Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541–750)

Marcel Keller[a,b,1,2], Maria A. Spyrou[a,1], Christiana L. Scheib[c,d], Andreas Kröpelin[a,e], Brigitte Haas-Gebhard[f], Bernd Päffgen[g], Jochen Haberstroh[h], Albert Ribera i Lacomba[i], Claude Raynaud[j], Craig Cessford[c], Peter Stadler[k], Kathrin Nägele[a], Gunnar U. Neumann[a], Jessica S. Bates[c], Bernd Trautmann[b], Sarah Inskip[l], Joris Peters[b,m], John E. Robb[c], Toomas Kivisild[c,n], Michael McCormick[o,p], Kirsten I. Bos[a], Michaela Harbeck[b,2], Alexander Herbig[a,2], Johannes Krause[a,p,2]

[a]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, 07745 Jena, Germany

[b]SNSB, State Collection of Anthropology and Palaeoanatomy Munich, Karolinenplatz 2a, 80333 Munich, Germany

[c]Department of Archaeology, University of Cambridge, Downing St, Cambridge CB2 3ER, United Kingdom

[d]Institute of Genomics, University of Tartu, Riia 23b, 51010 Tartu, Estonia

[e]Friedrich Schiller University Jena, Fürstengraben 1, 07743 Jena, Germany

[f]Archaeological Collection of the Bavarian State, Himbselstraße 1, 80538 Munich, Germany

[g]Institute for Pre- and Protohistoric Archaeology and Archaeology of the Roman Provinces, Ludwig Maximilian University Munich, Schellingstraße 12, 80799 Munich, Germany

[h]Bavarian State Department of Monuments and Sites, Hofgraben 4, 80539 Munich, Germany

[i]Department for Municipal Archaeology, Valencia City Council, Calle Traginers s/n 46014 Valencia, Spain

[j]CNRS, UMR5140, Archéologie des Sociétés Méditerranéennes, route de Mende, 34199 Montpellier, France

[k]Department of Pre- and Protohistory, University of Vienna, Franz-Klein-Gasse 1, 1190 Vienna, Austria

[l]McDonald Institute for Archaeological Research, University of Cambridge, Downing St Cambridge CB2 3ER, United Kingdom

[m]ArchaeoBioCenter and Department of Veterinary Sciences, Institute of Palaeoanatomy, Domestication Research and the History of Veterinary Medicine, Ludwig Maximilian University Munich, Kaulbachstr. 37/III, 80539 Munich, Germany

[n]Department of Human Genetics, Katholieke Universiteit Leuven, Leuven, 3000, Belgium

35 [o]Initiative for the Science of the Human Past, Department of History, Harvard University,

36 Robinson Hall M-03, 35 Quincy Street, Cambridge, MA 02138, USA

37 [p]Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean

38

39 [1]These authors contributed equally to this work.

40

41 [2]To whom correspondence may be addressed:

42 Marcel Keller, keller@shh.mpg.de, Department of Archaeogenetics, Max Planck Institute for

43 the Science of Human History, Kahlaische Straße 10, 07745 Jena, Germany

44 Michaela Harbeck, harbeck@snsb.de, SNSB, State Collection of Anthropology and

45 Palaeoanatomy Munich, Karolinenplatz 2a, 80333 Munich, Germany

46 Alexander Herbig, herbig@shh.mpg.de, Department of Archaeogenetics, Max Planck Institute

47 for the Science of Human History, Kahlaische Straße 10, 07745 Jena, Germany

48 Johannes Krause, krause@shh.mpg.de, Department of Archaeogenetics, Max Planck Institute

49 for the Science of Human History, Kahlaische Straße 10, 07745 Jena, Germany

50

51 Author contributions:

52 M.McC., M.H., A.H. and J.K. designed the study. M.K., M.A.S, C.L.S., K.N., G.U.N. and

53 J.S.B. performed laboratory work. A.K. developed the new analytical tool. M.K. and M.A.S.

54 performed data analyses. B.T. and S.I. performed anthropological examination. C.L.S. and T.K.

55 provided genomic data. B.H.-G., B.P., J.H., A.R.iL., C.R., P.S., J.P., J.E.R., and M.H. identified

56 and provided access to archaeological material. B.H.-G., B.P., J.H., A.R.iL., C.R., C.C., P.S.

57 and M.McC. provided archaeological and historical information. M.K., M.A.S, M.McC. and

58 A.H. wrote the paper with contributions from all authors.

Abstract

The first historically documented pandemic caused by *Yersinia pestis* started as the Justinianic Plague in 541 within the Roman Empire and continued as the so-called First Pandemic until 750. Although palaeogenomic studies have previously identified the causative agent as *Y. pestis*, little is known about the bacterium's spread, diversity and genetic history over the course of the pandemic.

To elucidate the microevolution of the bacterium during this time period, we screened human remains from 20 sites in Austria, Britain, Germany, France and Spain for *Y. pestis* DNA and reconstructed six new genomes. We present a novel methodological approach assessing SNPs in ancient bacterial genomes, facilitating qualitative analyses of low coverage genomes from a metagenomic background. Phylogenetic analysis reveals the existence of previously undocumented *Y. pestis* diversity during the 6th–7th centuries, and provides evidence for the presence of multiple distinct *Y. pestis* strains in Europe. We offer genetic evidence for the presence of the Justinianic Plague in the British Isles, previously only hypothesized from ambiguous documentary accounts, as well as southern France and Spain, and that southern Germany seems to have been affected by at least two distinct *Y. pestis* strains. Four of the reported strains form a polytomy similar to others seen across the *Y. pestis* phylogeny, associated with the Second and Third Pandemics. We identified a deletion of a 45 kb genomic region in the most recent First Pandemic strain affecting two virulence factors, intriguingly overlapping with a deletion found in 17th–18th-century genomes of the Second Pandemic.

Significance Statement

The first historically reported pandemic attributed to *Yersinia pestis* started with the Justinianic Plague (541−544) and continued for around 200 years as the so-called First Pandemic. To date, only one *Y. pestis* strain from this pandemic has been reconstructed using ancient DNA. In this study, we present six new genomes from Britain, France, Germany and Spain, demonstrating the geographic range of plague during the First pandemic and showing microdiversity in the Early Medieval Period. Moreover, we detect similar genome decay during the First and Second Pandemic (17th to 18th century) that includes the same two virulence factors, thus providing an example of potential convergent evolution of *Y. pestis* during large scale epidemics.

Introduction

Yersinia pestis, the causative agent of plague, is a Gram-negative bacterium that predominantly infects rodents and is transmitted by their ectoparasites such as fleas. As a zoonosis, it is also able to infect humans with a mortality rate of 50–100 % without antibiotic treatment (1), manifesting as bubonic, septicaemic or bubonic plague. After the pathogen spread worldwide at the end of the 19th century in the so-called Third Pandemic that started in 1855 in Yunnan, China, it established new local foci in Africa and the Americas in addition to the ancient foci that exist in Central and East Asia. Today, *Y. pestis* causes sporadic infections every year and even local recurrent epidemics such as documented in 2017 in Madagascar (2).

Although recent palaeogenetic analyses have been able to reconstruct an ancient form of *Y. pestis* that infected humans as early as in prehistoric times (2,800 to 1,700 BCE (3–5)) the First Pandemic (541–750) is the earliest historically recorded pandemic that has been clearly attributed to *Y. pestis* (6, 7), starting with the fulminant Justinianic Plague (541–544). It was later followed by the Second Pandemic, which started with the Black Death of 1347–1353 (8, 9) and persisted in Europe until the 18th century (10–12).

First attempts in the 2000s aimed to amplify *Y. pestis*-specific DNA fragments from burials of the 6th century (13–15). Although some early studies are controversial due to methodological limitations (16) and proved inconsistent with later work (17), more recent studies have been successful in reconstructing and authenticating whole *Y. pestis* genomes from two early medieval burial sites in modern-day Bavaria, Germany (6, 7).

These genomic investigations identified a previously unknown lineage associated with the First Pandemic that was found to be genetically identical in both sites and falls within the modern diversity of *Y. pestis*. Moreover, this lineage is distinct from those associated with the Second Pandemic that started ca. 800 years later, indicating two independent emergence events.

Although these studies have unequivocally demonstrated the involvement of *Y. pestis* in the First Pandemic, the published genomes represent a single outbreak, leaving the genetic diversity of that time entirely unexplored. Here, we assess the diversity and microevolution of *Y. pestis* during that time by analysing multiple and mass burials in a broader temporal and spatial scope than previously attempted. After screening 167 samples from 20 archaeological sites, we were able to reconstruct six new genomes with higher than 5-fold mean coverage from Britain, France, Germany and Spain. Furthermore, we identified a large deletion in the most recent First Pandemic strain that affects the same region as a deletion observed in late Second Pandemic strains, suggesting similar mechanism of pathogen adaptation in the waning period of the two separate pandemics.

4

Results

<u>Screening and Capture</u>

We used a previously described quantitative PCR assay (18) that targets the *Y. pestis*-specific *pla* gene on the pPCP1 plasmid to test 145 teeth from a minimum of 96 individuals from 19 sites (Table S1). All 19 PCR-positive extracts were subsequently turned into double-stranded libraries and enriched for *Y. pestis* DNA following an in-solution capture approach (19). Whereas some samples reached up to 9.6-fold chromosomal mean coverage after whole genome capture, four of the PCR-positive samples yielded a coverage of lower than 0.1-fold. Since the qPCR assay can amplify non-specific products and subsequent capture can enrich for environmental DNA that sporadically maps to the *Y. pestis* reference, it is crucial to differentiate between samples that show low DNA preservation and those that are false positives.

False positive samples are unlikely to show similar mapping success on all genetic elements when compared to true positive samples. Therefore, mapping to all three plasmids was used in combination with a statistical outlier detection for verification of low coverage genomes. Ratios of reads mapping to the *Y. pestis* chromosome and the three individual plasmids were determined, and samples were authenticated by calculating the Mahalanobis distance to detect outliers (χ^2=9.210, df=2, p=0.01; Table S2). Two samples, DIR002.A and PEI001.A, were classified as outliers: despite having chromosomal coverage, they had no or only a few reads mapping to the plasmids and were therefore considered as *Y. pestis* negative. The remaining 17 samples come from four sites in Germany (Dittenheim [DIT], n=3; Petting, n=3; Waging [WAG], n=1; Unterthürheim [UNT], n=5), one in Spain (Valencia [VAL], n=1) and one in France (Lunel-Viel [LVC], n=6) (Table 1, Fig. 1).

