mumps was introduced into Washington multiple independent times**
450 Estimating the number and timing of viral introductions is important for estimating epidemiologic
451 parameters and evaluating surveillance networks, but is challenging with case count data alone.
452 The Washington Department of Health had identified a single potential index case in October of
453 2016. To determine whether the genomic data supported a single introduction, we separated
454 each introduction inferred in the maximum clade credibility tree and plotted each as its own
455 transmission chain (**Fig. 3a**). We then enumerated the number of transitions into Washington in
456 each tree in the posterior set, and plotted the distribution of Washington introductions consistent
457 with the phylogeny (**Fig. 3b**).
458
459 Genomic data show that mumps was introduced into Washington State approximately 13
460 independent times (95% highest posterior density, HPD: 11 - 16), from geographically disparate
461 locations (**Fig. 3**). Ten sampled tips descend from long branches (~22 years), suggesting likely
462 transmission from unsampled geographic locations. We infer introductions from Ontario and
463 Missouri that each lead to 1-3 sampled cases (**Fig. 3b**), suggesting limited onward transmission
464 following these introductions. In contrast, 4 introductions from Arkansas account for 92/110
465 sequenced cases, suggesting that these introductions led to more sustained chains of
466 transmission following introduction (**Fig. 3b**). We refer to the largest cluster as the “primary
467 outbreak clade,” and infer its introduction from Arkansas to Washington around August of 2016
468 (August 7, 2016, 95% HPD: July 11, 2016 to September 19, 2016, **Fig. 3b**), 3.5 months before

Washington's first reported case. These data reveal that what had appeared to be a single outbreak based on case surveillance data was in fact a series of multiple introductions, primarily from Arkansas, sparking overlapping and co-circulating transmission chains.

SH gene sequencing is insufficient for fine-grained geographic inference

Mumps virus surveillance and genotyping relies on the SH gene [34], a short, 316 bp gene that is simple and rapid to sequence. To determine whether SH gene sequencing would have produced similar results, we built a divergence tree using our set of North American full genomes (**Supplemental Figure 2a**), then truncated that data to include only SH gene sequences (**Supplemental Figure 2b**). Almost all North American SH genes were identical, resulting in a single, large polytomy (**Supplemental Figure 2b**). This indicates that SH sequences lack sufficient resolution to elucidate fine-grained patterns of geographic spread.

Quantifying differences in transmission patterns within Washington

In both Arkansas and Washington, Marshallese individuals comprised over 50% of mumps cases, despite accounting for a much lower proportion of the population in both states. Phylogenetic reconstruction links the outbreaks in Washington and Arkansas, placing most sampled mumps genomes in Washington as descendant from Arkansas. We sought to investigate how mumps transmission may have differed within Marshallese and non-Marshallese communities within the same outbreak. Phylogenetic trees reflect the transmission process and can be used to quantify differences in transmission patterns among population groups. If transmission rates were distinct between Marshallese and non-Marshallese mumps cases, we expect the following: 1. Sequences from the high-transmitting group should be more frequently detected upstream in transmission chains. 2. Introductions seeded into the high-transmitting group should result in larger and more diverse clades in the tree. 3. The internal

494 nodes of the phylogeny should be predominantly composed by members of the high-
495 transmitting group, while members of the low-transmitting group should primarily be found at
496 terminal nodes.
497
498 **Marshallese cases are enriched upstream in transmission chains**
499 We developed a transmission metric to quantify whether Marshallese cases were enriched at
500 the beginnings of successful transmission chains. We traverse the full genome divergence
501 phylogeny (**Supplemental Figure 2a**) from root to tip. When we encounter a tip that lies on an
502 internal node, we enumerate the number of tips that descend from its parent node. We then
503 classify each tip in the phylogeny as either having descendants or not, and compare the
504 proportion of tips with and without descendants among groups (**Fig. 4a**, see Methods for more
505 details). Given our sampling proportion (110 sequences/889 total cases, ~12%), we do not
506 expect to have captured true parent/child infection pairs. Rather, we expect to have
507 preferentially sampled long, successful transmission chains within the state. This allowed us to
508 assess whether infections with particular attributes (community membership, vaccination status,
509 age) are predictors for being upstream in these chains, and thus associated with sustained
510 transmission. We evaluated the probability of having descendants in the tree as a function of
511 vaccination status, age, and community status with logistic regression (see Methods for details
512 and full model). Neither age nor vaccination status were significantly associated with the
513 presence of downstream tips in the tree (**Supplemental Table 3**). However, Marshallese cases
514 were significantly more likely to have downstream descendants than non-Marshallese cases
515 (odds ratio = 3.2, p = 0.00725, **Supplemental Table 3**). While only 27% (14/52) of non-
516 Marshallese tips were ancestral to downstream samples, 56% (32/57) of Marshallese tips had
517 downstream descendents. These results suggest that community membership was a significant
518 determinant of sustained transmission while controlling for vaccination status and age.

519

520 **Longer transmission chains are associated with community status**

521 In the absence of recombination, closely-linked infections will cluster together on the tree, while
522 unrelated infections should fall disparately on the tree, forming multiple smaller clusters. We
523 inferred the number of Washington-associated clades in the tree as a function of whether
524 sampled infections came from Marshallese or non-Marshallese individuals. Using the full North
525 American phylogeny, we removed all Washington sequences and separated them into viruses
526 sampled from cases noted as Marshallese or non-Marshallese. Then, separately for each
527 group, we added sequences back into the tree one by one, until all sequences for that group
528 had been added. For each number of sequences, we performed 10 independent trials (see
529 Methods for complete details), and at each step, enumerated the number of inferred
530 Washington clusters in the phylogeny. For comparison, we also grouped tips by vaccination
531 status and repeated this analysis.

532

533 For tips from non-Marshallese individuals, the number of inferred clusters increases linearly as
534 tips are added to the tree (**Fig. 4b**). This suggests that these infections are not directly related,
535 and are not part of sustained transmission chains (**Fig. 4b**). In contrast, the number of inferred
536 clusters for Marshallese tips stabilizes after ~10 tips are added, even as almost 50 more
537 sequences are added to the tree. This pattern likely arises because many Marshallese
538 infections are part of the same long transmission chain, such that newly added tips nest within
539 existing clusters. We do not observe similar differences among vaccination groups
540 (**Supplemental Figure 3**). These findings are consistent with distinct patterns of transmission
541 among Marshallese versus non-Marshallese cases: transmission among Marshallese
542 individuals resulted in a small number of large clusters, while transmission among non-

Marshallese individuals are generally the result of disparate introductions that generate shorter transmission chains.

We next separated each Washington introduction and colored each tip by community membership. Every introduction that was not seeded from Arkansas led to exclusively non-Marshallese infections, while introductions from Arkansas were longer and enriched with Marshallese tips (**Fig. 4c**). The primary outbreak clade is particularly enriched, containing 43 Marshallese tips and 26 non-Marshallese tips, hinting that transmission chains are longer when Marshallese cases are present in a cluster.

