bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 **Targeted sequence capture of *Orientia tsutsugamushi* DNA from chiggers**
2 **and humans**
3
4 Ivo Elliott[1,2], Neeranuch Thangnimitchok[1], Mariateresa de Cesare[3], Piyada
5 Linsuwanon[4], Daniel H. Paris[5,6], Nicholas PJ Day[2,7], Paul N. Newton[1,2,7], Rory
6 Bowden[3], Elizabeth M. Batty[2,4]
7
8 Affiliations:
9 1. Lao-Oxford-Mahosot Hospital-Wellcome Trust Research Unit, Microbiology
10 Laboratory, Mahosot Hospital, Vientiane, Lao PDR
11 2. Centre for Tropical Medicine and Global Health, Nuffield Department of
12 Medicine, University of Oxford, Oxford, United Kingdom
13 3. Wellcome Centre for Human Genetics, University of Oxford, Oxford, United
14 Kingdom
15 4. Department of Entomology, Armed Forces Research Institute of Medical
16 Sciences, Bangkok, Thailand
17 5. Department of Medicine, Swiss Tropical and Public Health Institute, Basel,
18 Switzerland
19 6. Department of Clinical Research, University of Basel, Basel, Switzerland
20 7. Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine,
21 Mahidol University, Bangkok, Thailand
22
23 Corresponding author: Ivo Elliott ivo@tropmedres.ac
24
25 Word count: 5,438
26

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

27 **Abstract**

28 Scrub typhus is a febrile disease caused by *Orientia tsutsugamushi*, transmitted
29 by larval stage Trombiculid mites (chiggers), whose primary hosts are small
30 mammals. The phylogenomics of *O. tsutsugamushi* in chiggers, small mammals
31 and humans remains poorly understood. To combat the limitations imposed by
32 the low relative quantities of pathogen DNA in typical *O. tsutsugamushi* clinical
33 and ecological samples, along with the technical, safety and cost limitations of
34 cell culture, a novel probe-based target enrichment sequencing protocol was
35 developed. The method was designed to capture variation among conserved
36 genes and facilitate phylogenomic analysis at the scale of population samples. A
37 whole-genome amplification step was incorporated to enhance the efficiency of
38 sequencing by reducing duplication rates. This resulted in on-target capture
39 rates of up to 93% for a diverse set of human, chigger, and rodent samples, with
40 the greatest success rate in samples with real-time PCR C_t values below 35.
41 Analysis of the best-performing samples revealed phylogeographic clustering at
42 local, provincial and international scales. Applying the methodology to a
43 comprehensive set of samples could yield a more complete understanding of the
44 ecology, genomic evolution and population structure of *O. tsutsugamushi* and
45 other similarly challenging organisms, with potential benefits in the
46 development of diagnostic tests and vaccines.

47

48 **Introduction**

49 Scrub typhus is a vector-borne zoonotic disease risking life-threatening febrile
50 infection in humans. The disease is caused by an obligate intracellular Gram-
51 negative bacterium, *Orientia tsutsugamushi.* Scrub typhus has an expanding

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

52 known distribution, with most disease occurring across South and East Asia and
53 parts of the Pacific Rim.
54 The genus *Orientia* is classified in the family Rickettsiaceae, a member of the
55 order Rickettsiales. Two species of *Orientia* are currently recognised - *O.*
56 *tsutsugamushi* and *O. chuto*, the latter known solely from a patient infected in the
57 United Arab Emirates [1]. Recent molecular identification of *O. tsutsugamushi* in
58 humans in Chile [2] and 16S sequences with close homology to *O. tsutsugamushi* in
59 dogs in South Africa [3] and small mammals in Senegal and France [4], and to *O.*
60 *chuto* in chiggers in Kenya [5], suggest the possibility of further species and future
61 taxonomic re-evaluation.
62 Larval trombiculid mites (chiggers) transmit *Orientia* to vertebrates, including
63 man. The organism appears to be maintained by transovarial (vertical) and
64 transstadial (between life-stages) transmission in chiggers, suggesting that they
65 act as both vector and reservoir [6-9]. There is good evidence for the transmission
66 of *O. tsutsugamushi* to man by at least 10 species of chiggers [10]. The ecology of the
67 disease and the interaction of *Orientia* between vectors, small mammals and
68 humans are complex and relatively poorly understood [11].
69 A high degree of phenotypic and genotypic diversity has been reported in *O.*
70 *tsutsugamushi*. Several antigenic types appear to be widely present throughout
71 Southeast Asia, with one (TA716) making up over 70% of isolates from several
72 countries [12]. More recently, genetic analysis of highly variable single genes for
73 outer membrane proteins such as the 56kDa and 47kDa antigens or more
74 conserved genes (e.g. GroEL) have been used to define genotypic variation. A
75 recent detailed analysis of 56kDa sequences from across South and East Asia
76 identified at least 17 clusters of genotypes belonging to 5 identifiable groups [13].

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

77 Several multi-locus sequence typing (MLST) schemes using sets of housekeeping
78 genes have been proposed, though no single scheme has been universally
79 accepted [14-18]. Using one MLST scheme, human isolates from 3 regions of Laos
80 and an isolate from nearby Udon Thani in Northeast Thailand were compared.
81 Low levels of population differentiation were reported between geographically
82 close (Vientiane and Udon Thani) strains, while isolates from southern Laos
83 formed a distinct population [16]. In that study, 8% of isolates appeared to
84 represent mixed infection, and in Thailand 25% of infections were reportedly
85 mixed [18]. Recent whole-genome phylogenetic comparisons between 8 well-
86 characterised strains revealed relationships that were significantly different
87 from phylogenies created from single-gene or MLST schemes, illustrating the
88 increased resolution achievable from whole-genome sequencing [19]. At the level
89 of individual genes such as 56kDa, enormous genetic variability is seen, while at
90 the MLST level only a few clonal clusters are evident.
91 Several factors combine to make genomic studies of *Orientia* infection
92 challenging. The bacterium is an obligate intracellular pathogen, necessitating
93 cell culture for laboratory propagation [20]. *Orientia* is typically collected from a
94 range of specimen types including human whole blood, buffy coat and eschar
95 tissue, rodent blood and organs, and chiggers, and the absolute quantity of *O.*
96 *tsutsugamushi* DNA present in these specimen types is variable, but frequently
97 low. *Orientia* can only be propagated in cell culture, which is technically
98 demanding [20] and costly and must be performed in biosafety level 3 facilities [21].
99 In one study of 155 infected human blood samples tested by 16S PCR, the median
100 pathogen genome load was 0.013 copies/μL, the interquartile range 0-0.334 and
101 the maximum 310 [22], while a recent study from Thailand reported a range of 13.8

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

102 to 2,252 copies/μL [23]. Very few data are available for the quantity of *O.*
103 *tsutsugamushi* in individual chiggers and there are no published data from
104 rodents. The *O. tsutsugamushi* genome is relatively poorly defined, with just nine
105 complete genome sequences, and shows a high density of repetitive elements
106 and extreme rates of genomic rearrangement, two added challenges that make
107 innovative approaches to sample preparation, sequencing and analysis essential
108 [19,24,25].

109 Next-generation sequencing (NGS) techniques have become the gold standard for
110 revealing the genetic variation of organisms [26]. Culture of *O. tsutsugamushi* in
111 eukaryotic cells can increase the quantity and concentration of DNA available for
112 downstream whole-genome sequencing by thousands of fold. This technique is
113 technically demanding, costly, time-consuming and prone to contamination.
114 Handling infected-cell cultures is also hazardous and carries a risk of infection in
115 those accidentally exposed [21]. The entire process must be undertaken in biosafety
116 level 3, with all its associated costs and complications.

117 Targeted enrichment sequencing is a tool whereby certain pre-selected regions of
118 the genome are targeted for sequencing, via hybridisation to a set of probes
119 corresponding to the sequences of interest. The method is akin to, and works
120 similarly to, whole-exome sequencing where just the "exome" or coding portion of
121 the human genome is sequenced. Targeted enrichment can be useful where the
122 whole genome is not required, or a particular genome of interest is selected from
123 contaminating DNA [27,28], for example in the metagenomic analysis of multiple
124 virus species, where culture is difficult and costly [29-31], and for *Neisseria*
125 *meningitidis* directly from cerebrospinal fluid, where culture often fails due to

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

126 prior antibiotic treatment [32]. Thus, the method in principle provides an efficient
127 alternative to cell culture combined with whole-genome sequencing for *Orientia.*

128 In summary, the many difficulties associated with conducting a large-scale study
129 at the whole-genome level of *O. tsutsugamushi* in human, chiggers and small
130 mammal samples prompted the development of a probe-based targeted
131 enrichment sequencing strategy, which was used to examine phylogeographical
132 relatedness of samples collecting in Northern Thailand and elsewhere.