After mapping to the chromosome, seven genomes showed a higher than 5-fold mean coverage and were used for downstream analyses. These were DIT003.B (9.4-fold), VAL001.B (9.6-fold), PET004.A (5.6-fold) as well as UNT003.A and UNT004.A (7.6-fold and 5.2-fold respectively) (Table S3). Six positive samples of the individuals LVC001, LVC005 and LVC006 were merged to yield a mean coverage of 6.7-fold for the site of Lunel-Viel. For the phylogenetic analysis, we omitted the lower-covered UNT004.A after assuring that there are no conflicting positions with UNT003.A that derives from the same archaeological site.

For the 22 samples from Edix Hill, Britain, only shotgun sequencing data was available and, therefore, pathogen DNA screening was performed using the metagenomic tool MALT (19). This analysis revealed six putatively *Y. pestis*-positive samples after visual inspection of aligned reads in MEGAN (20) (Table S4). The sample EDI001.A had more than 9000 reads assigned

160   to *Y. pestis* and was sequenced to a greater depth without enrichment to yield a mean

161   chromosomal coverage of 9.1-fold.

162

163   <u>SNP Evaluation</u>

164   In the context of ancient pathogen DNA, there are three possible sources for false positive

165   SNPs: First, DNA damage such as deamination of cytosine to uracil can lead to

166   misincorporation of nucleotides during sample processing (21). Second, the mapping of closely

167   related environmental species to the reference sequence of the target organism is likely,

168   especially for conserved regions of the genome (22). Third, mapping of short reads is more

169   prone to mismapping and calling of false positive SNPs generated at sites of genome

170   rearrangement. Whereas the first source can be circumvented via *in vitro* protocols like UDG

171   treatment (23), the latter two can be reduced but not eliminated with strict mapping parameters

172   and exclusion of problematic regions (24) as applied here. A fourth source for false SNP

173   assignments could result from multiple genetically distinct strains that would lead to a chimeric

174   sequence. The later was not observed in our data (Fig. S1) and this phenomenon might be

175   limited to chronic infections with pathogens such as *Mycobacterium tuberculosis,* where mixed

176   infections have been previously documented (25).

177   The retrieval of genomes that span a wide geographic area gives us the opportunity to assess *Y.*

178   *pestis* microdiversity present in Europe during the First Pandemic. Given that our genomes are

179   of relatively low genomic coverage, we critically evaluated uniquely called and shared SNPs

180   among the First Pandemic genomes in order to accurately determine their phylogenetic position.

181   This analysis was performed for all genomes retrieved from UDG-treated libraries with higher

182   than 5-fold mean coverage, including the previously published high-quality Altenerding

183   genome (17.2-fold mean coverage).

184   For this, we developed the tool 'SNPEvaluation' and defined three different criteria, all

185   applying for a 50 bp window surrounding the SNP: (A) Comparing the mean coverage after

186   BWA mapping with high and low stringency and excluding all SNPs that showed a higher

187   coverage under low stringent mapping than in high stringent mapping. In metagenomic

188   datasets, reads of related species map frequently to conserved regions in the reference genome.

189   When the position is not covered by reads from the target organism (*Y. pestis*) but the genomic

190   region is similar enough in other environmental organisms so that their reads can map, they

191   might mimic a SNP in *Y. pestis* when the contaminant species carries a different allele in that

192   position. (B) Excluding all SNPs for which heterozygous calls were identified in the

193   surrounding regions. Heterozygous calls accumulate in conserved regions due to the above-

194   described effect. (C) Excluding all SNPs within regions that include positions that lack genomic

195   coverage. Variants in genome architecture often appear as gaps in mapped data and are likely

196   to cause mapping errors, potentially resulting in false positive SNPs.

197   This evaluation was applied to all SNPs identified as unique to the First Pandemic lineage,

198   totalling between one and 15 per genome, respectively (Table S5). 17 authentic derived

199   chromosomal SNPs and an additional one detected on the pMT1 plasmid were found across all

200   seven genomes (Table S6). The Altenerding genome (AE1175) as well as the genomes of

201   Unterthürheim (UNT003.A, UNT004.A) and Dittenheim (DIT003.B) appear identical after

202   SNP evaluation at all positions, with the exception of one SNP which is covered in both UNT

203   samples, and by only one read in DIT003.B, but not covered in Altenerding. The genomes from

204   Petting (PET004.A), Valencia (VAL001.B) and Lunel-Viel (LVC00_merged) appear distinct,

205   each occupying a unique branch comprised of two, three and ten unique SNPs, respectively

206   (Fig. 2B and Table S6). One additional SNP was found on the pMT1 plasmid in the Valencia

207   genome (VAL001.B). An analysis of the Aschheim genome as well as a SNP effect analysis is

208   presented in the SI.

209   Since the Altenerding genome shows the highest coverage (6), all SNPs previously presented

210   as unique SNPs for this genome were evaluated as potentially shared SNPs when they appeared

211   in at least one of the new genomes – excluding the position shared exclusively with DIT003.B

212   and UNT004.A. We applied the exact same parameters as for the unique SNPs, but also

213   considered positions with less than 3-fold coverage (Table S7). Only SNPs that pass all three

214   criteria of our SNP evaluation in at least half of the analysed genomes (i.e., four out of seven)

215   were accepted as true shared SNPs, reducing the number from 53 identified in a previous study

216   (6) to 43.

217   The Waging sample (WAG001.A) had a genomic coverage too low for inclusion in our

218   phylogenetic analysis. Since it was the only sample giving evidence for *Y. pestis* presence at

219   this site, it was assessed for all SNPs that were either shared or unique in the other First

220   Pandemic genomes. Visual inspection revealed seven of the 43 shared SNPs to be present in

221   the WAG001.A genome at low coverage (<3-fold), but none of the unique ones. For both shared

222   and unique SNPs, no conflicting positions were found. This strain could, therefore, be attributed

223   to the First Pandemic lineage without, however, resolving its exact position (Table S5).

224   The genome reconstructed from the non-UDG library of EDI001.A was not considered in the

225   presented SNP analysis, since damaged sites interfere with the defined SNP evaluation criteria.

226   With a mean coverage of 9.1, we accept the two relevant positions (one ancestral and one

227 derived SNP) as true positive after visual inspection. Regardless, metrics determined in the SNP

228 evaluation are still reported for comparison.

229

230 <u>Phylogenetic analysis</u>

231 A set of 233 modern *Y. pestis* genomes (Table S8) as well as seven Second Pandemic genomes,

232 including a representative of the Black Death strain (London) and six post-Black Death

233 genomes (16th-century Ellwangen (11); 18th-century Marseille (12)), and an ancient genome

234 from Tian Shan (DA101, 2nd to 3rd century (26)) were used for phylogenetic analyses alongside

235 our First Pandemic genomes presented here (Table S3) and the previously published genome

236 of Altenerding. The *Y. pseudotuberculosis* isolate IP32953 (27) was used as an outgroup.

237 Our maximum likelihood tree (28) constructed from the full SNP alignment reveals that all of

238 the genomes presented here occupy positions on the same lineage (Fig. 2A, Fig. S2). This

239 confirms their authenticity and is congruent with previous association of this lineage to the First

240 Pandemic (541–750). In addition, the previously reported genome from Altenerding (2148) is

241 identical to the new genomes from Dittenheim (DIT003.B) and Unterthürheim (UNT003.A).

242 Moreover, the genomes of Petting (PET004.A), Valencia (VAL001.B) and Lunel-Viel

243 (LVC_merged) seem to diverge from the Altenerding cluster through a polytomy (Fig. 2C,

244 bootstrap support 98 %). The British genome of EDI001, however, branches off one SNP

245 ancestral to this polytomy (bootstrap support 100 %) and possesses one unique SNP. This is

246 remarkable, since the British Isles are one of the most remote places where the First Pandemic

247 has been suspected to have reached in relation to the presumed starting point in Egypt.

248

249 <u>Virulence factor and deletion analysis</u>

250 We screened for the presence/absence of 80 chromosomal and 42 plasmid associated virulence

251 genes (29, 30) in all First Pandemic genomes with higher than 5-fold coverage (Fig. 3 and Fig.

252 S3). Only the filamentous prophage was consistently found absent in all presented genomes.

253 This is expected, since it has integrated into the genome of only a number of modern Branch 1

254 genomes (31). Reduced coverages for a set of virulence factors can be seen in the Altenerding

255 (AE1175) and Ellwangen genomes due to a capture bias, since the capture probe set in the

256 respective studies was designed on the basis of *Y. pseudotuberculosis* rather than of *Y. pestis*

257 (6, 11).

258 Intriguingly, the most derived First Pandemic genome from Lunel-Viel shows a deletion of two

259 chromosomal virulence factors, *mgtB* and *mgtC* (Fig. 3). These magnesium transporters are part

260 of the PhoPQ regulon, which is important for survival of *Y. pestis* in the magnesium-deficient

8

261   environment of macrophages. However, functional studies on *mgtB* hint at an important role

262   during macrophage invasion rather than intracellular survival (32).

263   A second deletion was observed for the gene YPO2258, categorized as a potential virulence

264   factor based on the presence of a frame shift mutation in the avirulent 0.PE2_Microtus91001

265   strain (30). Its inactivation in the 2.ANT1_Nepal516 strain, isolated from a human patient,

266   nevertheless indicates that this gene is not essential for virulence in humans (33).