Mumps transmitted efficiently within the Marshallese community

Internal nodes on a phylogeny represent ancestors to subsequently sampled tips, while terminal nodes represent viral infections that did not give rise to sampled progeny. If the mumps outbreak were primarily sustained by transmission within one group, the backbone of the phylogeny and the majority of internal nodes should be inferred as that group. We selected the 4 introductions that contained both Marshallese and non-Marshallese tips (**Fig. 4c**, the 4 Arkansas introductions), and reconstructed ancestral states along the phylogeny and migration/transmission rates between Marshallese and non-Marshallese groups using a structured coalescent model.

74/88 internal nodes were inferred to circulate within the Marshallese community with posterior probability of at least 0.95 (**Fig. 5a, b**). Transmission events from Marshallese into non-Marshallese demes resulted in short, terminal transmission chains (**Fig. 5a,** dark blue branches). This suggests that transmission was overwhelmingly maintained within the Marshallese community, and that infections seeded into the non-Marshallese community did not

sustain prolonged transmission chains. We estimate substantially more transmission from Marshallese to non-Marshallese groups than the opposite: within the primary outbreak clade, we estimate 29 transmission events from Marshallese to non-Marshallese groups (95% HPD: 21, 37), and only 6 (95% HPD: 0, 14) from non-Marshallese to Marshallese groups (**Fig. 5d**). This strongly suggests that transmission predominantly occurred in one direction: transmission events leading to non-Marshallese infections usually died out, and did not typically re-seed circulation within the Marshallese community. These results hold true regardless of migration rate prior (**Supplemental Figure 4**).

To ensure that our results were not driven by unequal sampling within the analyzed clades, we generated 3 datasets in which the number of Marshallese and non-Marshallese tips were subsampled to be equal. For each of these 3 subsampled datasets, we ran 3 independent chains under the same model described above. Chains converged for 2 of the 3 subsampled datasets. In the converged chains, we recover very similar tree topologies (**Supplemental Figure 5a**) with equivalent Marshallese and non-Marshallese samples. We also recovered maximum clade credibility trees in which the vast majority of the internal nodes are inferred to circulate within the Marshallese deme (**Supplemental Figure 5a,b**), confirming that our findings are robust to sampling, consistent with past observations of model performance [29,35,36].

The structured coalescent model requires both groups to be present in each cluster, excluding several small Washington introductions composed entirely of non-Marshallese tips (**Fig. 4c**). To account for this, we used all Washington genotype G sequences in our dataset and estimated a single tree using an approximate structured coalescent model [33]. All Washington sequences were annotated as either Marshallese or not Marshallese. To provide a “source” population for the extensive diversity among our disparate Washington introductions, we also specified a third,

unsampled deme, for which migration was only allowed to proceed outward. As above, we inferred very few non-Marshallese internal nodes (**Supplemental Figures 6 and 7**). All internal nodes in the primary outbreak group are inferred as Marshallese with high probability, while non-Marshallese cases are present as terminal nodes. We recovered support for a single non-Marshallese cluster, indicating limited sustained transmission in the non-Marshallese population.

Structured coalescent models infer the effective population size (N_e) for each group, which reflects the number of infections necessary to generate the observed genetic diversity. Differences in N_e can result from different transmission rates or different numbers of infected individuals [37], and can therefore approximate differences in disease frequency between groups. While the total number of Marshallese and non-Marshallese cases reported through the public health surveillance system were similar (**Supplemental Table 1**), we estimate that N_e for the non-Marshallese group is approximately 3 times higher than that of the Marshallese group. Assuming the same number of infected individuals in each group, lower N_e's suggest higher transmission rates [37], suggesting more transmission within the Marshallese deme. Taken together, our results suggest that the outbreak was primarily sustained by transmission within the Marshallese community. While this transmission sometimes spilled over into the non-Marshallese community, transmission was generally not as successful there and died out, resulting in short, terminal transmission chains.

Viruses infecting vaccination groups individuals are genetically similar

Although only 9.7% of reported mumps cases in Washington were not up-to-date for mumps vaccination, infection of these individuals could have disproportionately impacted transmission in the state. Emergence of an antigenically novel strain of mumps could also allow infection of

previously vaccinated individuals, and result in different virus lineages infecting individuals in different vaccination categories. We colored the tips of all Washington cases in our phylogeny to represent whether they were derived from individuals who were up-to-date, not up-to-date, or whose vaccination status was unknown. Mirroring overall vaccination coverage in Washington, the vast majority of samples in our dataset were from up-to-date individuals. The not up-to-date individuals present in our dataset are dispersed throughout the phylogeny and do not cluster together (**Fig. 6**), suggesting that there is no genetic difference between viruses infecting individuals with different vaccination statuses.

Discussion

The resurgence of mumps in North America has ushered renewed attention towards understanding post-vaccine era mumps transmission. We show that the Washington State outbreak was fueled by approximately 13 independent introductions, primarily from Arkansas, leading to multiple co-circulating transmission chains. Within Washington, transmission was more efficient within the Marshallese community. Marshallese individuals were more often sampled at the beginnings of transmission chains, contributed to longer transmission chains on average, and were overwhelmingly enriched on internal nodes within the phylogeny. We found no support for age or vaccination status as critical determinants of transmission in our outbreak, consistent with epidemiologic findings. Our data suggest that social networks can be critical determinants of mumps transmission. Future work exploring how social and economic disparities may amplify respiratory disease transmission will be necessary for updating outbreak mitigation and prevention strategies.

Our results highlight the utility of genomic data to clarify epidemiologic hypotheses. While genomic data and epidemiologic investigation (including case interviews and contact follow up) suggested an Arkansas introduction as the Washington outbreak's primary origin, sequence data revealed repeated and ongoing introductions into Washington, similar to patterns observed in Massachusetts [38]. Our finding that most introductions sparked short transmission chains suggests that mumps did not transmit efficiently among the general Washington populace, despite similar numbers of Marshallese and non-Marshallese cases. Although statewide vaccine coverage is high, the risk for acquiring and transmitting mumps is clearly not equal across all Washington population groups. Provision of an outbreak dose of mumps-containing vaccine to high-risk groups may therefore be especially effective for limiting mumps transmission in future outbreaks. Others have reported success in using outbreak dose mumps vaccinations to reduce mumps transmission on college campuses [3,39] and in the US army [40–43], and the CDC currently recommends providing outbreak vaccine doses to individuals with increased risk due to an outbreak [44]. Future work to quantify the interplay between contact rates and vaccine-induced immunity among different age and risk groups should be used to guide updated vaccine recommendations.

Recent research has focused on identifying groups at risk for mumps infection due to their age [4], with less attention to other factors that may make populations vulnerable. In the 2016/017 Washington State outbreak, Marshallese status was a crucial predictor of transmission risk. Mumps has historically caused outbreaks in communities with strong, interconnected contact patterns [5,45,46], and in dense housing environments [47]. While a combination of waning immunity and dense housing settings make college campuses ideal for mumps outbreaks, the Washington and Arkansas outbreaks show that populations other than young adults are at risk.