133 **Materials and Methods**

134 **Sample collection**

135 Small mammals were trapped alive in wire-mesh traps baited with corn. Animals
136 were killed using the inhalational anaesthetic isoflurane. Chiggers were collected
137 from rodents by removing the ears and placing into tubes containing 70% ethanol
138 and stored at 4°C. The rodent lung, liver and spleen were removed, preserved in
139 70% ethanol and stored at -80°C [33]. International standards were stringently
140 followed for animal-handling and euthanasia procedures [34,35]. Free-living chiggers
141 were collected using the black plate method [36,37]. Human blood and eschar
142 samples were collected during the non-malarial fever studies in Laos [16] and the
143 natural immune response to paediatric scrub typhus study in Thailand and stored
144 at -80°C [38]. Chiggers were identified using autofluorescence and bright-field
145 microscopy [39] with reference to a range of taxonomic keys [40-42]. Ethical approval
146 was obtained from Kasetsart University Animal Ethics Committee (EC), Bangkok,
147 Thailand for animal collection; the Faculty of Tropical Medicine EC, Mahidol
148 University, Bangkok, the Chiangrai Prachanukroh Hospital EC, the Chiangrai
149 Provincial Public Health EC and the Oxford Tropical Research EC for human

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

150 samples in Thailand and additionally the Lao National Committee for Health
151 Research for human samples in Laos.

152 **DNA extraction and PCR**

153 DNA was extracted from individual chiggers, pools of chiggers, rodent tissues and
154 human samples using the Qiagen Blood and Tissue Kit (Qiagen, USA). The
155 procedures prior to protein digestion were as follows. Chiggers were rinsed with
156 distilled water and individuals cut through the mid-gut using a sterile 30G needle
157 under a dissecting microscope and pools crushed using a sterile polypropylene
158 motorized pestle (Motorized pellet pestle Z35991, Sigma Aldrich, St Louis, MO).
159 Rodent tissues were cut into a small piece (≤10mg of spleen or ≤25mg of liver or
160 lung). Buffy coat or whole blood was extracted from a starting volume of 200 µl.
161 Eschars were collected either as pieces of crust in 70% ethanol or swabs. Chigger,
162 rodent and eschar swabs were incubated with proteinase K at 56°C for 3 hours.
163 Whole blood and buffy coat was incubated for 1 hour and eschar crust was
164 incubated overnight. The rest of the steps followed the manufacturer's protocol.
165 Chigger samples were eluted in 45 µl, while rodent and human samples were
166 eluted in 100 µl of buffer AE (Qiagen, Hilden, Germany). Samples were stored at -
167 20°C before PCR.

168 Quantitative real-time PCR targeting the 47kDa *O. tsutsugamushi* outer-membrane
169 protein was performed on all rodent, chigger and human samples [43]. A PCR master
170 mix was prepared by combining the following reagent volumes per sample: 15 µl
171 of Platinum PCR Supermix UDG (Sigma Aldrich, USA), 0.25 µl each of Forward and
172 Reverse Primers (10 µM) and 0.5 µl of Probe (10 µM). For chigger samples 4 µl of
173 sterile water and 5 µl of DNA was added. For rodent and human samples 8 µl of

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

174 sterile water and 1 µl of DNA added to complete the Master Mix. PCR was run with
175 the following conditions: 2 minutes at 50°C, then denaturation at 95°C for 2
176 minutes, followed by 45 cycles of 95°C for 15 seconds and 60°C for 30 seconds.
177 Real-time PCR was performed on a Bio Rad CFX96 (Bio Rad, USA) using in-house
178 quantitative standards. Duplicate 10-fold concentrations from 10^0 to 10^6 (1 µl
179 each) and two no-template controls were included on every run.

180 **Library preparation**

181 In the first round of sequencing in this study, the Nextera XT DNA library
182 preparation kit (Illumina Inc, San Diego, USA) methodology was used to prepare
183 libraries, predominantly for human-derived samples. High duplication rates and
184 relatively low coverage for this approach resulted in a switch to a whole-genome
185 amplification (WGA) step prior to a ligation-based library preparation method.

186 For Nextera XT libraries, DNA was normalized for an input of ≤1 ng in 5 µL across
187 all samples and libraries were prepared following the manufacturer's protocol.

188 For whole-genome amplified libraries, specimens from input volumes ranging
189 from 40 µL (chiggers) and ~50 µL (human samples), to 95 µL for small mammal
190 samples were dried using a Speed-Vac (Eppendorf, Hamburg, Germany) and
191 resuspended in 2.5 µL of TE. WGA was performed following the manufacturer's
192 protocol for the REPLI-g Single Cell Kit (Qiagen, Hilden, Germany).

193 The concentration of the amplified DNA was assessed using a Qubit dsDNA HS
194 Assay (Thermo Fisher, MA, USA). Samples were normalized to 500 ng mass in 34
195 µL DNA and fragmented using an Episonic instrument, (EpiGentek, NY, USA) with
196 the following settings: Amplitude 40, Process time 00:03:20, Pulse-ON time

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

197 00:00:20, Pulse-OFF time 00:00:20. The fragmented DNA was cleaned with a 1X
198 ratio of AMPure XP beads (Beckman Coulter, Indianapolis, USA), resuspended in
199 34 μl.

200 Libraries were prepared using the NEBNext Ultra DNA Library Prep Kit for
201 Illumina (New England Bioalabs) with a modified protocol. In detail, 6.5 μL
202 NEBNext End repair reaction buffer, 0.75 μL NEBNext End prep enzyme mix and
203 24.25 μL nuclease-free water were added to each sample and incubated at 20°C
204 for 30 mins and 65°C for 30 minutes. Next, ligation of an in-house Y-adapter was
205 performed by adding 3.75 μL of Blunt/TA Ligase master mix, 1 μL of Ligation
206 enhancer, 1.5 μL of 15 μM adapter and 12.25 μL of nuclease-free water to each
207 sample. This was then incubated for 15 minutes at 20°C, followed by an AMPure
208 XP bead clean-up using 86.5 μL of beads and finally eluted into 100 μL EB buffer.

209 For sequencing on the Illumina HiSeq4000, an AMPure XP size-selection was then
210 performed by adding 52 μL of AMPure XP to the DNA, mixing, incubating for 5
211 minutes at room temperature and then transferring to a magnet for 8 minutes. The
212 supernatant was then transferred to a fresh plate and the process repeated using
213 25 μL of AMPure XP. Finally, the beads were washed twice with ethanol and
214 resuspended in 20 μL of EB buffer.

215 PCR was then performed on the library using 10 μL of Pre-PCR library, 5 μL of
216 indexed primer i5 and i7, 10 μL water and 25 μL NEBNext Q5 PCR Master Mix. The
217 following conditions were used: 98°C for 30secs, 98°C for 10secs, 65°C for 30secs,
218 72°C for 30secs, 72°C for 5mins and 10 cycles performed.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

219 A final AMPure XP bead clean-up was carried out using 37.5 µL of beads and eluted
220 in 30 µL of EB buffer. Qubit and Tapestation DNA analysis was performed for all
221 libraries prior to target enrichment.

222 **Target enrichment**

223 Paired-end DNA libraries prepared using either WGA followed by an in-house
224 library preparation, or Nextera XT, were pooled for capture using pre-designed
225 Agilent SureSelectXT Custom 3-5.9Mb probes and the capture module of the
226 SureSelectXT Reagent Kit, HSQ (Agilent).

227 The pool of indexed libraries was first normalized to 750 ng in 3.4 µL. A Master
228 Mix containing 2.5 µL of SureSelect Indexing Block #1, 2.5 µL SureSelect Block #2,
229 3 µL IDT xGen Blocking Oligos was prepared. This was added to the sample, mixed
230 and placed on a thermocycler at 95°C for 5 minutes and then 65°C for 5 minutes.