267   Further exploration of the deletion of the two neighbouring genes *mgtB* and *mgtC* revealed that

268   they are part of a ca. 45 kb deletion (positions 1,883,402 to 1,928,869 in the CO92 reference),

269   affecting 34 genes including multiple motility (*motA*, *motB*) and chemotaxis genes (*cheA, cheB,*

270   *cheD, cheR, cheW, cheY, cheZ*) (Fig. S4). On the downstream end, the deletion is flanked by an

271   IS100 insertion element. A potential upstream insertion element might be undetectable at our

272   current resolution due to a genome rearrangement in the reference genome CO92. This is in

273   agreement with previous findings concerning the highly abundant IS100 element in *Y. pestis*,

274   responsible not only for disruptions of multiple genes caused by homologous recombination

275   (27), but also for the loss of the 102 kb long *pgm* locus containing a high-pathogenicity island

276   in several strains (34). To address the specificity of this deletion to the 6th–7th century strain

277   from France, we also investigated the presence of the two virulence factors in all other modern

278   and ancient strains in this study. Intriguingly, a similar deletion affecting the same region

279   including *mgtB* and *mgtC* was observed in the late Second Pandemic genomes from London

280   New Churchyard, (1560-1635 (35)) and Marseille L'Observance (1720-1722 (12)). However,

281   a full deletion of this 45 kb region was not found in any of the other ancient or modern genomes.

282   Therefore, the deletion appeared independently in the course of both the First and Second

283   Pandemics.

284

285   **Discussion**

286   Identifying *Y. pestis* DNA in low complexity specimens

287   In total, we screened 145 samples from 19 sites in France, Germany and Spain for *Y. pestis* with

288   a qPCR assay (18) and 22 additional samples from Edix Hill, Britain, with the metagenomic

289   tool MALT (19). While the most promising sample of Edix Hill was directly sequenced to

290   greater depth to reach a chromosomal mean coverage of 9.1-fold, all qPCR positive samples of

291   the other sites were turned into UDG libraries and subsequently enriched for *Y. pestis*, resulting

292   in mean coverages ranging from 0.01 to 9.6-fold.

293 The validation of ancient genomes with relatively low coverage as presented here is challenging

294 since the DNA extracted from archaeological remains results in metagenomic data and the

295 differentiation between target organism DNA and environmental background can be difficult.

296 The identification of *Y. pestis* DNA based on PCR targeting the *pla* locus on the pPCP1 plasmid

297 has theoretically been shown to be problematic (36), leading to discussions about false positive

298 results (16). However, assignment to *Y. pestis* based on reads retrieved from shotgun

299 sequencing and mapping to a reference genome also can be challenging in case of extremely

300 low genomic coverage (3, 4). Since all the presented genomes, except the one from Edix Hill,

301 are derived from DNA libraries specifically enriched for *Y. pestis* DNA and are thus biased

302 towards the target organism, a previously suggested competitive mapping approach (3) would

303 not be suitable. Instead, we considered the relative number of mapping reads to the plasmids

304 and chromosome to identify false positive samples from captured data. We were able to verify

305 that 17 out of 19 samples were positive for *Y. pestis* with as few as 4000 reads mapping to the

306 chromosome. Since the three plasmids pCD1, pMT1 and pPCP1 were already present in the

307 early divergent Neolithic and Bronze Age strains (3, 4) and loss of plasmids has only been

308 observed sporadically in attenuated strains (37), this method could be reliably applied to data

309 stemming from other branches in the *Y. pestis* phylogeny.

310

311 <u>Analysing microdiversity with low coverage genomes</u>

312 Reliable SNP calling is crucial for the phylogenetic analysis of verified low coverage genomes

313 and can be challenging when dealing with ancient pathogen DNA stemming from metagenomic

314 contexts. This has been demonstrated on *Y. pestis* genomes (6), but previously applied visual

315 inspections are time-consuming and not easily reproducible.

316 Here, we present a novel approach for SNP authentication using a semi-automated SNP

317 evaluation. We selected three criteria for our evaluation to assess the likelihood of mismapping.

318 We excluded all SNPs that (A) had higher coverage when mapped with less strict parameters,

319 (B) had 'heterozygous' positions in close proximity or (C) were flanked by gaps. With these

320 filters, we tolerate a loss of sensitivity to increase specificity, which is critical for detection and

321 characterization of microdiversity. Moreover, the tool 'SNPEvaluation' that was newly

322 developed for this analysis offers a highly flexible framework for the assessment of VCF files

323 and can be utilized also for a variety of analyses on different organisms.

324

325

326

10

327   <u>Phylogenetic analysis</u>

328   We were able to confidently reconstruct six new genomes from the First Pandemic in Britain,

329   France, Germany and Spain, providing insights into the microdiversity of *Y. pestis* in Europe

330   between the 6th and 7th centuries.

331   Our presented genomes add diversity to a phylogenetic lineage that was previously shown to

332   contain two identical 6th century genomes of southern Germany (Aschheim and Altenerding (6,

333   7)). It diverges between the 0.ANT1, 0.ANT2 and 0.ANT5 clades in the main *Y. pestis*

334   phylogeny and shares a short branch with a 2nd- to 3rd-century genome from the Tian Shan

335   mountains (26). Intriguingly, a single diversification event gave rise to the published as well as

336   three of the presented additional genomes, each defined by 1 to 10 derived SNPs. Similar

337   polytomies can be detected in other parts of the phylogeny of *Y. pestis* that have been related to

338   human epidemics (38): one gave rise to Branches 1 to 4 (including ancient Second Pandemic

339   genomes, Fig. 2A) and is dated to 1142-1339 (38), shortly before the European Black Death.

340   To date it is unknown if this event was restricted to a rodent reservoir, or if it was already

341   associated with a human epidemic. A second polytomy gave rise to the 1.ORI clade, which

342   includes strains related to the worldwide spread of plague during the Third Pandemic in the 19th

343   century (Fig. 2C).

344   Within the First Pandemic lineage, the genomes that derive from this polytomy display variable

345   terminal branch lengths (1-10 SNPs), which are likely concurrent with their different ages (see

346   below). Given that *Y. pestis* is a pathogen that can cover large geographic distances by

347   accumulating little to no genetic diversity (11), it is challenging to elucidate the geographic

348   origin for this diversification event. A first hypothesis suggests an origin of this diversification

349   event within the historically recorded geographic range of the First Pandemic, i.e. either in

350   Europe, the Mediterranean basin, or the Middle East. Our current data may lend some

351   credibility to this scenario for two reasons: First, we identify four European strains with short

352   genetic distances from this polytomy, the shortest of which is identified in three locations in

353   rural Bavaria, and second, we identify an almost direct ancestor of this polytomy to be present

354   in Europe during the 6th century, represented by a genome from Britain. Alternatively, the

355   bacterium may have been recurrently introduced to the affected regions from a single remote

356   reservoir.

357   The hypothesis of a single introduction would require the establishment of a local reservoir,

358   since we have to assume that at least the genome recovered from Lunel-Viel is not directly

359   associated with the initial outbreak in 541–544 but rather with a subsequent one (see below).

360   Possible locations for reservoirs during the First Pandemic have been suggested in the Iberian

11

361   peninsula and the Levant (39). There is also a growing body of evidence for the presence of

362   black rats (*Rattus rattus)* in Europe in late Antiquity and the Early Medieval Period (40, 41),

363   suspected to represent the main reservoir species during the Second Pandemic (40).

364   Such a scenario would be congruent with the Second Pandemic, where the phylogeny of ancient

365   genomes is in line with a single introduction and subsequent persistence in a local host (11, 35,

366   42), although this hypothesis was challenged by an alternative scenario claiming multiple

367   introductions on the basis of climatic data (43). Similar to the European Second Pandemic

368   lineage (11, 12), strains emerging from the First Pandemic lineage have so far been recovered

369   solely from ancient DNA of European plague burials, suggesting that the lineage either went

370   extinct or persists in a yet unsampled reservoir.

371

372   Origin of the Justinianic Plague

373   Based on available data, it has been suggested that the most parsimonious location for the

374   divergence event that gave rise to the First Pandemic lineage is Central Asia (26). All published

375   genomes of the branches 0.ANT1, 0.ANT2 and 0.ANT5 that frame the First Pandemic lineage

376   in the phylogenetic tree were sampled in the autonomous Xingjiang region in north-western

377   China or in Kyrgyzstan (38, 44). In addition, an ancient 2nd- to 3rd-century *Y. pestis* genome

378   from the Tian Shan mountains in Central Asia (26) branches off basal to all the First Pandemic

379   genomes. The resulting claim that the Huns might have brought plague to Europe is however

380   unsubstantiated due to the gap of more than three centuries prior to the onset of the First

381   Pandemic.

382   Since the long shared branch of the First Pandemic genomes (43 SNPs) does not have any

383   known extant descendants, this strain might have been maintained in a now extinct reservoir

384   after its emergence in Central Asia. The first outbreak is reported in Pelusium, Egypt; an

385   introduction from either Africa or Asia was presumed, given the sudden and dramatic onset of

386   the pandemic. Previous assumptions of an African origin were mainly based on a single deeply

387   diverging 0.PE strain 'Angola' (45) and the reports of the Byzantine historian Evagrius

388   Scholasticus, who wrote in his *Ecclesiastical History* that the plague began in "Ethiopia".

389   However, there are legitimate doubts about the characterization of the 'Angola' genome as a

390   genuine African strain (24, 46) and the account of Evagrius has been assessed critically with

391   historical and philological methods (47, 48). For an Asian origin, the sea route via the Red Sea

392   and the Indian Ocean is a plausible scenario since India was well connected by marine traffic

393   with the early Byzantine Empire (39). A suggested alternative scenario would require overland

394   transport from the Eurasian Steppe via Iran to the Red Sea that is, so far, not supported by any

12

395   data (49). In conclusion, we interpret the current data as insufficient to resolve the origin of the

396   Justinianic Plague as a human epidemic.

397

398   <u>Archaeological and historical context</u>

399   Here, we present the first genomic evidence for the First Pandemic reaching the British Isles in

400   the 6th century. This genome was recovered from a burial on the site of Edix Hill, close to

401   Cambridge (Roman *Duroliponte*) and near a Roman road running north from London

402   (*Londinium*) toward Lincoln (*Lindum Colonia*) via Braughing, all of which were Roman

403   settlements. Based on archaeological dating in combination with its rather basal position within

404   the clade, this genome is likely related to the very first occurrence of plague in Britain suggested

405   for 544 (see SI). The close proximity to the trade center of ancient London supports that the

406   strain was introduced by sea communications, e.g., with Brittany, following the outbreak in

407   central Gaul in 543 (Fig. 1, (50)). Interestingly, the genome was recovered from a single burial,

408   underlining that in small settlements, plague-induced mortality crises need not always involve

409   a radical change in mortuary practice towards multiple or mass burials. The fact that three of

410   the five additional Edix Hill individuals that appeared positive for plague in the MALT

411   screening were buried in two simultaneous double burials nevertheless suggests that multiple

412   burials in normal cemeteries are indeed a good indicator for epidemic events (17).