We speculate that within the Marshallese community, a combination of factors likely led to a high force of infection.

Multigenerational living is common in the Marshallese community [5], and Marshallese households tend to be larger on average (average household size = 5.28 [48], average household size for entire US populace = 2.52 [49]). Having more household contacts may have facilitated a greater number and higher intensity of interactions among individuals, allowing the force of infection to overcome pre-existing immunity. It is also possible that infection intensity within the Marshallese community was exacerbated by low rates of insurance coverage and poor access to healthcare [50,51], hesitancy to seek medical care [52], and health disparities stemming from US occupation, nuclear testing, and exclusion from healthcare services. As part of reparations for US nuclear testing, the US signed the Compact of Free Association Treaty (COFA)[53] with the Marshall Islands in 1989, permitting Marshallese residents to live and work in the US without visas. However, eligibility for Medicaid was revoked for COFA immigrants in 1996, and US-residing Marshallese remain economically disadvantaged and under-insured [17,50,54]. The passage of the Affordable Care Act (ACA) has not ameliorated these issues. Interviews with US-residing Marshallese note confusion among ACA staff regarding the legal status of COFA recipients, leading to drawn out enrollment processes that often leave individuals uninsured, frustrated [55], and far less likely to access care [51]. Marshallese trust in US medical institutions was further undermined by the unconsented use of Marshallese individuals for experiments on health impacts of nuclear exposure, with effects lingering today [6]. Banked historical samples confirm uptake of radioactive materials in Marshallese inhabitants of affected Islands [56], but there has been limited published data on long-term health impacts of nuclear exposure, and significant concern remains within the community [13]. Finally, when Marshallese individuals do access care, they report experiencing disdain from healthcare workers [57] and sub-optimal care

690 [55]. Interviews with medical workers show that blame for poor Marshallese health outcomes is
691 sometimes placed on host genetics or cultural practices [57], poor health literacy [58], or choosing to
692 delay care [58], with less consideration given to how the economic and legal impacts of US
693 occupation affect the health of Marshallese individuals. These factors compound, and
694 Marshallese individuals report hesitation to seek medical care, even when sick [55]. Hesitancy to
695 seek care could have contributed to mumps transmission if sick individuals were primarily cared
696 for at home without knowledge of or the ability to implement community-isolation protocols.
697
698 Our findings highlight that social networks can be the primary risk factor for a respiratory virus
699 outbreak, even when a vaccine is effective and widely used. Future work should explore
700 whether nuclear exposure has impacted Marshallese immune function and susceptibility to
701 infectious disease. The recent passing of WA Senate Bill 5683 marks an important step in
702 remedying the exclusion of Marshallese Washingtonians from healthcare services. Future work
703 to evaluate whether this change improves Marshallese access to healthcare and mitigates
704 increased disease risk within Washington's Marshallese community will be crucial follow-up.
705 The findings of this paper demonstrate the importance of expanding our understanding of
706 populations at risk for mumps re-emergence, so that rapid and comprehensive outbreak
707 response strategies can be implemented to mitigate negative health impacts for all affected
708 communities. Finally, future work to disentangle the complex interplay between healthcare
709 access, social and economic disparity, and respiratory virus risk will be essential for mitigating
710 health impacts of mumps and other respiratory viruses.
711

712 ## Acknowledgments

We would like to extend our sincerest thank you to Jiji Jally for her help and input on the project. Jiji Jally is an advocate for affordable access to healthcare services, supportive services for the Marshallese community, and works as a translator to assist the community in Washington State. These insightful discussions were absolutely critical for contextualizing our results. We would also like to sincerely thank Kelsey Florek for locating and sharing mumps samples from Wisconsin, Ohio, Missouri, Alabama, and North Carolina, which greatly enhanced the analyses presented here. We also thank Jeff Joy for graciously sharing mumps genomes from British Columbia. Finally, we would like to thank the Fred Hutchinson Cancer Research Center sequencing core for providing excellent sequencing services. LHM is an Open Philanthropy Project fellow of the Life Sciences Research Foundation. AB was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1256082. TB is a Pew Biomedical Scholar and is supported by NIH R35 GM119774-01.

Author contributions

LHM and AB generated and analyzed data and wrote the manuscript. CD and NRG provided data and contributed to writing the manuscript. ML, APO, and SL, provided samples and support for the study. DH provided conceptual input and contributed to the writing of the manuscript. NFM analyzed data and helped write the manuscript. TB planned the study, supported data analysis, and wrote the manuscript.

References

1. CDCMMWR. Notifiable Diseases and Mortality Tables. (2019).
2. Davidkin, I., Jokinen, S., Broman, M., Leinikki, P. & Peltola, H. Persistence of measles,

mumps, and rubella antibodies in an MMR-vaccinated cohort: a 20-year follow-up. *J. Infect. Dis.* **197**, 950–956 (2008).

3. Cardemil, C. V. *et al.* Effectiveness of a Third Dose of MMR Vaccine for Mumps Outbreak Control. *N. Engl. J. Med.* **377**, 947–956 (2017).
4. Lewnard, J. A. & Grad, Y. H. Vaccine waning and mumps re-emergence in the United States. *Sci. Transl. Med.* **10**, (2018).
5. Fields, V. S. *et al.* Mumps in a highly vaccinated Marshallese community in Arkansas, USA: an outbreak report. *Lancet Infect. Dis.* **19**, 185–192 (2019).
6. Barker, H. M. *Bravo for the Marshallese: Regaining Control in a Post-Nuclear, Post-Colonial World*. (Cengage Learning, 2012).
7. Palafox, N. A., Riklon, S., Alik, W. & Hixon, A. L. Health consequences and health systems response to the Pacific U.S. Nuclear Weapons Testing Program. *Pac. Health Dialog* **14**, 170–178 (2007).
8. Takahashi, T. *et al.* An investigation into the prevalence of thyroid disease on Kwajalein Atoll, Marshall Islands. *Health Phys.* **73**, 199–213 (1997).
9. Simon, S. L. A brief history of people and events related to atomic weapons testing in the Marshall Islands. *Health Phys.* **73**, 5–20 (1997).
10. Niedenthal, J. A history of the people of Bikini following nuclear weapons testing in the Marshall Islands: with recollections and views of elders of Bikini Atoll. *Health Phys.* **73**, 28–36 (1997).
11. Hallgren, E. A., McElfish, P. A. & Rubon-Chutaro, J. Barriers and opportunities: a community-based participatory research study of health beliefs related to diabetes in a US Marshallese community. *Diabetes Educ.* **41**, 86–94 (2015).
12. Abella, M. K. I. L., Molina, M. R., Nikolić-Hughes, I., Hughes, E. W. & Ruderman, M. A. Background gamma radiation and soil activity measurements in the northern Marshall

762 Islands. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15425–15434 (2019).

763 13. Bordner, A. S. *et al.* Measurement of background gamma radiation in the northern Marshall
764 Islands. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6833–6838 (2016).