231 Next the Hybridization Buffer Master Mix (SureSelect Hyb #1 to #4 and RNase
232 Block) in a total volume 13.5 µL was prepared. 5 µL of baits were aliquoted and
233 added to the Hybridization Buffer Master Mix. This was then transferred to the
234 samples held at 65°C and incubated for 24hrs.

235 Dynabeads MyOne Streptavidin T1 beads were prepared using the
236 manufacturer's standard protocol. The PCR plate was maintained at 65°C while
237 moving the samples to the bead plate and pipette mixing. Samples were then
238 incubated on a mixer at 1100 rpm for 30 minutes at room temperature. Samples
239 were then spun briefly, place on a magnetic rack and the supernatant removed
240 and saved. The beads were resuspended in 200 µL of SureSelect Wash Buffer 1
241 and incubated for 15 minutes at room temperature, replaced on the magnetic

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

242 rack and the supernatant discarded. The procedure was repeated with
243 SureSelect Wash Buffer 2, incubated for 10 minutes at 65°C and discarding the
244 supernatant as before. The process was repeated 3 times. The beads were then
245 resuspended in 30 µL of distilled water, of which 14 µL was transferred to a post-
246 hybridization PCR using the following PCR Master Mix (Herculase II Reaction
247 buffer, 100mM dNTP Mix, qPCR Library Quantification Primer Premix, nuclease
248 free water and Herculase II Fusion DNA Polymerase), with the cycle parameters
249 of: 98°C for 2mins then 14 cycles of 98°C for 30secs, 57°C for 30secs, 72°C for 1
250 min, followed by a final extension of 72°C for 10 minutes.

251 **Sequencing**

252 Sequencing was performed on the Illumina HiSeq4000 with paired-end 150 bp
253 reads.

254 **Bioinformatic analysis**

255 Raw reads generated from Illumina HiSeq4000 were mapped to the UT76
256 reference genome (GCF_900327255.1) using BWA MEM v0.7.12 [44]. Samtools
257 flagstat v1.8 was used to summarise the total number of reads and the proportion
258 mapping to the reference. The reads were then deduplicated using Picard
259 MarkDuplicates v2.0.1 and the same statistics were recalculated, along with the
260 total number of fragments present in the library. Depth of coverage across the
261 whole genome and the proportion of the core genome represented at 1x, 5x and
262 10x minimum per-base coverage was calculated using GATK v3.7 [45].

263 Haploid variant calling and core genome alignment was performed using Snippy
264 v4.3.6 [46]. The method identified single nucleotide polymorphisms (SNPs) between

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

265 the sequence reads and the reference genome. The variant calls were used as input
266 to construct maximum-likelihood (ML) phylogenetic trees using iqtree v1.3.11 [47].
267 The most suitable model was selected using ModelFinder Plus which computes
268 the log-likelihoods of an initial parsimony tree for many different models and the
269 Akaike information criterion (AIC), corrected AIC and Bayesian information
270 criterion (BIC) [48]. To estimate branch supports of the phylogenetic tree inferred
271 from the multiple sequence alignment, ultrafast bootstrap approximation was
272 used [49].

273 **Data availability**

274 The sequences uploaded to generate Agilent SureSelect capture probes are
275 available through Figshare at 10.6084/m9.figshare.12546377. The sequence
276 reads are available in the Sequence Read Archive under project PRJEB39975.
277 For sequence read sets obtained from human samples, reads mapping to the
278 human genome using Bowtie2 were removed from the data before uploading.

279 **Results**

280 A total of 184 small mammals were trapped at 5 sites in Northern Thailand: Ban
281 Thoet Thai (20.24°N, 99.64°E), Mae Fahluang district; Ban Song Kwair (20.02°N,
282 99.75°E) and Ban Mae Khao Tom (20.04°N, 99.95°E) and Ban Mae Mon
283 (19.85°N, 99.61°E), Meuang district in Chiang Rai Province and Ban Huay Muang
284 (19.14°N, 100.72°E), Tha Wang Pha district, Nan Province. One chigger sample
285 was collected on the Penghu Islands, Taiwan (23.57°N, 119.64°E). Human
286 samples were collected from Chiang Rai Province, Northern Thailand, across
287 Laos and one from Green Island, Taiwan (22.66°N, 121.49°E).

288 **Probe design**

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

289 The probes were designed in the following way, aiming to ensure that the full
290 diversity of the *O. tsutsugamushi* genome would be successfully captured. Two
291 finished reference strains (Boryong and Ikeda) plus seven other assemblies
292 available at the time of probe design were used (Gilliam: GCF_000964615.1, Karp:
293 GCF_000964585.1, Kato: GCF_000964605.1, TA716: GCF_000964855.1, TA763:
294 GCF_000964825.1, UT144: GCF_000965195.1, UT76: GCF_000964835.1). The
295 complete Boryong strain was used as a reference genome and the whole genome
296 was included in the probe design. To cover genes not found in the Boryong
297 genome, or which had high levels of divergence from the Boryong genome, the
298 genome assemblies were reannotated using Prokka v1.11 and predicted open
299 reading frames from all eight genomes were clustered into groups based on
300 >=80% identity at the protein sequence level using Roary v3.6.0 [50]. For each
301 cluster, an alignment of the corresponding DNA sequences (using Clustal Omega
302 [51]) was divided into windows of 120 nt in which every aligned sequence was a
303 candidate probe. Probes were then chosen until every sequence in each cluster
304 was represented by a probe with <10% DNA sequence , a strategy informed by
305 previous work demonstrating efficient capture with probe target divergence up to
306 20% [31] and the requirement to capture as-yet uncharacterised sequences. The
307 reference Boryong gene sequence was always included if it had a representative
308 in the cluster under consideration and sequences that would capture human and
309 rodent genomes (*Rattus norvegicus*) were excluded. The probe design strategy
310 generated a total sequence length of 4.7Mb which was synthesised as a single
311 Agilent SureSelect probe pool. The FASTA file containing the sequences uploaded
312 for probe design is available at 10.6084/m9.figshare.12546377.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

313 **Validation using spiked samples**

314 To create the spike-in solution, DNA was extracted from 20 chiggers of the genus
315 *Walchia* that had previously tested negative for *O. tsutsugamushi* using the 47 kDa
316 real-time PCR. DNA extraction was performed using the methods described
317 previously. The 20 extracted DNA samples (40µL each) of negative chiggers were
318 pooled and then split into 20 tubes, such that the sample was equivalent to the
319 mean amount of DNA extracted from a chigger.

320 *O. tsutsugamushi* (strains UT76 and CRF136) DNA extracted from cell culture
321 was used to create the dilution series. The concentration was 838 ng/µL with
322 82% of the DNA being from *O. tsutsugamushi* and 18% from contaminants, (as
323 estimated by qPCR and bulk sequencing of the isolate) giving a starting
324 concentration of *O. tsutsugamushi* of 687 ng/µL. 100,000 copies of *O.*
325 *tsutsugamushi* = 0.227 ng of DNA. 100,000 copies/µL = 0.42 ng/µL of UT76 stock
326 solution (assuming DNA is 82% *O. tsutsugamushi*). To create a final
327 concentration of 0.42 ng/µL equivalent to 100,000 copies/µL: 2 µL of *O.*
328 *tsutsugamushi* DNA was added to 18 µL of water, mixed thoroughly and 5 µL of
329 this removed and added to 45 µL of water, mixed again and then 2 µL added to
330 38 µL water. The following concentrations were made following a dilution series
331 using the prepared *O. tsutsugamushi* and chigger solutions: 100,000, 50,000,
332 25,000, 10,000, 5,000 and 1,000 copies.
333 The results of the spiked sample sequencing are shown in Figure 1 and
334 Supplementary Table 1. Total reads of 2.2×10^5 to 8.5×10^6 were obtained for each
335 sample, with 32-93% of reads mapping to the target genome. Due to the highly
336 repetitive nature of the *O. tsutsugamushi* genome, which varies hugely between

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

337 strains, we chose to measure coverage statistics by using coverage across 657
338 core genes previously identified as present in all samples [19], covering 685kb of
339 the 2.2Mb genome. The proportion of the core genome covered with ≤ 10 reads
340 ranged from 14.3 to 99.8. The percentage of reads which were identified as
341 sequencing duplicates ranged from 51 to 66%, with a greater duplication rate in
342 the samples with lower quantities of target DNA, as expected.