413   In addition, we were able to reconstruct two genomes from the Mediterranean basin, where the

414   historical records are more explicit about the presence of plague during the First Pandemic.

415   Regarding Spain, the radiocarbon dating of the *Y. pestis*-positive individual from Valencia

416   (432-610) would include the first outbreak reported for Spain in 543 in a contemporary

417   chronicle (see SI). The three unique SNPs identified in this genome, which separate it from the

418   identified polytomy may suggest its association with a later outbreak. Intriguingly, a canon of

419   a church council held in 546 in Valencia dealing with burial practices for bishops in case of

420   sudden death was recently connected with plague by philological and contextual analysis (51).

421   Later outbreaks within the relevant timeframe are documented in Spain's Visigothic kingdom,

422   e.g., in 584 and 588 by Gregory of Tours, and by a funerary inscription dated 609 at Cortijo de

423   Chinales 35 km northeast of Malaga (31).

424   The second Mediterranean genome from Lunel-Viel in southern France likely represents

425   another outbreak, since it forms an independent strain which derives from the same polytomy

426   as the Spanish and German genomes. The radiocarbon dates for the inhumations give an interval

427   of at least 567-618 (youngest lower and oldest upper boundary, Table S9) overlapping with

428   documented outbreaks in 571, 582, 588, 590 and possibly 599–600 in southern France (see

13

429   Fig.1A and C). Lunel-Viel's broader vicinity includes Arles, the seaport city of Marseille and
430   the Rhône mouth. Close to important coastal and fluvial shipping routes as well as Roman roads
431   that facilitated the spread of plague (39), Lunel-Viel could have been affected by all five
432   recorded epidemics. The initial outbreak, documented for Arles ca. 543, falls outside of some
433   of the radiocarbon intervals. This is consistent with the phylogenetic analysis that shows a
434   higher accumulation of SNPs in this genome. Thus the victims at Lunel-Viel can most likely
435   be attributed to one of the subsequent outbreaks.

436   Moreover, within Bavaria, Germany we detected *Y. pestis* in four sites (Dittenheim, Petting,
437   Unterthürheim, Waging) in addition to the two previously published sites (Altenerding (6),
438   Aschheim (7)). Two of the reconstructed genomes were identical to Altenerding and Aschheim,
439   suggesting that these four can be attributed to the same epidemic event. Some of the radiocarbon
440   intervals of these sites fall even slightly before the onset of the First Pandemic, suggesting an
441   association of this outbreak directly with the Justinianic Plague. Regarding the Edix Hill
442   genome, this would in turn necessitate the accumulation of one (Edix Hill) to two (Altenerding
443   cluster) SNPs within the onset of the First Pandemic between 541–544.

444   Intriguingly, the genome of Petting, Bavaria falls not with the Altenerding cluster but in a
445   distinct phylogenetic position. Since this strain also branches off from the common node with
446   the other Bavarian strain as well as the French and Spanish genomes, this shows the presence
447   of two independent strains and, therefore, presumably two independent epidemic events in early
448   medieval Bavaria. This is striking, since we lack any historical records of the First Pandemic
449   affecting southern Germany. The radiocarbon dates for the Bavarian sites are inconclusive and
450   do not allow for a clear temporal separation of the two events. The higher number of
451   accumulated SNPs nevertheless suggests a younger date for the epidemic represented by
452   Petting. Further phylogeographic analyses are presented in the SI.

453

454   Deletion analysis
455   The analysis of virulence factors revealed a deletion of a ca. 45 kb region in the most derived
456   and putatively most recent genome thus far identified for the First Pandemic. This deletion
457   contained two previously described virulence factors involved in host cell invasion and
458   intracellular growth (*mgtB* and *mgtC*). Intriguingly, a similar deletion covering the same
459   genomic region was detected in the most derived available Second Pandemic genomes from
460   London New Churchyard (1560-1635) and Marseille (1720-1722). Genome decay by deletion
461   or pseudogenization is a well-known trait of *Y. pestis* and has contributed to its distinct ecology
462   and pathogenicity (53). Both deletions from the First and Second Pandemic are observed in

14

463 genomes recovered from human victims. Therefore, it is reasonable to assume that the deletion

464 may not have reduced the bacterium's virulence. Moreover, it affects a number of cell surface

465 proteins – remnants of the motile lifestyle of non-pestis *Yersiniae* (54) – so the deletion might

466 have even facilitated immune evasion.

467 Because none of the investigated modern strains harboured this specific deletion, this possible

468 case of convergent evolution might be an adaptation to a distinct ecological niche in Europe or

469 the Mediterranean basin since an ancient local reservoir is the most parsimonious hypothesis

470 for both historical pandemics (12, 42).

471

472 Concluding remarks

473 Our study succeeds in offering new insights into the first historically documented plague

474 pandemic, complementing the limited power of conventional historical, archaeological or

475 palaeoepidemiological research. Moreover, we show the potential of palaeogenomic research

476 for understanding historical and modern pandemics by a comparative approach on genomic

477 features throughout millennia. Facing the problem of low coverage genomic data with a high

478 environmental background – a notorious challenge in ancient DNA research –, we have

479 developed new approaches to facilitate the authentication and confident phylogenetic

480 placement of such genomes.

481 In the future, more extensive sampling of putative plague burials will help to draw a more

482 comprehensive picture of the onset and persistence of the First Pandemic, especially on sites in

483 the eastern Mediterranean basin, where not only is the Justinianic Plague reported to have

484 started, but where also the 8th century outbreaks also clustered. This will contribute to the

485 comparative exploration of *Y. pestis'* microevolution and human impact in the course of past

486 and present pandemics.

487

488 **Material and Methods**

489 Sites and Samples

490 The acquisition and selection of samples followed two approaches: Focussing on Bavaria, we

491 concentrated on one region, where the two previously reconstructed *Y. pestis* genomes

492 attributed to the Justinianic Plague had been found (6, 7). Additionally, given the absence of

493 robust genetic evidence from the Mediterranean basin, which historical records depict as the

494 epicenter of the pandemic, and the controversial presence of plague on the British Isles during

495 the Justinianic Plague, we extended our screening to three sites with multiple burials in a

496   broader geographical scope on the Mediterranean coast in France and Spain and inland Britain.

497   Table 1 gives an overview of all tested sites.

498   For the first focus, we collected samples of 79 individuals from 46 burials belonging to 16

499   archaeological sites in Bavaria, Germany and one site in Austria (See Fig. 1B). The dating of

500   the burials spans the 4th to 10th century, including also burials dating before (8 individuals on 3

501   sites) and after (17 individuals on 5 sites) the Justinianic Plague (541–544). Since mass graves

502   that could be indicative of an epidemic are unsurprisingly rare for the small settlements

503   associated with early medieval cemeteries in Bavaria, we followed the approach of the previous

504   successful studies (6, 7, 17): we systematically screened multiple burials, i.e., where two or

505   more individuals were found in a context indicating a simultaneous burial, such as a common

506   grave pit and articulated remains on the same level. Single burials were sporadically tested, if

507   the context suggested a close connection to a multiple burial. Burials with indications of a

508   violent death of the interred were excluded, since a coincidental acute infection with *Y. pestis*

509   seems unlikely.

510   Within the Mediterranean basin, we tested inhumations from Valencia, Spain and Lunel-Viel

511   (Hérault), France. A contemporary chronicler records that bubonic infection devastated Spain

512   during the first phase of the Justinianic Plague (541–544), and new interpretation of a

513   contemporary record argues that it reached Valencia presumably before 546 (51). Further

514   textual references, including an epitaph dating to 609, document later Iberian outbreaks (52)

515   (See Fig. 1). In the Visigothic levels of the *Plaça de l'Almoina* in Valencia, several collective

516   burials in an intramural cemetery were interpreted as possible plague burials (52, 55).

517   The historical evidence for the First Pandemic in France is more substantial, mainly based on

518   the contemporary bishop and historian Gregory of Tours (56). He reports several plague

519   outbreaks spanning from ca. 543 in the province of Arles through 588 in Marseille to 590 in

520   Avignon (See Fig. 1C). The site of Lunel-Viel, around 30 km southwest of the ancient Roman

521   city of Nîmes and less than 100 km from the mentioned cities, revealed eight exceptional

522   inhumations in demolition trenches unrelated to the nearby contemporary cemeteries (57).

523   For the British Isles, the historical evidence for plague presence in the 6th century is

524   controversial. Unlike later outbreaks in 7th-century Britain that are reported, e.g., by Bede, the

525   identification of a disease occurring in the 540s and called *blefed* in Irish chronicles as bubonic

526   plague is mainly based on coincidence with the Continental European outbreaks and thus

527   uncertain. The same is true for Britain, where a great mortality (*mortalitas magna*) is reported

528   in the *Annales Cambriae* (see SI). For this study, we screened 22 individuals from the Anglo-

16

529   Saxon cemetery of Edix Hill, well-connected to the Roman road network and Roman towns,

530   and characterized by a number of multiple burials.

531   For the screening, one tooth (preferentially molar) per individual was used for every individual

532   of a multiple burial, if available. For a number of individuals, additional teeth were tested, if

533   sequencing the first gave a weak positive. For the collective burials from Valencia, a clear

534   attribution to individuals was not assured, so multiple teeth were sampled per feature number,

535   if possible. Detailed site descriptions can be found in the SI, including a table with all screened

536   samples (Table S1).

537

538   <u>Sample Preparation, DNA Extraction, qPCR and MALT Screening</u>

539   The sample preparation and DNA extraction for samples from Austria, France, Germany and

540   Spain was done in the ancient DNA facilities of the ArchaeoBioCenter of the University of

541   Munich, Germany, and the Max Planck Institute for the Science of Human History in Jena,

542   Germany.

543   All teeth were cut along the cementoenamel junction and the surface of the pulp chamber was

544   drilled out with a dental drill from the crown and in some cases the root, aiming for 30 to 50 mg

545   of bone powder. DNA was extracted based on the protocol published in (58): The powder was

546   suspended in 1 ml of extraction buffer (0.45 M EDTA pH 8.0, and 0.25 mg/ml Proteinase K in

547   UV-irradiated HPLC water) and incubated at 37 °C overnight on a rotor. After centrifugation,

548   the supernatant was mixed with 10 ml binding buffer (5 M guanidinium hydrochlorid, 40 %

549   isopropanol and 90 mM sodium acetate) to bind the DNA on a silica column of either the

550   MinElute purification kit (Qiagen) or the High Pure Viral Nucleic Acid Kit (Roche). After

551   purification with washing buffer of the respective kit, the DNA was eluted in 100 µl TET buffer

552   (10 mM Tris-HCl, 1 mM EDTA pH 8.0, 0.05 % Tween20).