765 14. Wong, D. C., Purcell, R. H. & Rosen, L. Prevalence of antibody to hepatitis A and hepatitis
766 B viruses in selected populations of the South Pacific. *Am. J. Epidemiol.* **110**, 227–236
767 (1979).

768 15. Adams, W. H., Fields, H. A., Engle, J. R. & Hadler, S. C. Serologic markers for hepatitis B
769 among Marshallese accidentally exposed to fallout radiation in 1954. *Radiat. Res.* **108**, 74–
770 79 (1986).

771 16. Yamada, S., Dodd, A., Soe, T., Chen, T.-H. & Bauman, K. Diabetes mellitus prevalence in
772 out-patient Marshallese adults on Ebeye Island, Republic of the Marshall Islands. *Hawaii*
773 *Med. J.* **63**, 45–51 (2004).

774 17. McElfish, P. A., Hallgren, E. & Yamada, S. Effect of US health policies on health care
775 access for Marshallese migrants. *Am. J. Public Health* **105**, 637–643 (2015).

776 18. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*
777 **9**, 357–359 (2012).

778 19. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
779 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

780 20. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for
781 virology research. *Nucleic Acids Res.* **40**, D593–8 (2012).

782 21. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**,
783 4121–4123 (2018).

784 22. Katoh, K., Misawa, K., Kuma, K. K.-I. & Miyata, T. MAFFT: a novel method for rapid
785 multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–
786 3066 (2002).

787 23. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
788 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**,
789 268–274 (2015).
790 24. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic
791 analysis. *Virus Evolution* **4**, (2018).
792 25. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure
793 of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007
794 (2016).
795 26. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with
796 BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
797 27. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography
798 finds its roots. *PLoS Comput. Biol.* (2009) doi:10.1371/journal.pcbi.1000520.
799 28. McLean, H. Q., Amy Parker Fiebelkorn, M. S. N., Temte, J. L. & Wallace, G. S. Prevention
800 of Measles, Rubella, Congenital Rubella Syndrome, and Mumps, 2013.
801 https://www.cdc.gov/mmwr/preview/mmwrhtml/rr6204a1.htm (2013).
802 29. Vaughan, T. G., Kühnert, D., Popinga, A., Welch, D. & Drummond, A. J. Efficient Bayesian
803 inference under the structured coalescent. *Bioinformatics* (2014)
804 doi:10.1093/bioinformatics/btu201.
805 30. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary
806 analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
807 31. Müller, N. F. & Bouckaert, R. Coupled MCMC in BEAST 2. *bioRxiv* 603514 (2019)
808 doi:10.1101/603514.
809 32. Müller, N. F., Rasmussen, D. A. & Stadler, T. The Structured Coalescent and Its
810 Approximations. *Molecular Biology and Evolution* vol. 34 2970–2981 (2017).
811 33. Müller, N. F., Rasmussen, D. & Stadler, T. MASCOT: parameter and state inference under

812 the marginal structured coalescent approximation. *Bioinformatics* (2018)
813 doi:10.1093/bioinformatics/bty406.
814 34. Health Departments: How to Optimize Mumps Testing | CDC.
815 https://www.cdc.gov/mumps/health-departments/optimize-testing.html (2019).
816 35. Dudas, G., Carvalho, L. M., Rambaut, A. & Bedford, T. MERS-CoV spillover at the camel-
817 human interface. *Elife* (2018) doi:10.7554/eLife.31257.
818 36. De Maio, N., Wu, C. H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A
819 Bayesian Structured Coalescent Approximation. *PLoS Genet.* (2015)
820 doi:10.1371/journal.pgen.1005421.
821 37. Volz, E. M. Complex population dynamics and the coalescent under neutrality. *Genetics*
822 **190**, 187–201 (2012).
823 38. Wohl, S. *et al.* Combining genomics and epidemiology to track mumps virus transmission in
824 the United States. *PLoS Biol.* **18**, e3000611 (2020).
825 39. Shah, M. *et al.* Mumps Outbreak in a Highly Vaccinated University-Affiliated Setting Before
826 and After a Measles-Mumps-Rubella Vaccination Campaign-Iowa, July 2015-May 2016.
827 *Clin. Infect. Dis.* **66**, 81–88 (2018).
828 40. Vaccine Policy and Guidance for Adults and Accessions_1.pdf.
829 41. Arday, D. R., Kanjarpane, D. D. & Kelley, P. W. Mumps in the US Army 1980-86: should
830 recruits be immunized? *Am. J. Public Health* **79**, 471–474 (1989).
831 42. Kelley, P. W., Petruccelli, B. P., Stehr-Green, P., Erickson, R. L. & Mason, C. J. The
832 susceptibility of young adult Americans to vaccine-preventable infections. A national
833 serosurvey of US Army recruits. *JAMA* **266**, 2724–2729 (1991).
834 43. Eick, A. A., Hu, Z., Wang, Z. & Nevin, R. L. Incidence of mumps and immunity to measles,
835 mumps and rubella among US military recruits, 2000-2004. *Vaccine* **26**, 494–501 (2008).
836 44. Marlow, M. A., Marin, M., Moore, K. & Patel, M. CDC Guidance for Use of a Third Dose of

MMR Vaccine During Mumps Outbreaks. *J. Public Health Manag. Pract.* **26**, 109–115 (2020).

45. Nelson, G. E. *et al.* Epidemiology of a mumps outbreak in a highly vaccinated island population and use of a third dose of measles-mumps-rubella vaccine for outbreak control--Guam 2009 to 2010. *Pediatr. Infect. Dis. J.* **32**, 374–380 (2013).
46. Barskey, A. E. *et al.* Mumps outbreak in Orthodox Jewish communities in the United States. *N. Engl. J. Med.* **367**, 1704–1713 (2012).
47. Snijders, B. E. P. *et al.* Mumps vaccine effectiveness in primary schools and households, the Netherlands, 2008. *Vaccine* **30**, 2999–3002 (2012).
48. We the People: Pacific Islanders in the United States.
49. US Census Bureau. Historical Households Tables.
50. McElfish, P. A. *et al.* Diabetes and Hypertension in Marshallese Adults: Results from Faith-Based Health Screenings. *J Racial Ethn Health Disparities* **4**, 1042–1050 (2017).
51. Towne, S. D. *et al.* Inequities in Access to Medical Care Among Adults Diagnosed with Diabetes: Comparisons Between the US Population and a Sample of US-Residing Marshallese Islanders. *J Racial Ethn Health Disparities* (2020) doi:10.1007/s40615-020-00791-x.
52. Williams, D. P. & Hampton, A. Barriers to health services perceived by Marshallese immigrants. *J. Immigr. Health* **7**, 317–326 (2005).
53. Congress, 108th United States. Compact of Free Association Amendments Act of 2003. (2003).
54. McElfish, P. A. Marshallese COFA Migrants in Arkansas. *J. Ark. Med. Soc.* **112**, 259–60, 262 (2016).
55. McElfish, P. A. *et al.* Interpretive policy analysis: Marshallese COFA migrants and the Affordable Care Act. *Int. J. Equity Health* **15**, 91 (2016).