343 **Validation of real samples**

344 The low-input Nextera library preparation method was subsequently applied to
345 human samples. This provided inconsistent results, thought to be driven by low
346 and inconsistent amounts of input DNA leading to low-complexity libraries,
347 highly variable pooling and high duplication rates. We therefore altered the
348 library preparation to include a whole-genome amplification step and re-
349 validated using spiked samples, which resulted in lower duplication rates
350 (Supplementary Figure 1 and Supplementary Table 1); all subsequent batches
351 were sequenced with an initial whole-genome amplification step.

352 Sixty-nine human *O. tsutsugamushi* PCR positive samples from scrub typhus
353 patients were selected from retrospective collections, covering a wide
354 geographical range: 33 from Chiang Rai province in Thailand, 39 from Laos and 1
355 from Taiwan (Figure 2). Among these, 31 were buffy coat samples, 18 whole
356 blood and 20 eschars (including eschar tissue and eschar swabs). The samples
357 include 11 paired samples with whole-blood/buffy coat and eschar samples from
358 patients collected in Chiang Rai (9 pairs) and Laos (2 pairs).

359 Ninety-one *O. tsutsugamushi* PCR positive pooled chigger samples (mean 26
360 individuals per pool) were selected. These were composed of both pure and
361 mixed species pools collected from 36 small mammals, with multiple pools from

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

362 some animals (Supplementary Table 1). A total of 27 *O. tsutsugamushi* PCR
363 positive individual chiggers collected from rodents were included of 8 species in
364 5 genera. These included *L. deliense*, *L. imphalum, Walchia kritochaeta* and *W.*
365 *micropelta*. Chiggers were collected from 5 sites in Northern Thailand and the
366 Penghu Islands, Taiwan. A single free-living chigger (*L. imphalum*) from Ban
367 Thoet Thai was included. *O. tsutsugamushi*-infected colony chiggers from 3
368 different species were included, provided by the Armed Forces Research
369 Institute for Medicine (AFRIMS) in Bangkok, Thailand. Six lung and 3 liver tissue
370 samples were included from 7 small mammals of 3 species. Both liver and lung
371 from the same animal were tested in 2 cases. These were collected from 4 sites in
372 Chiang Rai Province (Figure 2).
373 All samples were PCR positive for the 47kDa gene. The C_t values for the samples
374 ranged from 24.6 to 41.3 cycles.
375 We assessed the sample sequencing based on the number and proportion of
376 reads generated which map to the reference genome, the coverage of the core
377 genes, and the sequence duplication rate (Figure 3). In most samples, only a
378 small proportion of reads mapped to the reference genome, reflecting the
379 performance of the methodology on samples that in general had very small
380 amounts of *O. tsutsugamushi* sequences. Among the different chigger sample
381 types, colony chiggers performed well, with a high percentage of reads mapped
382 to the reference genome likely reflecting their higher input total copy number
383 and corresponding lower C_t (mean 29.4, range 28.6-30.2). Chigger pools and
384 individual chiggers from rodents had high variability but with some samples
385 having high levels of reads mapped to the reference genome and
386 correspondingly a high percentage of the genome covered at 10X coverage. C_t

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

387 values for individual chiggers were higher (mean 36.4, median 37, range 30.2-
388 40.2) compared to chigger pools (mean 31.3, median 30.9, range 24.6-40.3).
389 Among the human samples, buffy coat and eschar samples gave more variable
390 performance, with very few samples having sufficient genome coverage to be
391 used in variant calling, and whole blood performed least well with percentage of
392 the core genome covered at 10X or more under 1% in all samples and median
393 percentage of reads mapped to the reference genome of 0.72%. Rodent tissue
394 samples performed poorly in all cases. The relatively low C_t values for colony
395 chiggers and their high core genome coverage may reflect the unusual ecological
396 scenario of long-term colony chiggers that may result in higher loads of *O.*
397 *tsutsugamushi* than wild chiggers.
398 We expected a positive association between the rate of reads matching *Orientia*
399 sequences and the number of *Orientia* genome copies detectable by qPCR. We
400 compared the fraction of reads which mapped to the C_t values (Supplementary
401 Figure 2). Colony chiggers had the highest fraction of reads mapped to the
402 reference genome and tended to have the lowest C_t (Supplementary Figure 2). A
403 lower C_t (higher input number of genomes) was correlated with the percentage
404 of reads mapped to the reference (Spearman's rank order correlation=-0.70,
405 $p=1.05\times10^{-35}$) (Supplementary Figure 2).
406 The multiple sample types had a wide range of estimated genome copies, as well
407 as different properties such as total DNA content, which change the ratio of
408 target to non-target DNA. Many samples fell near the lower limit of detection of
409 the qPCR assay, with 69/205 (34%) had a C_t value of >35 It appears that a C_t of
410 ≥35 results in poor coverage and low percentage mapping to the reference.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

411 Variant calling was performed on the entire set of sequenced samples. Due to the
412 low sequence coverage for many samples, phylogenetic comparisons were
413 attempted only for a set of 31 samples with >50,000 bases called: 4 chigger pools
414 from Ban Mae Mon, Thailand, 1 human buffy coat sample from Na Meuang, Laos,
415 1 individual chigger from the Penghu Islands, Taiwan, 4 individual chiggers and
416 17 chigger pools from Ban Thoet Thai, Thailand, and 4 colony chiggers. The
417 median C_t value for these samples was 29.0 (range 25.4-34.2). The distribution of
418 bases called for these 31 samples is shown in Supplementary Figure 4. Coverage
419 for each core gene is shown in the heatmap in Supplementary Figure 5. For
420 almost all samples, there is some sequence coverage for each of the core genes,
421 and for those with fewer positions called it is due to incomplete coverage across
422 the genome rather than genes which are completely uncovered in sequencing. A
423 notable exception is sample C0546, which has many genes which have no
424 coverage at all but sufficient coverage in the remaining genes to meet the
425 50,000bp threshold. A small number of genes were completely uncovered in
426 multiple samples, most notably several genes which have no coverage in any of
427 the samples taken from the R240 pools from a rodent in Ban Mae Mon.
428 The phylogeny is show in Figure 4. Branch bootstrap values, which can be
429 interpreted as the relative (%) support of the data for the tree topology
430 represented by the pairings of isolates or groups of isolates on either side of the
431 labelled branch, are plotted on the tree and fall below 70% support for some
432 branches, indicating some uncertainty in tree topology. The samples include two
433 colony chiggers from the same *L. deliense* colony. These samples are closely
434 related but not identical (35 SNPs between the two samples).
435

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

436 **Discussion**

437 We have successfully developed and tested the first whole-genome sequencing of
438 *O. tsutsugamushi* performed without prior cell culture. The sequence data
439 generated provided an opportunity to compare *O. tsutsugamushi* strains with
440 greater resolution than previously possible.

441 The sequencing results displayed great variability, with sufficient success to call
442 variants and perform phylogenetic analysis in a proportion of samples from
443 individual chiggers and chigger pools. The yield of unique on-target reads,
444 particularly at the low copy number dilutions (5,000 and 10,000 copies) was
445 higher for WGA before library preparation than for Nextera XT, and the
446 duplication rate was also improved. The low success rate likely reflects very low
447 quantities of *O. tsutsugamushi* DNA present in many samples, especially human
448 samples, and reflects the current limit of our enrichment method which cannot
449 enrich sufficiently to overcome the low levels of input DNA. While no firm C_t
450 cutoff value can be established above which target enrichment sequencing
451 cannot be successfully performed, samples with a C_t value of 35 or less are
452 candidates for sequencing. Methods for human and rodent DNA depletion prior
453 to sequence capture may improve the performance of enrichment. The first full
454 genome of *L. deliense* has been published since this array was designed, and this
455 could be used to check for any sequences in the array design which may capture
456 off-target chigger DNA [52].