553   All extracts were tested with the qPCR assay targeting a 52 bp region on the pPCP1 plasmid

554   published in (18) with minor changes (0.75 mg/ml BSA, additional 5 % DMSO, EVA green

555   instead of SYBR green, annealing for 30 s, elongation for 30 s, gradient from 60 to 90 °C). All

556   samples showing an amplification with a melting peak between 74 and 80 °C were captured for

557   *Y. pestis*.

558   The samples of Edix Hill, UK were prepared in the ancient DNA facility of the University of

559   Cambridge, Department of Archaeology. Root portions of teeth were removed with a sterile

560   drill wheel. These root portions were briefly brushed with 5 % w/v NaOCl using a UV-

561   irradiated toothbrush that was soaked in 5 % w/v NaOCl for at least 1 min between samples.

562   Roots were then soaked in 6 % w/v bleach for 5 min. Samples were rinsed twice with ddH$_2$O

17

563 and soaked in 70 % Ethanol for 2 min, transferred to a clean paper towel on a rack inside the

564 glove box, UV irradiated for 50 min on each side, and then allowed to dry. They were weighed

565 and transferred to clean, UV-irradiated 5 ml or 15 ml tubes for chemical extraction. Per 100 mg

566 of each sample, 2 ml of EDTA Buffer (0.5 M pH 8.0) and 50 µl of Proteinase K (10 mg/ml)

567 were added. Tubes were rocked in an incubator for 72 h at room temperature. Extracts were

568 concentrated to 250 µl using Amplicon Ultra-15 concentrators with a 30 kDa filter. Samples

569 were purified according to manufacturer's instructions using the Minelute^{TM} PCR Purification

570 Kit with the only change that samples were incubated with 50 µl Elution Buffer at 37 °C for

571 10 min prior to elution.

572

573 <u>Library Preparation</u>

574 Of putatively positive extracts in the qPCR screening, 50 µl were turned into Illumina double-

575 stranded DNA libraries with initial USER treatment (New England Biolabs) to remove post-

576 mortem damage in form of deaminated Cytosines by consecutive incubation with uracil-DNA-

577 glycosylase (UDG) and endonuclease VIII(23). To enhance the efficiency of subsequent double

578 indexing, UDG-treated libraries were quantified by qPCR using IS7/IS8 primer and split for a

579 maximum of 2×10^{10} DNA molecules. Every library was indexed with a unique index

580 combination in a 10-cycle amplification reaction using *Pfu Turbo Cx Hotstart DNA Polymerase*

581 (Agilent) (59, 60). The amplification products were purified using the MinElute DNA

582 purification kit (Qiagen) and eluted in TET (10 mM Tris-HCl, 1 mM EDTA pH 8.0, 0.05 %

583 Tween20). For the capture, the indexed libraries were amplified to 200-300 ng/µl using

584 *Herculase II Fusion DNA Polymerase* (Agilent) and purified a second time as described.

585 The non-UDG library preparation for all Edix Hill samples was conducted using a protocol

586 modified from the manufacturer's instructions included in the NEBNext® Library Preparation

587 Kit for 454 (E6070S, New England Biolabs, Ipswich, MA) as detailed in (61). DNA was not

588 fragmented and reactions were scaled to half volume, adaptors were made as described in (59)

589 and used in a final concentration of 2.5 µM each. DNA was purified on MinElute columns

590 (Qiagen, Germany). Libraries were amplified using the following PCR set up: 50 µl DNA

591 library, 1x PCR buffer, 2.5 mM MgCl$_2$, 1 mg/ml BSA, 0.2 µM in PE 1.0, 0.2 mM dNTP each,

592 0.1 U/µl HGS Taq Diamond, and 0.2 µM indexing primer. Cycling conditions were: 5' at

593 94 °C, followed by 18 cycles of 30 seconds each at 94 °C, 60 °C, and 68 °C, with a final

594 extension of 7 min at 72 °C. Amplified products were purified using MinElute columns and

595 eluted in 35 µl EB. Samples were quantified using Quant-iT™ PicoGreen® dsDNA kit (P7589,

596 Invitrogen™ Life Technologies) on the Synergy™ HT Multi-Mode Microplate Reader with

597 Gen5™ software.

598

599 <u>In-Solution Capture</u>

600 For the in-solution capture, a probe set was generated using a fragment size of 52 bp and a tiling

601 of 1 bp with the following genomes as templates: CO92 chromosome (NC_003143.1), CO92

602 plasmid pMT1 (NC_003134.1), CO92 plasmid pCD1 (NC_003131.1), KIM 10 chromosome

603 (NC_004088.1), Pestoides F chromosome (NC_009381.1) and *Y. pseudotuberculosis* IP 32953

604 chromosome (NC_006155.1). The capture was performed as previously described (62) on 96-

605 well plates with a maximum of two samples pooled per well and all blanks with unique index

606 combinations in one well.

607

608 <u>Sequencing and Data Processing</u>

609 All captured products were sequenced either on a Illumina NextSeq500 or HiSeq4000 platform

610 at the Max Planck Institute for the Science of Human History in Jena, Germany. Edix Hill

611 libraries were sequenced on Illumina NextSeq500 at the University of Cambridge Biochemistry

612 DNA Sequencing Facility and the FastQ files were processed on the Estonian Biocenter server.

613 De-multiplexed reads were processed with the EAGER pipeline (63) starting Illumina adapter

614 removal, sequencing quality filtering (minimum base quality of 20) and length filtering

615 (minimum length of 30 bp). Sequencing data of paired end and single end sequencing were

616 concatenated after adapter removal and merging. The same was done for samples from the same

617 individual (DIT004) and all data from Lunel-Viel (LVC) due to low genomic coverage. The

618 sequencing results are shown in Table S3.

619 All Edix Hill samples were screened using MALT (19) against a reference set including full

620 bacterial and viral genomes with 85 % identity, the strong positive sample EDI001.A was then

621 sequenced deeper to a coverage of 9.1-fold without enrichment.

622 After clipping of 3 bases on each end with fastx_trimmer of the FASTX toolkit

623 (https://github.com/agordon/fastx_toolkit) to remove the majority of damaged sites for the non-

624 UDG library of EDI001.A, the sample was processed along the UDG treated libraries. Mapping

625 against reference genomes of CO92 (chromosome NC_003143.1, plasmid pMT1

626 NC_003134.1, plasmid pCD1 NC_003131.1, plasmid pPCP1 NC_003132.1) was done with

627 BWA using stringent parameters (-n 0.1, -l 32). Reads with low mapping quality were removed

628 with Samtools (-q 37) and duplicates were removed with MarkDuplicates. For the plasmids, a

629 merged reference was used, consisting of the CO92 reference of pCD1 (NC_003131.1), pMT1

630   (NC_003134.1) and pPCP1 (NC_003132.1, with base pairs 3,000 to 4,200 masked (18)), to

631   avoid overestimation of coverage due to homologous regions. For the verification of positive

632   qPCR results, we normalized the number of reads mapping to each plasmid with reads mapping

633   to the chromosome and calculated the Mahalanobis distance for each sample to detect outliers.

634   Based on this, we excluded the samples PEI001.A and DIR002.A as false positives (Table S2).

635   The raw data of the Aschheim and Altenerding genomes were processed identically, however

636   considering only the A120 sample for Aschheim (6, 7).

637

638   <u>SNP Calling and Evaluation</u>

639   All genomes recovered from UDG-libraries with higher than 5-fold mean coverage including

640   the Altenerding genome were assessed in the SNP analysis. Additionally, the sample

641   WAG001.A was evaluated to explore its phylogenetic position, since it was the only positive

642   sample of the relevant site.

643   The UnifiedGenotyper within the Genome Analysis Toolkit was used for SNP calling and

644   creating VCF files for all genomes, using 'EMIT_ALL_SITES' to generate calls for all

645   positions in the reference genome. For the subsequent analyses, 233 previously published

646   modern *Y. pestis* genomes (Table S8), one genome from 2nd to 3rd century Tian-Shan mountains

647   (DA101 (26)) one genome representing the Black Death from London East Smithfield (8291-

648   11972-8124 (12)), and six Second Pandemic genomes (Ellwangen, Marseille L'Observance

649   OBS107, OBS110, OBS116, OBS124, OBS137 (11, 12)) were taken along together with *Y.*

650   *pseudotuberculosis* (IP32953) as an outgroup. Previously identified problematic regions (24,

651   38) as well as regions annotated as repeat regions, rRNAs, tRNAs and tmRNAs were excluded

652   for all following analyses. MultiVCFAnalyzer v0.85 (64) was used for generating a SNP table

653   with the following settings: Minimal coverage for base call of 3 with a minimum genotyping

654   quality of 30 for homozygous positions, minimum support of 90 % for calling the dominant

655   nucleotide in a 'heterozygous' position. All positions failing these criteria would be called 'N'

656   in the SNP table. For the SNP evaluation, all 'N' positions of unique SNPs within the First

657   Pandemic lineage were re-evaluated, replacing 'N' by '0' for not covered and lower case letters

658   for homozygous positions with max. 2-fold coverage. To test for possible mixed infections of

659   elevated contamination, all SNPs not passing the 90 % threshold were plotted (Fig. S1).

660   For the evaluation of unique and shared SNPs of First Pandemic genomes retrieved from non-

661   UDG   libraries,   we   used   the   newly   developed   tool   'SNPEvaluation'

662   (https://github.com/andreasKroepelin/SNP_Evaluation) and a comparative mapping, using

663   BWA with high stringent (-n 0.1, -l 32) and low stringent (-n 0.01, -l 32) mapping parameters,

20

664   allowing for more mismatches in the latter. SNPs were called true positive when meeting the

665   following criteria within a 50 bp window: (A) the ratio of mean coverage of low stringent to

666   high stringent mapping is not higher than 1, (B) no 'heterozygous' positions, (C) no non-

667   covered positions (Table S5). SNP evaluation on the plasmids was done using the same criteria

668   after mapping to the individual references as described above. For the SNP effect analysis, the

669   remaining unique true SNPs were compared to the genome annotations of the CO92 *Y. pestis*

670   reference genome (Table S6).