862 56. Simon, S. L., Bouville, A., Melo, D., Beck, H. L. & Weinstock, R. M. Acute and chronic
863 intakes of fallout radionuclides by Marshallese from nuclear weapons testing at Bikini and
864 Enewetak and related internal radiation doses. *Health Phys.* **99**, 157–200 (2010).
865 57. Duke, M. R. Neocolonialism and Health Care Access among Marshall Islanders in the
866 United States. *Med. Anthropol. Q.* **31**, 422–439 (2017).
867 58. McElfish, P. A., Chughtai, A., Low, L. K., Garner, R. & Purvis, R. S. 'Just doing the best we
868 can': health care providers' perceptions of barriers to providing care to Marshallese patients
869 in Arkansas. *Ethn. Health* 1–14 (2018).

870

871 ## Figure Legends

872 **Figure 1: Washington case counts and genomic sampling**

873 The first mumps case in Washington was reported on October 30, 2016, and case counts
874 peaked in winter of 2017. The x-axis represents the date of the corresponding epidemiologic
875 week, and the y-axis represents the number of confirmed and probable cases for that week.
876 Blue dots above the case count plot represent the number of Washington genome sequences in
877 our dataset sampled in that week.

878

879 **Figure 2: North American mumps outbreaks are related**

880 We combined all publicly available North American mumps genomes and built a time-resolved
881 phylogeny. Here, we display the maximum clade credibility tree, where each color represents a
882 unique US state or Canadian province and the x-axis represents the collection date (for tips), or
883 the inferred time to the most recent common ancestors (for internal nodes). We inferred
884 geographic history using a discrete trait model. The color of each internal node represents the
885 posterior probability of the inferred geographic location, where increasingly grey tone represents

886 decreasing probability.

887

888 **Figure 3: The mumps outbreak in Washington was seeded by approximately 13**
889 **introductions**

890 **a.** We separated each introduction inferred on the maximum clade credibility tree (Figure 2) and
891 plotted them independently. Large, colored dots represent the inferred geographic location that
892 the Washington introduction was seeded from. Branches that extend earlier than July of 2016
893 are dotted to represent that transmission likely occurred via other, unsampled locations. For
894 reference, the cumulative case counts from Arkansas and Washington are plotted below. **b.** For
895 each tree in the posterior set, we inferred the number of introductions into Washington. We plot
896 the proportion of trees in the posterior set in which that number of introductions was inferred.

897

898 **Figure 4: Marshallese individuals sustain longer transmission chains**

899 **a.** A schematic for quantifying tips that lie “upstream” in transmission chains. For tips that lie on
900 an internal node, i.e., have a branch length of zero, we infer the number of child tips that
901 descend from that tip’s parental node. For each tip in the example tree, the number of
902 descendants we would infer is annotated along side it. All tips that have a nonzero branch
903 length are annotated as having 0 descendants. We can then compare whether sequences of
904 particular groups (here, blue vs. red) are more likely to have descendants in the tree via logistic
905 regression. **b.** We separated all Washington tips and classified them into Marshallese and not
906 Marshallese. We then performed a rarefaction analysis and plot the number of inferred
907 Washington clusters (y-axis) as a function of the number of sequences included in the analysis
908 (x-axis). Dark blue represents not Marshallese sequences, and light blue represents
909 Marshallese sequences. Each dot represents the number of trials in which that number of
910 clusters was inferred, and the solid line represents the mean across trials. **c.** The exploded tree

as shown in Figure 3a is shown, but tips are now colored by whether they represent Marshallese or non-Marshallese cases. For reference, the number of Washington cases (y-axis) is plotted over time (x-axis), where bar color represents whether those cases were Marshallese or not.

Figure 5: The Washington outbreak was sustained by transmission in the Marshallese community

a. Using the 4 Washington clusters that had a mixture of Marshallese and non-Marshallese cases, we inferred phylogenies using a structured coalescent model. Each group of sequences shared a clock model, migration model, and substitution model, but each topology was inferred separately, allowing us to incorporate information from all 4 clusters into the migration estimation. For each cluster, the maximum clade credibility tree is shown, where the color of each internal node represents the posterior probability that the node is Marshallese. **b.** For each internal node shown in panel a, we plot the posterior probability of that node being Marshallese. Across all 4 clusters, 74 out of 88 internal nodes (84%) are inferred as Marshallese with a posterior probability of at least 0.95. **c.** The posterior distribution of the number of “jumps” or transmission events from Marshallese to not Marshallese (light blue) and not Marshallese to Marshallese (dark blue) inferred for the primary outbreak clade.

Figure 6: Individuals in different vaccination groups are infected by genetically similar viruses.

The exploded tree as shown in Figure 3a is shown, but tips are now colored by whether they represent cases from individuals who are up-to-date for mumps vaccination, not up-to-date, or cases for which vaccination status was unknown. The color of the large dot represents the inferred geographic location from which the Washington introduction was seeded.

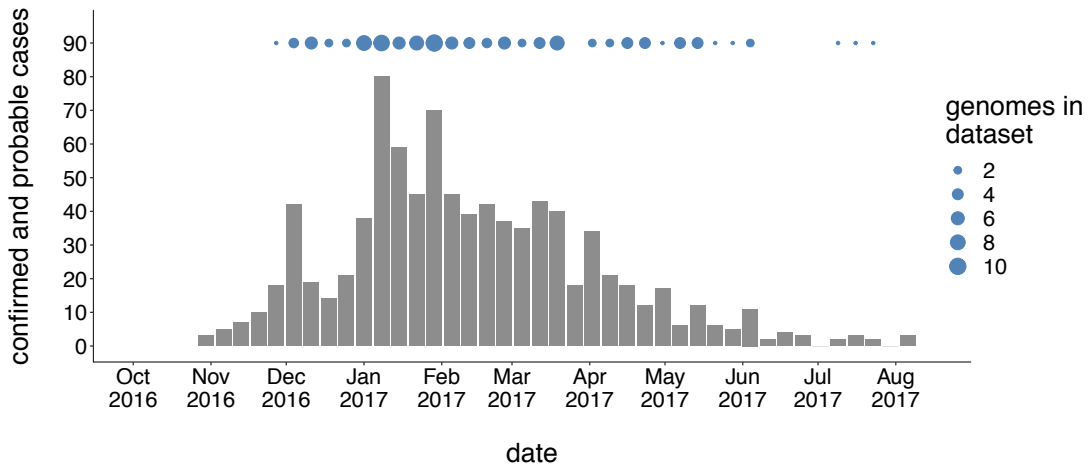

Figure 1: Washington case counts and genomic sampling
The first mumps case in Washington was reported on October 30, 2016, and case counts peaked in winter of 2017. The x-axis represents the date of the corresponding epidemiologic week, and the y-axis represents the number of confirmed and probable cases for that week. Blue dots above the case count plot represent the number of Washington genome sequences in our dataset sampled in that week.