457 A recent study has reported phylogenetic comparisons of *O. tsutsugamushi*
458 strains from chiggers collected from the same host animal, based on sequencing
459 of a single gene (encoding the 56 kDa antigen) [53]. Results revealed mixed
460 infections; with some chiggers containing a single genotype and others mixed

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

461 genotypes. There is also evidence of different *O. tsutsugamushi* 56 kDa type-
462 specific antigen genotypes being maintained and transmitted transovarially in
463 colony chiggers [54].
464 The sequence capture probes used in this experiment were designed when only
465 two complete genomes were available to use in the design process. Of the
466 incomplete assemblies included in the design process, two strains have been
467 removed from RefSeq due to problems with the assembly, and more complete
468 genomes are now available. A new probe design using the same approach but
469 more genomes may improve the capture efficiency.
470 Despite the poor performance for the target enrichment sequencing on some
471 samples, we were able to generate a phylogeny using 30 chigger samples, 1
472 human sample, and 8 complete reference genomes, which represents the first
473 phylogenetic analysis of *O. tsutsugamushi* from chiggers. Among the 31 best-
474 sequenced samples, >98.5% of the core genes of the reference sequence were
475 covered by at least one read at all positions. For most samples, the regions of no
476 coverage were confined to a very few genes, some of which were present in all
477 samples. Intriguingly, for chigger pools from Ban Mae Mon (R240), more genes
478 were incompletely covered, and most of these were present in all samples, even
479 though the total volume of on-target reads (equivalently, the average coverage of
480 the core genome) was similar in these samples as in other high-performing
481 samples. This could be due to diversity in these genes beyond the limits of what
482 our probes are able to capture; however, the sequence capture probes have been
483 shown to be effective at up to 20% sequence divergence [31], and the overall
484 diversity between our phylogenetic samples is well below this limit. It is more
485 likely that the set of core genes determined from the known complete genomes is

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

486 not universally present in all strains.
487 The study included strains sequenced from chiggers collected from a single host
488 animal, strains from chiggers from several animals at a single study site of
489 <10km^2 and from two sites 45km apart. Samples from Ban Mae Mon are clearly
490 distinct from samples from Ban Thoet Thai, which group together (Figure 4). All
491 the chigger pools and individuals from Ban Thoet Thai consisted of the known
492 vector *L. imphalum* (with or without some *Walchia* species). The Taiwanese
493 chigger was the known human vector *L. deliense*. The R240 pools from Ban Mae
494 Mon, which form a distinct cluster separate from all other samples, were
495 collected from the scansorial tree shrew *Tupaia glis* and consisted of *L. turdicola*
496 and *Helenicula naresuani* chiggers – neither known to be human vectors nor
497 previously reported as being infected with *O. tsutsugamushi*. The reference
498 genomes, which were collected from five different countries between 1943 and
499 2010, are spread throughout the tree and many are more closely related to the
500 samples from Ban Thoet Thai than the samples from Ban Mae Mon are to those
501 samples. A possible explanation for this is that *O. tsutsugamushi* has been
502 previously introduced into these two locations from divergent sources and
503 continues to evolve locally on a small scale, and larger-scale *O, tsutsugamushi*
504 movement between locations is a rare event due to the restricted range of the
505 host species.
506 Important questions remain about the role of recombination between strains in
507 infected chiggers and to what extent the accessory genome of *Orientia* is open or
508 closed. The sequence capture approach used in this study does not recover the
509 complete accessory genome, and hence cannot assist with the latter question.
510 The accumulation of more high-quality sequences may allow characterization of

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

511 the recombination landscape. However, *O. tsutsugamushi* genomes are known to
512 have poorly conserved synteny, which is likely to complicate analysis of
513 incomplete genomes.

514 Among captured sequences, pairwise divergences were in the range of 0-4%,
515 well within the reach of probe-based sequence enrichment for pathogen
516 genomics [31]. This illustrates the robustness and adaptability of probe-based
517 sequence enrichment, providing a means for genome-wide amplification of
518 sequence information without the need to validate a very large number of PCR
519 primers, any of which could fail because of hitherto uncharacterised sequence
520 variation.

521 The methods developed in this project have, for the first time in scrub typhus
522 research, demonstrated phylogeographic clustering of *O. tsutsugamushi* strains
523 at international, provincial and highly local scales. This shows that both closely
524 related and more distantly related strains may co-exist in one site. As methods
525 improve and can be applied to a greater range of samples, particularly sympatric
526 rodents and exposed humans, further insights into this fascinating
527 phylogeographic variation will be revealed with important consequences for
528 diagnostic tests and vaccine development strategies.

529 **Acknowledgments**

530 We are very grateful to Associate Professor Bounthaphany Bounxouei, ex-
531 Director of Mahosot Hospital, the Director and staff of the Microbiology
532 Laboratory, LOMWRU and wards, Assistant Professor Chanphomma
533 Vongsamphan, ex-Director of Department of Health Care, Ministry of Health, and
534 H.E. Professor Bounkong Syhavong, Minister of Health, Laos, for their help and
535 support. We thank the Director and staff of the Microbiology Laboratory and the

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

536 staff of LOMWRU for their wonderful help, Dr Chi-Chien Kuo and colleagues at
537 the National Taiwan Normal University, Taipei for facilitating chigger collection
538 on the Penghu Islands. We thank Sebastiaan Van Hal for providing the human
539 sample from Taiwan. We are very grateful to Rawadee Kumlert at Mahidol
540 University for her assistance in mite morphotyping, and all the staff at the
541 Chiangrai Clinical Research Unit and field teams, in particular Piangnet Jaiboon,
542 Dr Tri Wangrangsimakul, Dr Serge Morand and Dr Kittipong Chaisiri. We thank
543 Prof Alistair Darby for comments on the manuscript.
544 Material has been reviewed by the Walter Reed Army Institute of Research.
545 There is no objection to its presentation and/or publication. The opinions or
546 assertions contained herein are the private views of the author, and are not to be
547 construed as official, or as reflecting true views of the Department of the Army or
548 the Department of Defense. Research was conducted under an approved animal
549 use protocol in an AAALACi accredited facility in compliance with the Animal
550 Welfare Act and other federal statutes and regulations relating to animals and
551 experiments involving animals and adheres to principles stated in the Guide for
552 the Care and Use of Laboratory Animals, NRC Publication, 2011 edition.
553

554 **Financial Support**

555 This study was supported by Ivo Elliott's Wellcome Trust Research Training
556 Fellowship (105731/Z/14/Z and in part by Core Awards to the Wellcome Centre
557 for Human Genetics (090532/Z/09/Z and 203141/Z/16/Z) and by the Wellcome
558 Trust Core Award Grant Number 203141/Z/16/Z with additional support from
559 the NIHR Oxford BRC. The views expressed are those of the author(s) and not
560 necessarily those of the NHS, the NIHR or the Department of Health.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

561

562 **CReDiT Author statement**

563 **Ivo Elliott:** Conceptualization, Methodology, Formal analysis, Investigation,
564 Writing -Original Draft, Funding acquisition. **Neeranuch Thangnimitchok:**
565 Investigation, Writing – Review & Editing. **Mariateresa de Cesare:** Investigation,
566 Validation, Writing – Review & Editing. **Piyada Linsuwanon:** Resources. **Daniel**
567 **Paris:** Conceptualization, Methodology, Supervision, Writing – Review & Editing.
568 **Nicholas Day:** Conceptualization, Writing – Review & Editing, Supervision. **Paul**
569 **Newton:** Conceptualization, Writing – Review & Editing, Supervision. **Rory**
570 **Bowden:** Conceptualization, Methodology, Formal analysis, Writing – Review &
571 Editing, Supervision, Funding acquisition. **Elizabeth Batty:** Conceptualization,
572 Methodology, Software, Validation, Formal analysis, Data Curation, Writing –
573 Original Draft, Supervision.