671   Shared SNPs (Table S7) were evaluated with the same criteria with minor modifications: The

672   minimum threshold for calling a position was set to 1 read covering and SNPs were called true

673   positive, if the SNP passed the criteria in more than half of the genomes under examination.

674   The genome of EDI001.A from Edix Hill was only included in the presented SNP evaluation

675   for comparison, since it is derived from a non-UDG library and damaged sites interfere with

676   both the comparative mapping and the count of heterozygous positions.

677   The Aschheim genome was evaluated separately (Table S8) but with the same criteria. As

678   previously addressed (6), the enormously high number of false positive SNPs might not be

679   explained solely by contamination by soil bacteria or sequencing errors but additionally by PCR

680   or capture artefacts.

681

682   <u>Phylogenetic Analyses</u>

683   For the phylogenetic analyses we aimed for one high coverage genome per site to minimize

684   missing data in the SNP alignment, excluding the genome of UNT004.A after assuring no

685   conflicting positions with UNT003.A in the SNP evaluation. A maximum likelihood tree

686   (RAxML 8 (28) using the GTR substitution model, Fig. 2A, for full tree see Fig. S2) was

687   generated without exclusion of missing and ambiguous data (full SNP alignment), resulting in

688   a total number of 6496 SNPs. Robustness of all trees was tested by the bootstrap methods using

689   1000 pseudo-replicates.

690   A detailed tree of the First Pandemic lineage was drawn manually based on the performed SNP

691   evaluation, excluding all false positive SNPs (Fig. 2B).

692

693   <u>Analysis of virulence factors and genome decay</u>

694   The presence/absence analysis for genes was performed with BEDTools (65) by calculating the

695   percentage across each gene (3). Since gene duplications can affect the mapping quality, the

696   mapping quality filter of BWA was set to 0 (-q = 0) to generate a bam-file as input. For the

697   heatmap of virulence factors (Fig. 3), a collection of proven and putative virulence genes (29,

21

698   30) was evaluated. The more extensive analysis on genome decay was based on the annotation

699   file for the reference genome CO92 (54) by extracting all regions annotated as 'gene'.

700   For the exact determination of the start and end positions of deletions, mapping with

701   BWA_MEM was performed (66).

702

703   <u>Radiocarbon Dating</u>

704   At least one individual per burial was sampled for radiocarbon dating for all burials that tested

705   positive for *Y. pestis*, assuming simultaneity of interment for the multiple burials. Samples were

706   dated at the CEZ Archaeometry gGmbH, Mannheim, Germany. The raw radiocarbon dates

707   were calibrated with IntCal13 (67) in OxCal v4.3.2 (68). All raw and calibrated dates are given

708   in Table S9; Fig. S5 shows the respective probability distributions. Some of the intervals

709   completely pre-date the onset of Justinianic Plague (541) which could be explained by a marine

710   or freshwater reservoir effect (69, 70) or human bone collagen offset (71). In the absence of

711   C/N isotope data and a well-established method for addressing the human bone collagen offset,

712   we report calibrated dates without any correction.

713

714   <u>Cartography</u>

715   All maps were generated in ArcGIS 10.4.1 (ESRI) using the 'World Ocean Basemap' without

716   references. The sources for all historical occurrences are given in the SI. The mapped regions

717   in Fig. 1 are primarily based on the Digital Atlas of Roman and Medieval Civilizations

718   (DARMC; https://darmc.harvard.edu) maps "Provinces AD303-324" for the western Europe

719   and "Provinces ca. AD500" for eastern Europe, Middle East and Africa. The provinces in Fig.

720   S6 are based on a georeferenced map by Rettner and Steidl (72), the Roman roads are combined

721   from Rettner and Steidl and the DARMC map "Roman Roads". The main rivers in Figs. 1C

722   and 3 were taken from Natural Earth (ne_10m_river_lake_centerlines,

723   http://www.naturalearthdata.com), based on data provided by the European Commission, Joint

724   Research Centre, Institute for Environment and Sustainability (JRC IES).

725

726   **Acknowledgements**

736

Data availability

738   The raw sequencing data of the relevant plague positive samples will be available on the

739   European Nucleotide Archive under project accession number PRJEB29991 upon publication.

References

1. Yang R, Anisimov A (2016) *Yersinia pestis: Retrospective and Perspective* (Springer, Dordrecht).

2. Burki T (2017) Plague in Madagascar. *Lancet Infect Dis* 17(12):1241.

3. Andrades Valtueña A, et al. (2017) The Stone Age Plague and Its Persistence in Eurasia. *Curr Biol* 27:1–9.

4. Rasmussen S, et al. (2015) Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago. *Cell* 163:571–582.

5. Spyrou MA, et al. (2018) Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague. *Nat Commun* 9(2234):1–10.

6. Feldman M, et al. (2016) A High-Coverage Yersinia pestis Genome from a Sixth-Century Justinianic Plague Victim. *Mol Biol Evol* 33(11):2911–2923.

7. Wagner DM, et al. (2014) Yersinia pestis and the Plague of Justinian 541-543 AD: A genomic analysis. *Lancet Infect Dis* 14(4):319–326.

8. Benedictow OJ (2004) *The Black Death, 1346-1353 : The Complete History* (Boydell & Brewer, Woodsbridge).

9. Bos KI, et al. (2011) A draft genome of Yersinia pestis from victims of the Black Death. *Nature* 478(7370):506–10.

10. Biraben J-N (1975) *Les hommes et la peste en France et dans les pays européens et méditerranéens* (Mouton, Paris).

11. Spyrou MA, et al. (2016) Historical Y. pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics. *Cell Host Microbe* 19(6):874–881.

12. Bos KI, et al. (2016) Eighteenth century Yersinia pestis genomes reveal the long-term persistence of an historical plague focus. *Elife* 5:e12994.

13. Wiechmann I, Grupe G (2005) Detection of Yersinia pestis DNA in two early medieval skeletal finds from Aschheim (Upper Bavaria, 6th century A.D.). *Am J Phys Anthropol* 126(1):48–55.

14. Drancourt M, et al. (2004) Genotyping, orientalis-like Yersinia pestis, and plague pandemics. *Emerg Infect Dis* 10(9):1585–1592.

15. Drancourt M, et al. (2007) Yersinia pestis Orientalis in remains of ancient plague patients. *Emerg Infect Dis* 13(2):332–333.

16. Gilbert MTP, et al. (2004) Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims. *Microbiology* 150:341–

774   354.

775   17.   Harbeck M, et al. (2013) Yersinia pestis DNA from Skeletal Remains from the 6th
776   Century AD Reveals Insights into Justinianic Plague. *PLoS Pathog* 9(5):e1003349.

777   18.   Schuenemann VJ, et al. (2011) Targeted enrichment of ancient pathogens yielding the
778   pPCP1 plasmid of Yersinia pestis from victims of the Black Death. *Proc Natl Acad Sci*
779   *USA* 108(38):746–752.

780   19.   Vågene ÅJ, et al. (2018) Salmonella enterica genomes from victims of a major
781   sixteenth-century epidemic in Mexico. *Nat Ecol Evol* 2(3):520–528.

782   20.   Huson DH, et al. (2016) MEGAN Community Edition - Interactive Exploration and
783   Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput Biol*
784   12(6):e1004957.

785   21.   Briggs AW, et al. (2007) Patterns of damage in genomic DNA sequences from a
786   Neandertal. *Proc Natl Acad Sci U S A* 104(37):14616–14621.

787   22.   Warinner C, et al. (2017) A Robust Framework for Microbial Archaeology. *Annu Rev*
788   *Genomics Hum Genet* 18(1):13.1-13.36.

789   23.   Briggs AW, et al. (2010) Removal of deaminated cytosines and detection of in vivo
790   methylation in ancient DNA. *Nucleic Acids Res* 38(6):e87–e87.

791   24.   Morelli G, et al. (2010) Yersinia pestis genome sequencing identifies patterns of global
792   phylogenetic diversity. *Nat Genet* 42(12):1140–1143.

793   25.   Kay GL, et al. (2015) Eighteenth-century genomes show that mixed infections were
794   common at time of peak tuberculosis in Europe. *Nat Commun* 6(6717):6717.

795   26.   Damgaard P de B, et al. (2018) 137 ancient human genomes from across the Eurasian
796   steppes. *Nature*.

797   27.   Chain PSG, et al. (2004) Insights into the evolution of Yersinia pestis through whole-
798   genome comparison with Yersinia pseudotuberculosis. *Proc Natl Acad Sci*
799   101(38):13826–13831.

800   28.   Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-
801   analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

802   29.   Zhou D, Yang R (2009) Molecular Darwinian Evolution of Virulence in Yersinia
803   pestis. *Infect an* 77(6):2242–2250.

804   30.   Zhou D, et al. (2004) Genetics of Metabolic Variations between Yersinia pestis Biovars
805   and the Proposal of a New Biovar, microtus. *J Bacteriol* 186(15):5147–5152.

806   31.   Derbise A, Carniel E (2014) YpfΦ: a filamentous phage acquired by Yersinia pestis.
807   *Front Microbiol* 5(December):701.

808   32.   Ford DC, Joshua GWP, Wren BW, Oyston PCF (2014) The importance of the
809   magnesium transporter MgtB for virulence of Yersinia pseudotuberculosis and
810   Yersinia pestis. *Microbiology* 160:2710–2717.

811   33.   Chain PSG, et al. (2006) Complete genome sequence of Yersinia pestis strains antiqua
812   and Nepal516: Evidence of gene reduction in an emerging pathogen. *J Bacteriol*
813   188(12):4453–4463.

814   34.   Fetherston JD, Perry RD (1994) The pigmentation locus of Yersinia pestis KIM6+ is
815   flanked by an insertion sequence and inciudes the structural genes for pesticin
816   sensitivity and HMWP2. *Mol Microbiol* 13(4):697–708.