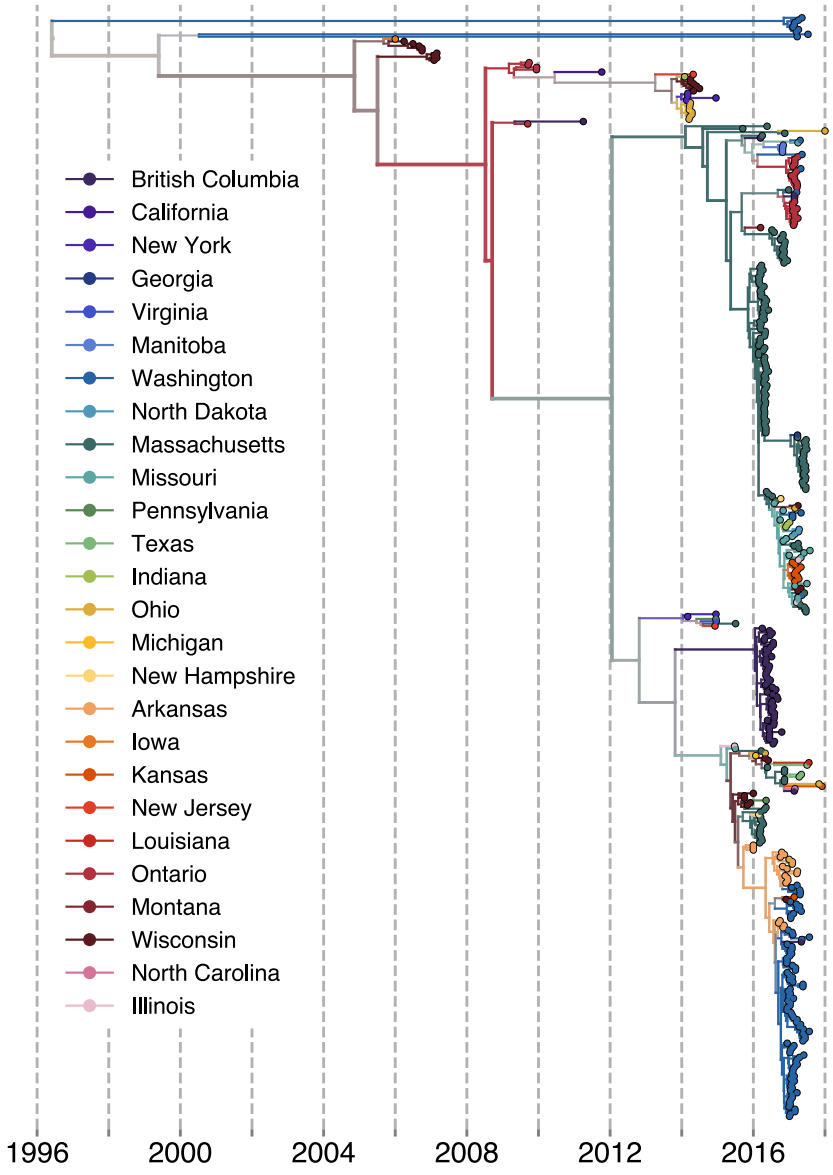

Figure 2: North American mumps outbreaks are related
We combined all publicly available North American mumps genomes and built a time-resolved phylogeny. Here, we display the maximum clade credibility tree, where each color represents a unique US state or Canadian province and the x-axis represents the collection date (for tips), or the inferred time to the most recent common ancestors (for internal nodes). We inferred geographic history using a discrete trait model. The color of each internal node represents the posterior probability of the inferred geographic location, where increasingly grey tone represents decreasing probability.

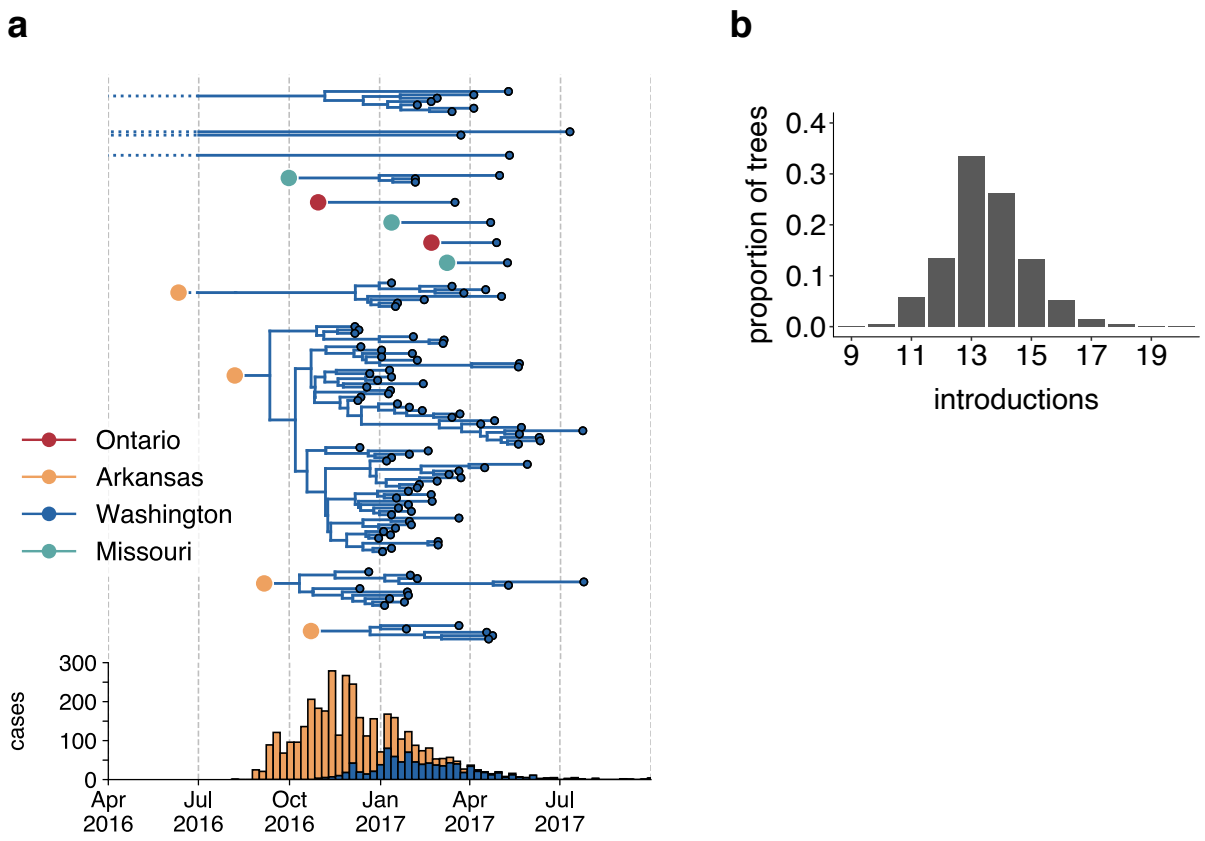

Figure 3: The mumps outbreak in Washington was seeded by approximately 13 introductions

a. We separated each introduction inferred on the maximum clade credibility tree (Figure 2) and plotted them independently. Large, colored dots represent the inferred geographic location that the Washington introduction was seeded from. Branches that extend ealier than July of 2016 are dotted to represent that transmission likely occurred via other, unsampled locations. For reference, the cumulative case counts from Arkansas and Washington are plotted below. **b.** For each tree in the posterior set, we inferred the number of introductions into Washington. We plot the proportion of trees in the posterior set in which that number of introductions was inferred.