574

575 **Conflict of interests**:

576 IE, NT, MdC, PL, DHP, NDJP, PNN, RB, EMB - none

577

578 **References**

579 1. Izzard L, Fuller A, Blacksell SD, et al. Isolation of a novel Orientia species
580 (O. chuto sp. nov.) from a patient infected in Dubai. *J Clin Microbiol* 2010; **48**(12):
581 4404-9.
582 2. Weitzel T, Dittrich S, Lopez J, et al. Endemic Scrub Typhus in South
583 America. *N Engl J Med* 2016; **375**(10): 954-61.
584 3. Kolo AO, Sibeko-Matjila KP, Maina AN, Richards AL, Knobel DL, Matjila PT.
585 Molecular Detection of Zoonotic Rickettsiae and Anaplasma spp. in Domestic
586 Dogs and Their Ectoparasites in Bushbuckridge, South Africa. *Vector Borne*
587 *Zoonotic Dis* 2016; **16**(4): 245-52.
588 4. Cosson JF, Galan M, Bard E, et al. Detection of Orientia sp. DNA in rodents
589 from Asia, West Africa and Europe. *Parasit Vectors* 2015; **8**: 172.
590 5. Masakhwe C, Linsuwanon P, Kimita G, et al. Identification and
591 characterization of Orientia chuto in trombiculid chigger mites collected from
592 wild rodents in Kenya. *J Clin Microbiol* 2018; **56**: e01124-18.
593 6. Rapmund G, Upham RW, Jr., Kundin WD, Manikumaran C, Chan TC.
594 Transovarial development of scrub typhus rickettsiae in a colony of vector mites.
595 *Trans R Soc Trop Med Hyg* 1969; **63**(2): 251-8.
596 7. Frances SP, Watcharapichat P, Phulsuksombati D. Vertical transmission of
597 Orientia tsutsugamushi in two lines of naturally infected Leptotrombidium
598 deliense (Acari: Trombiculidae). *J Med Entomol* 2001; **38**(1): 17-21.
599 8. Urakami H, Okubo K, Misumi H, Fukuhara M, Takahashi M. Transovarial
600 transmission rates of Orientia tsutsugamushi in naturally infected

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

601 Leptotrombidium colonies by immunofluorescent microscopy. *Med Entomol Zool*
602 2013; **64**(1): 43-6.
603 9. Lerdthusnee K, Khuntirat B, Leepitakrat W, et al. Scrub typhus: vector
604 competence of Leptotrombidium chiangraiensis chiggers and transmission
605 efficacy and isolation of Orientia tsutsugamushi. *Ann N Y Acad Sci* 2003; **990**: 25-
606 35.
607 10. Santibanez P, Palomar AM, Portillo A, Santibanez S, Oteo JA. The role of
608 chiggers as human pathogens. In: Samie A, ed. An overview of tropical diseases:
609 InTech; 2015: 173-202.
610 11. Elliott I, Pearson I, Dahal P, Thomas NV, Roberts T, Newton PN. Scrub
611 typhus ecology: a systematic review of Orientia in vectors and hosts. *Parasit*
612 *Vectors* 2019; **12**(1): 513.
613 12. Kelly DJ, Fuerst PA, Ching WM, Richards AL. Scrub typhus: the geographic
614 distribution of phenotypic and genotypic variants of Orientia tsutsugamushi. *Clin*
615 *Infect Dis* 2009; **48 Suppl 3**: S203-30.
616 13. Kim G, Ha NY, Min CK, et al. Diversification of Orientia tsutsugamushi
617 genotypes by intragenic recombination and their potential expansion in endemic
618 areas. *PLoS Negl Trop Dis* 2017; **11**(3): e0005408.
619 14. Arai S, Tabara K, Yamamoto N, et al. Molecular phylogenetic analysis of
620 Orientia tsutsugamushi based on the groES and groEL genes. *Vector Borne*
621 *Zoonotic Dis* 2013; **13**(11): 825-9.
622 15. Duong V, Blassdell K, May TT, et al. Diversity of Orientia tsutsugamushi
623 clinical isolates in Cambodia reveals active selection and recombination process.
624 *Infect Genet Evol* 2013; **15**: 25-34.
625 16. Phetsouvanh R, Sonthayanon P, Pukrittayakamee S, et al. The Diversity
626 and Geographical Structure of Orientia tsutsugamushi Strains from Scrub Typhus
627 Patients in Laos. *Plos Negl Trop Dis* 2015; **9**(8): e0004024.
628 17. Jiang J, Paris DH, Blacksell SD, et al. Diversity of the 47-kD HtrA nucleic
629 acid and translated amino acid sequences from 17 recent human isolates of
630 Orientia. *Vector Borne Zoonotic Dis* 2013; **13**(6): 367-75.
631 18. Sonthayanon P, Peacock SJ, Chierakul W, et al. High rates of homologous
632 recombination in the mite endosymbiont and opportunistic human pathogen
633 Orientia tsutsugamushi. *PLoS Negl Trop Dis* 2010; **4**(7): e752.
634 19. Batty EM, Chaemchuen S, Blacksell S, et al. Long-read whole genome
635 sequencing and comparative analysis of six strains of the human pathogen
636 Orientia tsutsugamushi. *PLoS Negl Trop Dis* 2018; **12**(6): e0006566.
637 20. Giengkam S, Blakes A, Utsahajit P, et al. Improved quantification,
638 propagation, purification and storage of the obligate intracellular human
639 pathogen Orientia tsutsugamushi. *PLoS Negl Trop Dis* 2015; **9**(8).
640 21. Blacksell SD, Robinson MT, Newton PN, Day NPJ. Laboratory-acquired
641 scrub typhus and murine typhus infections: The argument for risk-based
642 approach to biosafety requirements for Orientia tsutsugamushi and Rickettsia
643 typhi laboratory activities. *Clin Infect Dis* 2018.
644 22. Sonthayanon P, Chierakul W, Wuthiekanun V, et al. Association of high
645 Orientia tsutsugamushi DNA loads with disease of greater severity in adults with
646 scrub typhus. *J Clin Microbiol* 2009; **47**(2): 430-4.
647 23. Linsuwanon P, Krairojananan P, Rodkvamtook W, Leepitakrat S, Davidson
648 S, Wanja E. Surveillance for Scrub Typhus, Rickettsial Diseases, and Leptospirosis

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

649 in US and Multinational Military Training Exercise Cobra Gold Sites in Thailand.
650 *US Army Med Dep J* 2018; (1-18): 29-39.
651 24. Darby AC, Cho NH, Fuxelius HH, Westberg J, Andersson SG. Intracellular
652 pathogens go extreme: genome evolution in the Rickettsiales. *Trends Genet* 2007;
653 **23**(10): 511-20.
654 25. Nakayama K, Kurokawa K, Fukuhara M, et al. Genome comparison and
655 phylogenetic analysis of Orientia tsutsugamushi strains. *DNA Res* 2010; **17**(5):
656 281-91.
657 26. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of
658 next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**(6): 333-51.
659 27. Mertes F, Elsharawy A, Sauer S, et al. Targeted enrichment of genomic
660 DNA regions for next-generation sequencing. *Brief Funct Genomics* 2011; **10**(6):
661 374-86.
662 28. Summerer D. Enabling technologies of genomic-scale sequence
663 enrichment for targeted high-throughput sequencing. *Genomics* 2009; **94**(6):
664 363-8.
665 29. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing
666 using targeted sequence capture. *Genome Res* 2015; **25**(12): 1910-20.
667 30. O'Flaherty BM, Li Y, Tao Y, et al. Comprehensive viral enrichment enables
668 sensitive respiratory virus genomic identification and analysis by next
669 generation sequencing. *Genome Res* 2018; **28**(6): 869-77.
670 31. Bonsall D, Ansari MA, Ip C, et al. ve-SEQ: Robust, unbiased enrichment for
671 streamlined detection and whole-genome sequencing of HCV and other highly
672 diverse pathogens. *F1000Res* 2015; **4**: 1062.
673 32. Clark SA, Doyle R, Lucidarme J, Borrow R, Breuer J. Targeted DNA
674 enrichment and whole genome sequencing of Neisseria meningitidis directly
675 from clinical specimens. *Int J Med Microbiol* 2018; **308**(2): 256-62.
676 33. Herbreteau V, Jittapalapong S, Rerkamnuaychoke W, Chaval Y, Cosson JF,
677 Morand S. Protocols for field and laboratory rodent studies. Bangkok, Thailand:
678 Kasetsart University Press; 2011.
679 34. Sikes RS. The animal care and use committee of the American Society of
680 Mammalogists. 2016 Guidelines of the American Society of Mammalogists for the
681 use of wild mammals in research and education. *J Mammal* 2016; **97**(3): 663-88.
682 35. AVMA Panel on Euthanasia. AVMA Guidelines for the euthanasia of
683 animals: American Veterinary Medical Association; 2013.
684 36. Gentry JW. Black plate collections of unengorged chiggers. *Singapore Med*
685 *J* 1965; **1**(1): 46.
686 37. Uchikawa K, Kawamori F, Kawai S, Kumada N. Suzuki's method (Mitori-
687 ho) a recommended method for the visual sampling of questing
688 Leptotrombidium scutellare larvae in the field (Trombidiformes, Trombiculidae).
689 *J Acarol Soc Jpn* 1993; **2**(2): 91-8.
690 38. Wangrangsimakul T, Greer RC, Chanta C, et al. Clinical Characteristics and
691 Outcome of Children Hospitalized With Scrub Typhus in an Area of Endemicity. *J*
692 *Pediatric Infect Dis Soc* 2020; **9**(2): 202-9.
693 39. Kumlert R, Chaisiri K, Anantatat T, et al. Autofluorescence microscopy for
694 paired-matched morphological and molecular identification of individual chigger
695 mites (Acari: Trombiculidae), the vectors of scrub typhus. *PLoS One* 2018; **13**(3):
696 e0193163.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