817   35.   Spyrou MA, et al. (2018) A phylogeography of the second plague pandemic revealed
818   through the analysis of historical Y. pestis genomes. *bioRxiv*.

819   36.   Hänsch S, et al. (2015) The pla gene, encoding plasminogen activator, is not specific to
820   Yersinia pestis. *BMC Res Notes* 8(535):1–3.

821   37.   Garcia E, et al. (2007) Pestoides F, an atypical Yersinia pestis strain from the former
822   Soviet Union. *Adv Exp Med Biol* 603:17–22.

823   38.   Cui Y, et al. (2013) Historical variations in mutation rate in an epidemic pathogen,
824   Yersinia pestis. *Proc Natl Acad Sci USA* 110(2):577–82.

825   39.   Harper K (2017) *The Fate of Rome: Climate, Disease, and the End of an Empire*
826   (Princeton University Press, Princeton).

827   40.   McCormick M (2003) Rats, Communications, and Plague: Toward an Ecological
828   History. *J Interdiscip Hist* 34(1):1–25.

829   41.   Audoin-Rouzeau F, Vigne J-D (1997) Le rat noir (Rattus rattus) en Europe antique et
830   mediévale: Les voies du commerce et l'expansion de la peste. *Anthropozoologica*
831   25(26):399–404.

832   42.   Seifert L, et al. (2016) Genotyping Yersinia pestis in Historical Plague: Evidence for
833   Long-Term Persistence of Y. pestis in Europe from the 14th to the 17th Century. *PLoS*
834   *One* 11(1):e0145194.

835   43.   Schmid B V, et al. (2015) Climate-driven introduction of the Black Death and
836   successive plague reintroductions into Europe. *Proc Natl Acad Sci USA* 112(10):3020–
837   5.

838   44.   Eroshenko GA, et al. (2017) Yersinia pestis strains of ancient phylogenetic branch
839   0.ANT are widely spread in the high- mountain plague foci of Kyrgyzstan. *PLoS One*
840   12(10):e0187230.

841   45.   Eppinger M, et al. (2010) Genome sequence of the deep-rooted Yersinia pestis strain

842   angola reveals new insights into the evolution and pangenome of the plague bacterium.

843   *J Bacteriol* 192(6):1685–1699.

844   46.   Green MH (2014) Taking "Pandemic" Seriously: Making the Black Death Global.

845   *Pandemic Disease in the Medieval World: Rethinking the Black Death*, ed Green MH

846   (Kalamazoo, Bradford).

847   47.   Sarris P (2002) The Justinianic plague: origins and effects. *Contin Chang* 17(2):169–

848   182.

849   48.   Allen P (1979) The "Justinianic" Plague. *Byzantion* 49:5–20.

850   49.   Schamiloglu U (2016) The Plague in the Time of Justinian and Central Eurasian

851   History: An agenda for Research. *Central Eurasia in the Middle Ages. Studies in*

852   *Honour of Peter B. Golden*, eds Zimonyi I, Karatay O (Harrasowitz Verlag,

853   Wiesbaden).

854   50.   Little LK, et al. (2007) *Plague and the End of Antiquity: The Pandemic of 541-750* ed

855   Little LK (Cambridge University Press).

856   51.   Gruber H (2018) Indirect Evidence for the Social Impact of the Justinianic Pandemic:

857   Episcopal Burial and Conciliar Legislation in Visigothic Hispania. *J Late Antiq*

858   11(1):193–215.

859   52.   Kulikowski M (2007) Plague in Spanish Late Antiquity. *Plague and the End of*

860   *Antiquity: The Pandemic of 541-750*, ed Little LK (Cambridge University Press,

861   Cambridge), pp 150–170.

862   53.   McNally A, Thomson NR, Reuter S, Wren BW (2016) "Add, stir and reduce": Yersinia

863   spp. as model bacteria for pathogen evolution. *Nat Rev Microbiol* 14(3):177–190.

864   54.   Parkhill J, et al. (2001) Genome sequence of Yersinia pestis, the causative agent of

865   plague. *Nature* 413(6855):523–527.

866   55.   Alapont Martín L, Ribera i Lacomba AV (2009) Topografía y jerarquía funeraria en la

867   Valencia tardo-antigua. *Morir en el Mediterráneo Medieval: Actas del III Congreso*

868   *Internacional de Arqueología, Arte e Historia de la Antigüedad Tardía y Alta Edad*

869   *Media Peninsular* (Oxford), pp 59–88.

870   56.   Stoclet AJ (2007) Consilia humana, ops divina, superstitio: Seeking Succor and Solace

871   in Times of Plague, with Particular Reference to Gaul in the Early Middle Ages.

872   *Plague and the End of Antiquity: The Pandemic of 541-750*, ed Little LK (Cambridge

873   University Press, Cambridge), pp 135–149.

874   57.   Raynaud C (2010) *Les nécropoles de Lunel-Viel (Hérault) de l'Antiquité au Moyen*

875   *Âge. Revue archéologique de Narbonnaise Supplément 40* (Presses universitaires de la

876   Méditerranée, Montpellier).

877 58. Dabney J, et al. (2013) Complete mitochondrial genome sequence of a Middle
878   Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc Natl Acad*
879   *Sci USA* 110(39):15758–15763.

880 59. Meyer M, Kircher M (2010) Illumina sequencing library preparation for highly
881   multiplexed target capture and sequencing. *Cold Spring Harb Protoc* 2010(6).

882 60. Kircher M, Sawyer S, Meyer M (2012) Double indexing overcomes inaccuracies in
883   multiplex sequencing on the Illumina platform. *Nuc* 40(1):1–8.

884 61. Rasmussen M, et al. (2014) The genome of a Late Pleistocene human froma Clovis
885   burial site inwesternMontana. *Nature* 506:225–229.

886 62. Fu Q, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave,
887   China. *Proc Natl Acad Sci USA* 110(6):2223–2227.

888 63. Peltzer A, et al. (2016) EAGER: efficient ancient genome reconstruction. *Genome Biol*
889   17:60.

890 64. Bos KI, et al. (2014) Pre-Columbian mycobacterial genomes reveal seals as a source of
891   New World human tuberculosis. *Nature* 514(7523):494–497.

892 65. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing
893   genomic features. *Bioinformatics* 26(6):841–842.

894 66. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with
895   BWA-MEM. *arXiv*:1303.3997.

896 67. Reimer PJ, et al. (2013) IntCal13 and Marine13 Radiocarbon Age Calibration Curves
897   0–50,000 Years cal BP. *Radiocarbon* 55(4):1869–1887.

898 68. Ramsey CB (2017) Methods for Summarizing Radiocarbon Datasets. *Radiocarbon*
899   59(6):1809–1833.

900 69. Ascough PL, et al. (2010) Temporal and spatial variations in freshwater 14C reservoir
901   effects: Lake Mývatn, northern Iceland. *Radiocarbon* 52(2–3):1098–1112.

902 70. Ascough P, Cook G, Dugmore A (2005) Methodological approaches to determining the
903   marine radiocarbon reservoir effect. *Prog Phys Geogr* 4:532–547.

904 71. Barta P, Štolc S (2007) HBCO Correction: Its Impact on Archaeological Absolute
905   Dating. *Radiocarbon* 49(2 "Proceedings of the 19th international Radiocarbon
906   Conference, Keble College, Oxford, England"):465–472.

907 72. Wamser L (2010) *Karfunkelstein und Seide. Neue Schätze aus Bayerns Frühzeit*
908   (Pustet, Regensburg).