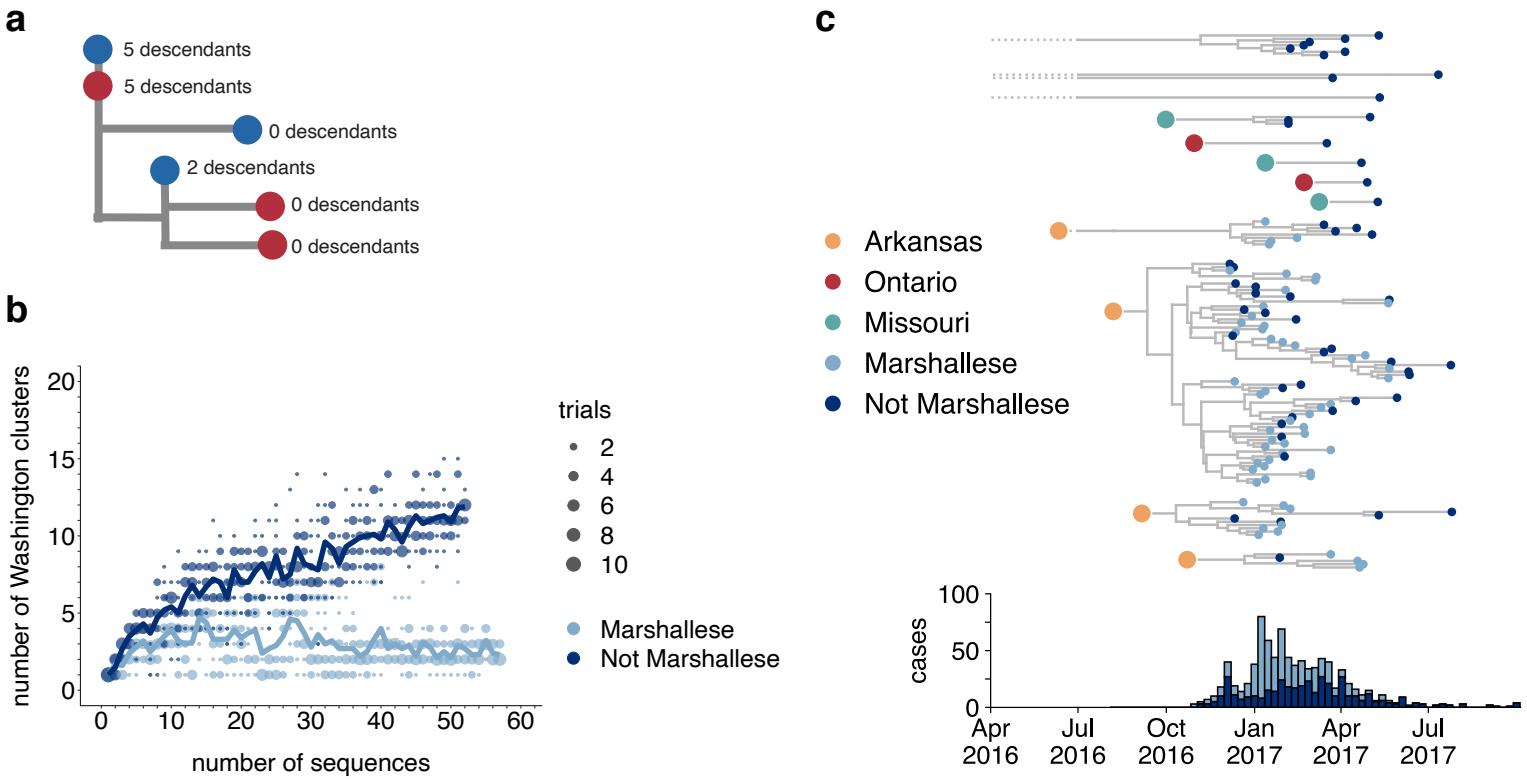

Figure 4: Marshallese individuals sustain longer transmission chains
a. A schematic for quantifying tips that lie "upstream" in transmission chains. For tips that lie on an internal node, i.e., have a branch length of zero, we infer the number of child tips that descend from that tip's parental node. For each tip in the example tree, the number of descendants we would infer is annotated along side it. All tips that have a nonzero branch length are annotated as having 0 descendants. We can then compare whether sequences of particular groups (here, blue vs. red) are more likely to have descendants in the tree via logistic regression. **b.** We separated all Washington tips and classified them into Marshallese and not Marshallese. We then performed a rarefaction analysis and plot the number of inferred Washington clusters (y-axis) as a function of the number of sequences included in the analysis (x-axis). Dark blue represents not Marshallese sequences, and light blue represents Marshallese sequences. Each dot represents the number of trials in which that number of clusters was inferred, and the solid line represents the mean across trials. **c.** The exploded tree as shown in Figure 3a is shown, but tips are now colored by whether they represent Marshallese or non-Marshallese cases. For reference, the number of Washington cases (y-axis) is plotted over time (x-axis), where bar color represents whether those cases were Marshallese or not.