697 40. Nadchatram M, Dohany AL. A pictorial key to the subfamilies, genera and
698 subgenera of Southeast Asian chiggers (Acari, Prostigmata, Trombiculidae).
699 *Institute for Medical Research, Kuala Lumpur, Malaysia* 1974; **Bulletin number**
700 **16**.
701 41. Vercammen-Grandjean PH. The chigger mites of the Far East. Special
702 study. Washington D.C.: U.S. Army Medical Research and Development
703 Command; 1968.
704 42. Stekolnikov AA. Leptotrombidium (Acari: Trombiculidae) of the World.
705 *Zootaxa* 2013; **3728**(1): 1-173.
706 43. Jiang J, Chan TC, Temenak JJ, Dasch GA, Ching WM, Richards AL.
707 Development of a quantitative real-time polymerase chain reaction assay specific
708 for Orientia tsutsugamushi. *Am J Trop Med Hyg* 2004; **70**(4): 351-6.
709 44. Li H. Aligning sequence reads, clone sequences and assembly contigs with
710 BWA-MEM. 2013. https://arxiv.org/abs/1303.3997 (accessed 30/01/2018.
711 45. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a
712 MapReduce framework for analyzing next-generation DNA sequencing data.
713 *Genome Res* 2010; **20**(9): 1297-303.
714 46. Seemann T. Snippy: fast bacterial variant calling from NGS reads. 2012.
715 https://github.com/tseemann/snippy.
716 47. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and
717 effective stochastic algorithm for estimating maximum-likelihood phylogenies.
718 *Mol Biol Evol* 2015; **32**(1): 268-74.
719 48. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS.
720 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
721 *Methods* 2017; **14**(6): 587-9.
722 49. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2:
723 Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 2018; **35**(2):
724 518-22.
725 50. Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote
726 pan genome analysis. *Bioinformatics* 2015; **31**(22): 3691-3.
727 51. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality
728 protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 2011; **7**:
729 539.
730 52. Dong X, Chaisiri K, Xia D, et al. Genomes of trombidid mites reveal novel
731 predicted allergens and laterally-transferred genes associated with secondary
732 metabolism. *Gigascience* 2018; **7**(12).
733 53. Takhampunya R, Korkusol A, Promsathaporn S, et al. Heterogeneity of
734 Orientia tsutsugamushi genotypes in field-collected trombiculid mites from wild-
735 caught small mammals in Thailand. *PLoS Negl Trop Dis* 2018; **12**(7): e0006632.
736 54. Takhampunya R, Tippayachai B, Korkusol A, et al. Transovarial
737 Transmission of Co-Existing Orientia tsutsugamushi Genotypes in Laboratory-
738 Reared Leptotrombidium imphalum. *Vector Borne Zoonotic Dis* 2016; **16**(1): 33-
739 41.
740