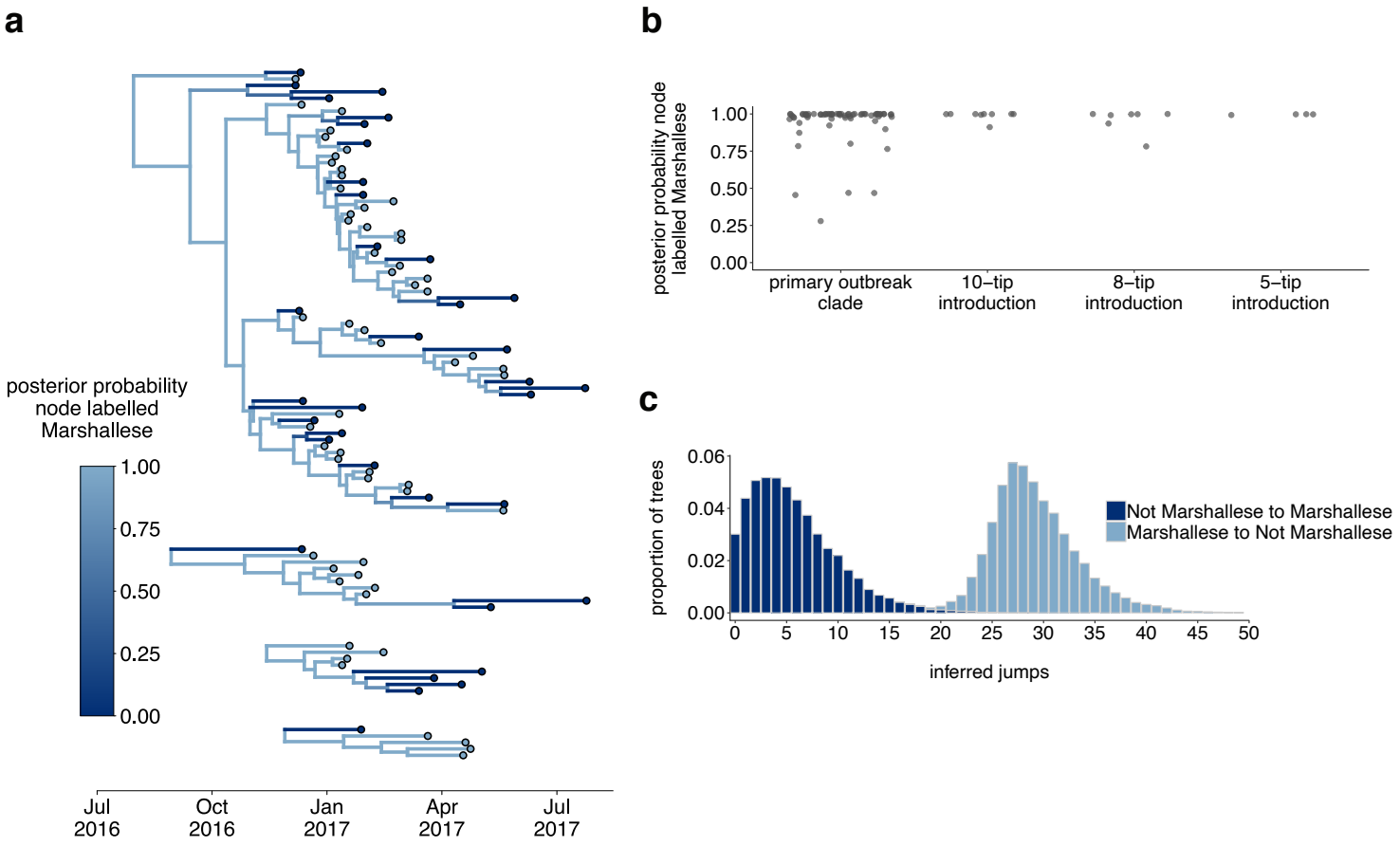

Figure 5: The Washington outbreak was sustained by transmission in the Marshallese community
a. Using the 4 Washington clusters that had a mixture of Marshallese and non-Marshallese cases, we inferred phylogenies using a structured coalescent model. Each group of sequences shared a clock model, migration model, and substitution model, but each topology was inferred separately, allowing us to incorporate information from all 4 clusters into the migration estimation. For each cluster, the maximum clade credibility tree is shown, where the color of each internal node represents the posterior probability that the node is Marshallese. **b.** For each internal node shown in panel **a**, we plot the posterior probability of that node being Marshallese. Across all 4 clusters, 74 out of 88 internal nodes (84%) are inferred as Marshallese with a posterior probability of at least 0.95. **c.** The posterior distribution of the number of "jumps" or transmission events from Marshallese to not Marshallese (light blue) and not Marshallese to Marshallese (dark blue) inferred for the primary outbreak clade.

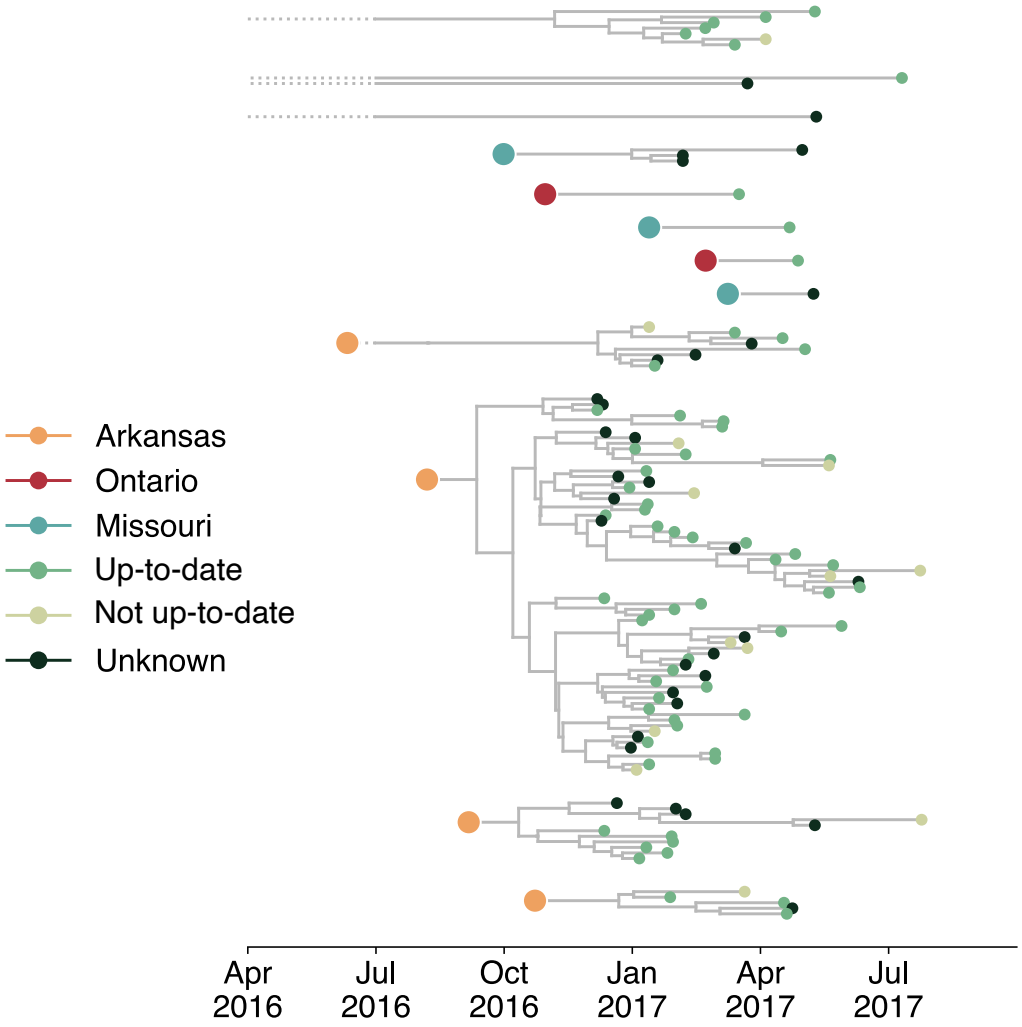

Figure 6: Individuals in different vaccination groups are infected by genetically similar viruses. The exploded tree as shown in Figure 3a is shown, but tips are now colored by whether they represent cases from individuals who are up-to-date for mumps vaccination, not up-to-date, or cases for which vaccination status was unknown. The color of the large dot represents the inferred geographic location from which the Washington introduction was seeded.