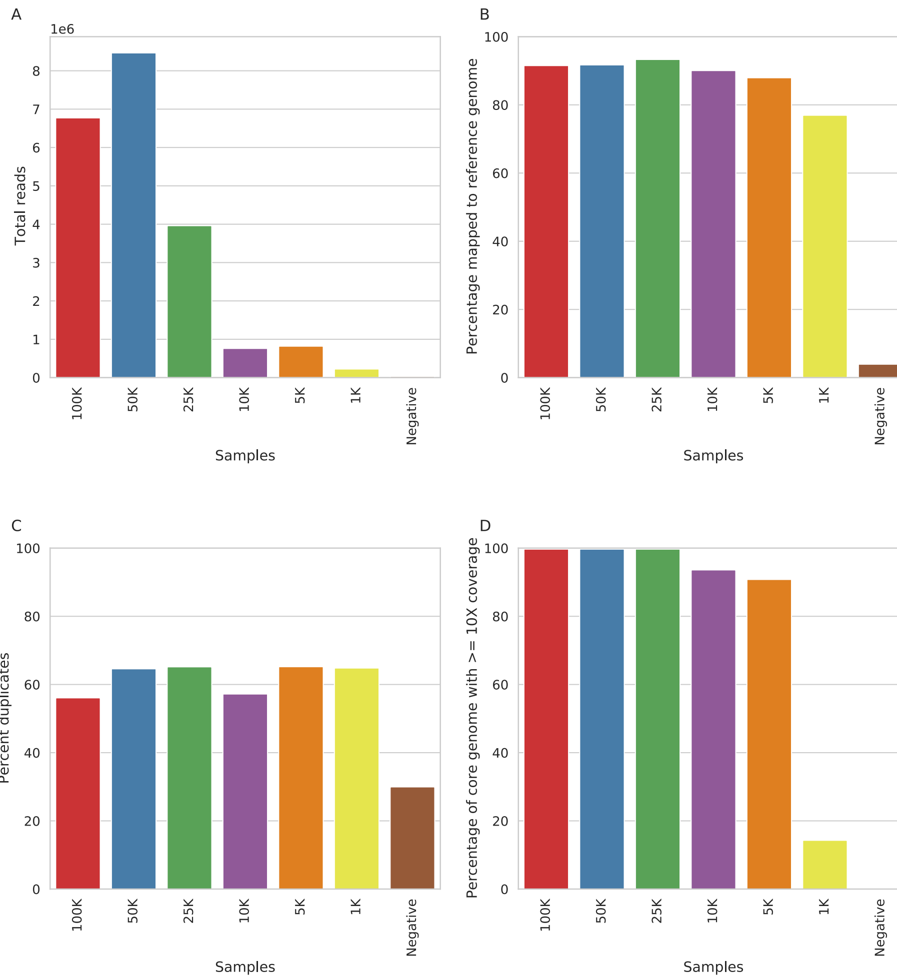
741

742

743

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

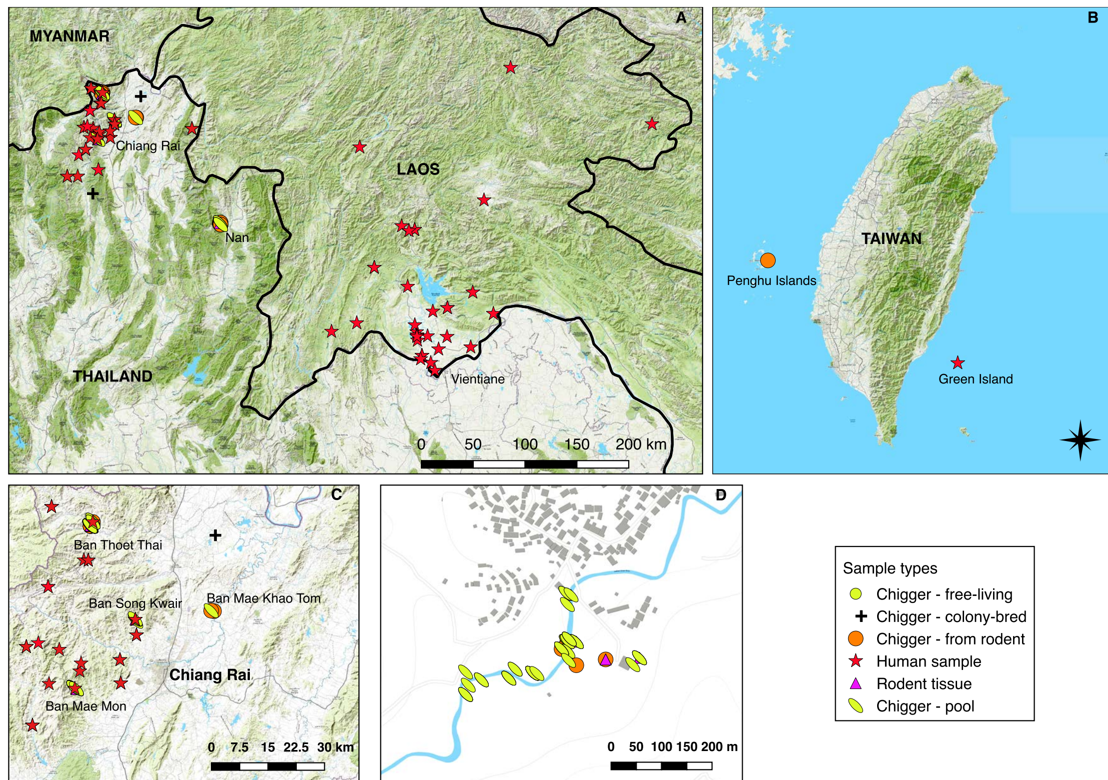
744 **Figures**



745

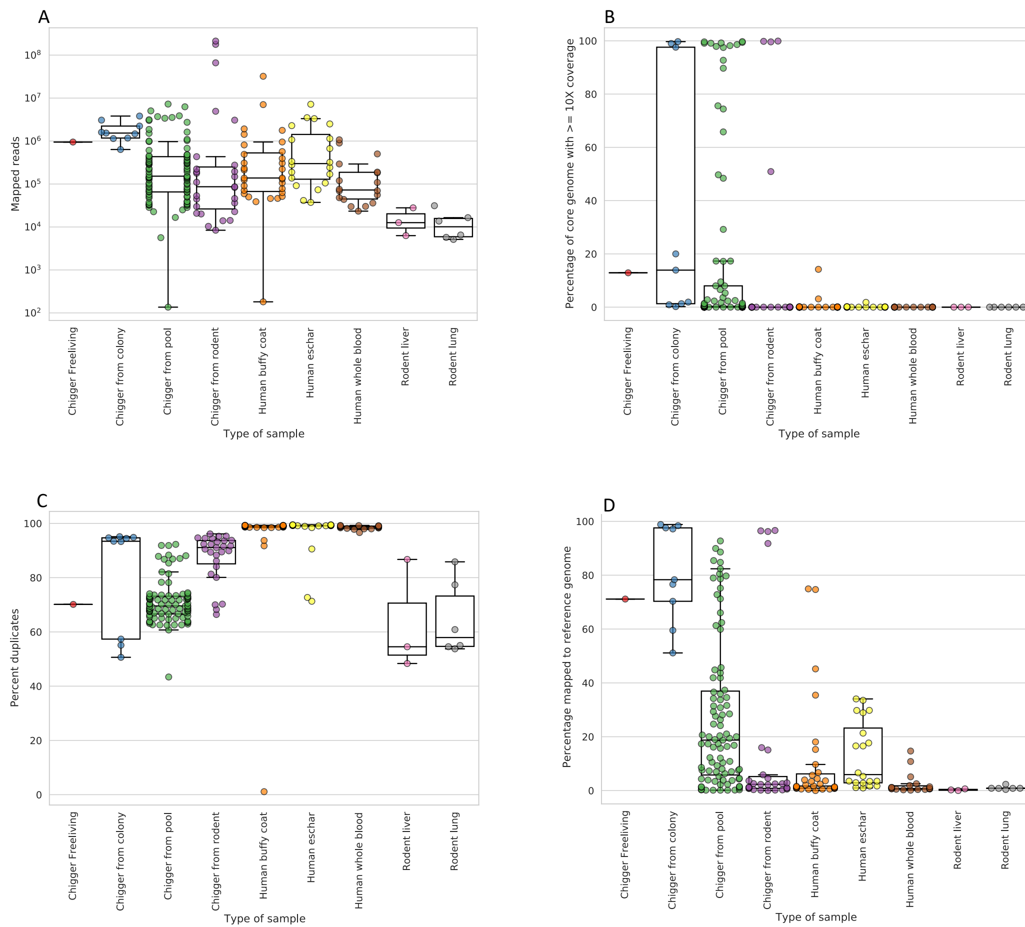
746 Figure 1. Results from sequencing of spike-in control samples showing a) total
747 reads produced b) percentage of those reads which mapped to the reference
748 genome c) percentage of the reads which were duplicates and d) the percentage
749 of the core genome covered by 10 or more reads.
750

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



751
752
753 Figure 2. Sample collection locations. A) Southeast Asia with locations in Laos
754 and Northern Thailand, B) Taiwan, C) Chiang Rai Province, with key field sites
755 named, D) Ban Thoet Thai, Chiang Rai Province, site of the greatest number of *O.*
756 *tsustsugamushi* PCR positive chigger and rodent samples.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

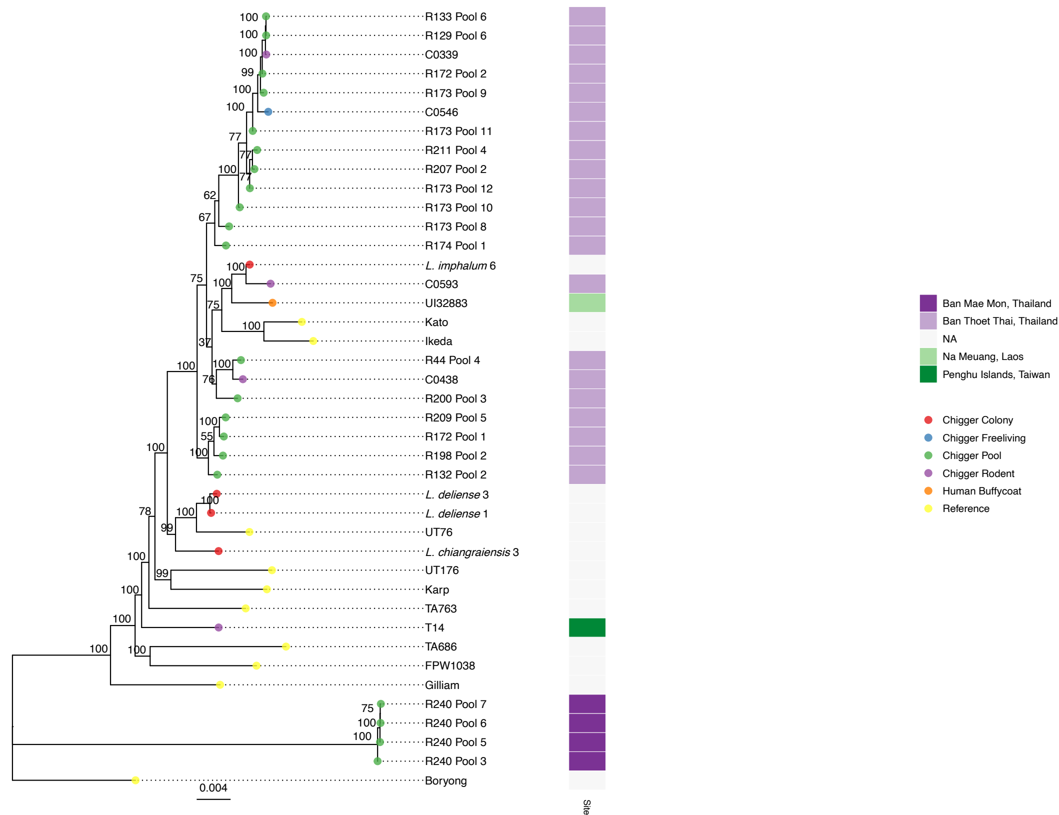


757

758

759 Figure 3. Sequencing statistics for human, chigger, and rodent samples. Panels
760 show a) total number of reads and b) the percentage of reads which were
761 mapped to the reference genome. Panel c) shows the sequence duplication rate
762 and d) shows the coverage of the core genome.
763

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.07.425812; this version posted January 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



764
765 Figure 4: A maximum-likelihood phylogenetic tree produced using IQTREE from
766 all samples that have >50kb of called positions. Tip colors represent the source
767 of each sample, and the heatmap shows the site where samples were collected.
768 The node labels show ultrafast bootstrap support values.