909

Fig. 1: Geographic extent of the First Pandemic and sampled sites. A: Map of historically documented occurrences of plague (regions shaded, cities depicted by circles, both with respective years of occurrence) between 541 and 650 in Europe and the Mediterranean basin. All sources are given in the SI. Sites with genomic evidence for *Y. pestis* are shown as pink (previously published) and yellow squares (presented here). B: Enlarged rectangular space of A (right) showing all sites in Germany and Austria that were included in this study. Sites tested negative are depicted in black upward-pointing triangles (burials dating before 541), squares (dating around 541–544) downward-pointing triangles (dating after 544). C: Enlarged inset of A (left) shows reported occurrences in France and main rivers.
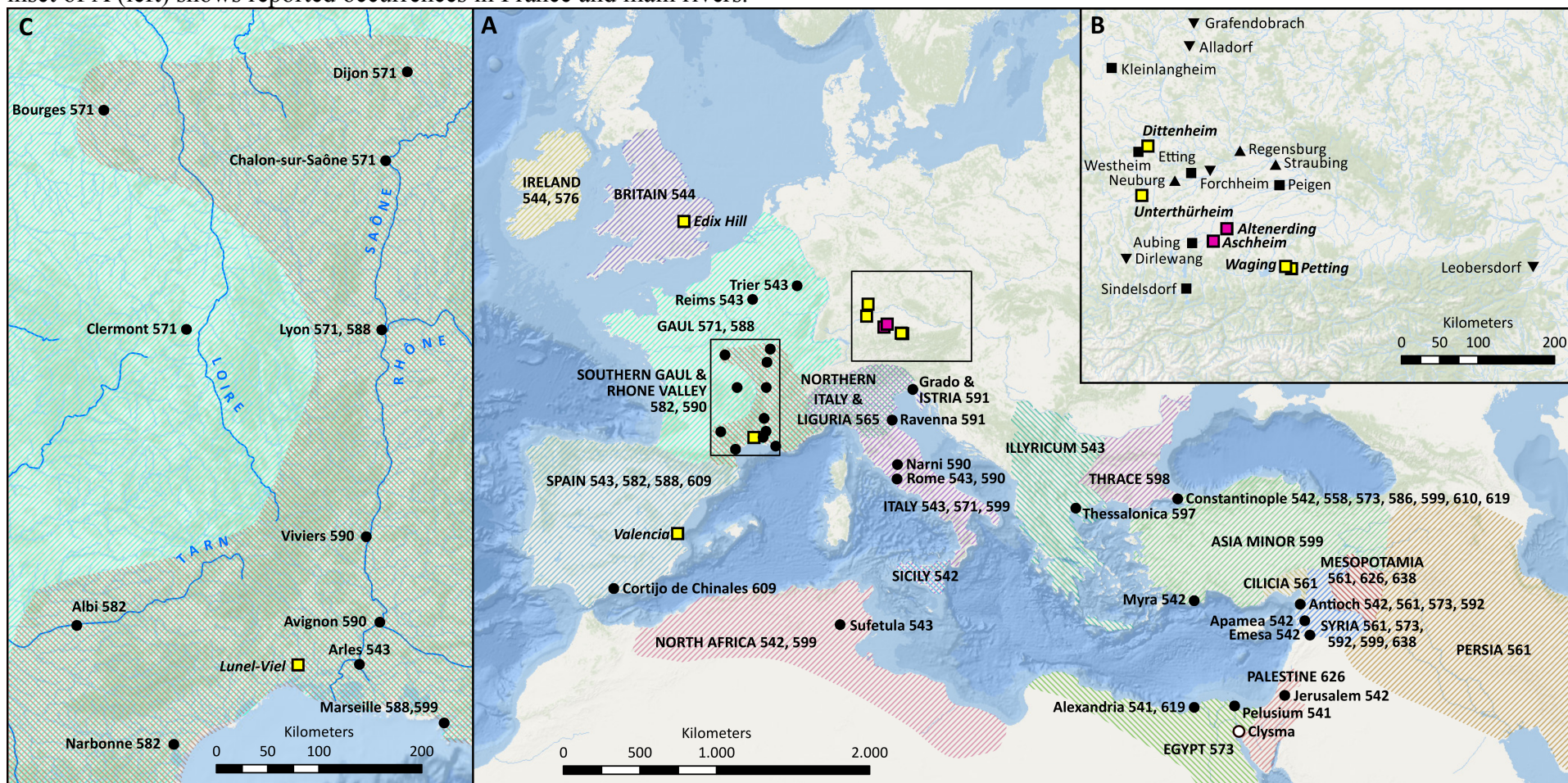
Fig. 2: Phylogenetic tree. **A:** Maximum Likelihood tree with full SNP alignment (6496 positions) of 233 modern *Y. pestis* and one *Y. pseudotuberculosis* genome, nine published (2nd to 3rd century Tian Shan in orange; Altenerding in blue; Second Pandemic in red) and six genomes presented here (green) with country given in brackets (DEU=Germany, ESP=Spain, FRA=France, GBR=Great Britain). Numbers and origins of modern genomes are given in brackets (CHN=China, COG=Congo, FSU=Former Soviet Union, IND=India, IRN=Iran, MDG=Madagascar, MMR=Myanmar, MNG=Mongolia, NPL=Nepal, UGA=Uganda). Numbers on nodes are showing bootstrap values (1000 iterations). B: Detailed, manually drawn tree of the First Pandemic genomes showing all remaining SNP positions after SNP evaluation (number of SNPs given in italics). C: Detailed tree of the 1.ORI clade within Branch 1, showing the polytomy.
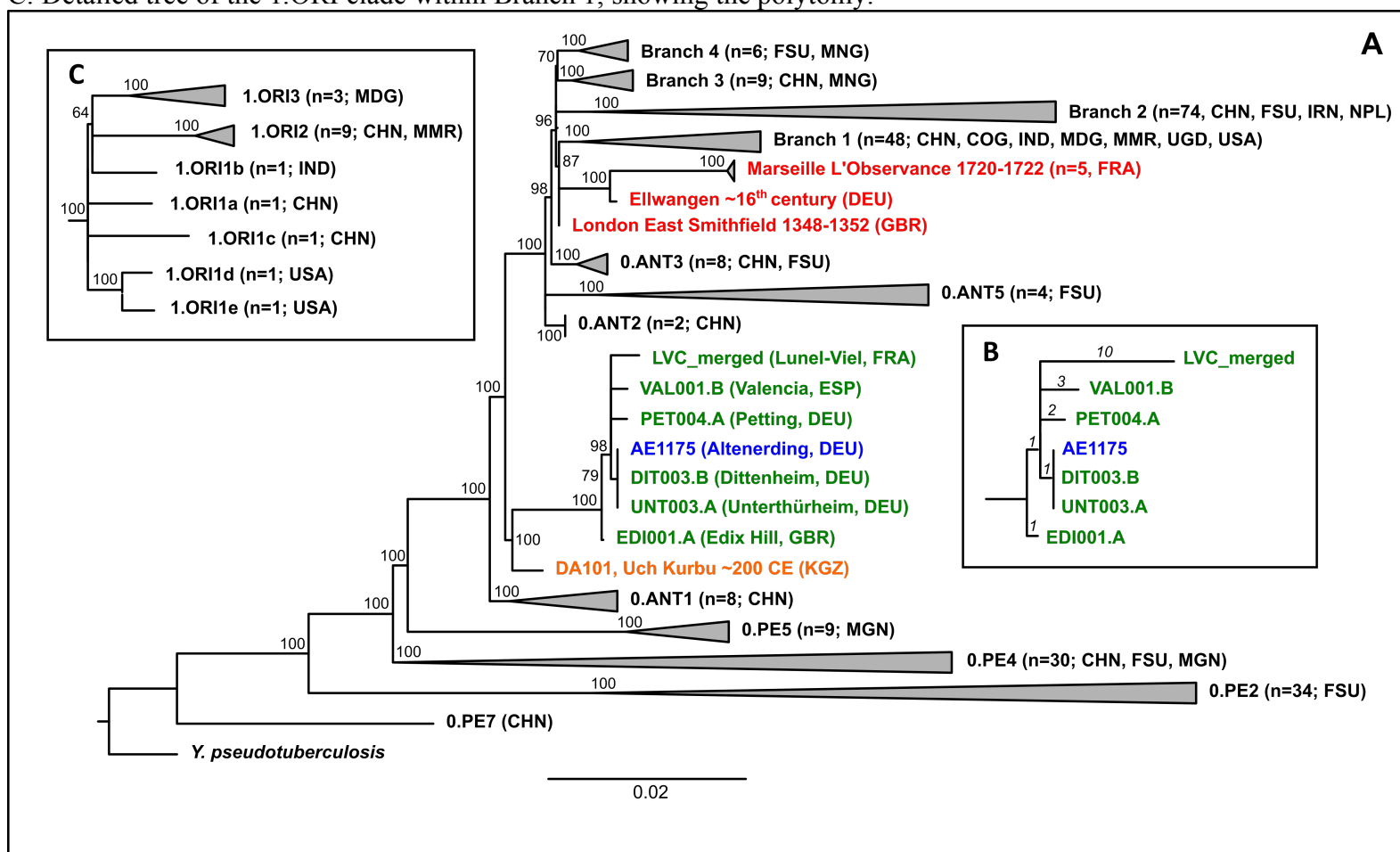
Fig. 3: Heatmap showing the percentage of coverage of chromosomal virulence factors. First Pandemic genomes (blue and green) and Second Pandemic genomes (red) are shown in combination with selected strains of main clades of modern *Y. pestis* diversity on Branch 0 as well as the reference genomes of *Y. pseudotuberculosis* and *Y. pestis* (CO92).
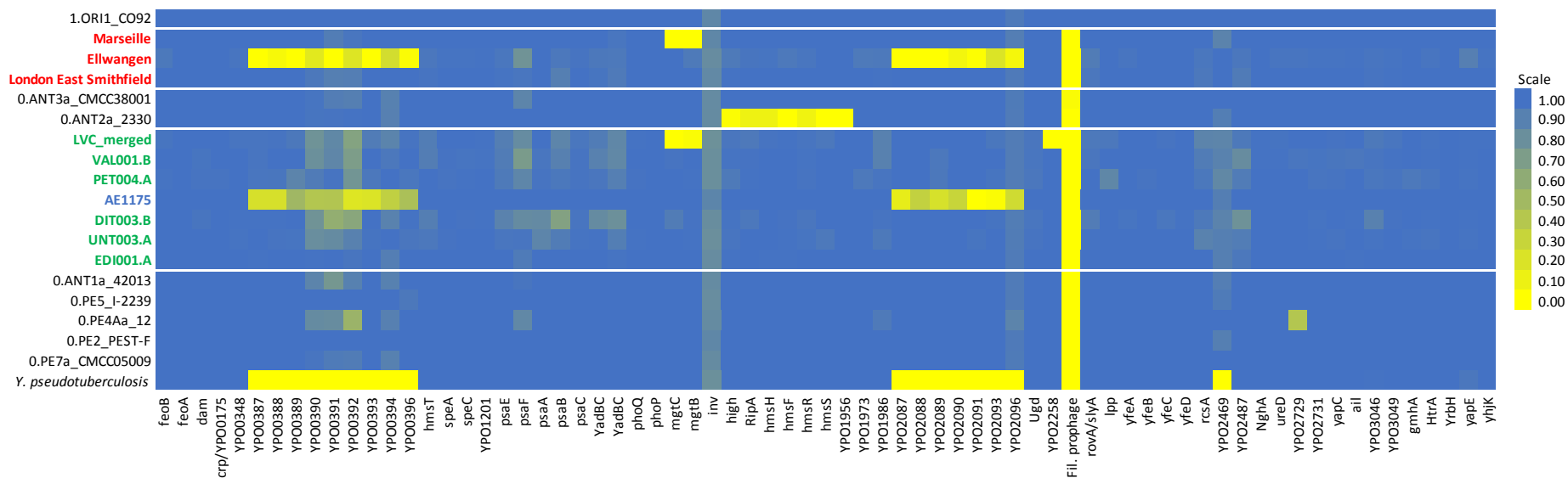
929 **Table 1:** List of all sites that were tested with country in brackets (AUS=Austria, DEU=Germany, ESP=Spain, FRA=France). The number of graves
930 is counting multiple burials as single graves; cremations are counted separately. Multiple burials are listed as number of graves times number of
931 individuals (5x2 translates to 5 double burials, 1x2+1 to one double burial associated with a single burial). Detailed site descriptions are given in the
932 SI, a table of all screened samples in Table S1.
933

Site	Lab ID	Context	Graves in total	Multiple burials	Timeframe	Positive/total samples
Alladorf (DEU)	ALL	Separate burial area (*Hofgrablege*)	163	5x2	630-720	0/6
Dirlewang (DEU)	DIR	Early medieval cemetery	40	2x2	650-700	0/2
Dittenheim (DEU)	DIT	Early medieval cemetery	238, 10 crem.	4x2	550-700	3/9
Edix Hill	EDI	Early medieval cemetery	115	1x4, 9x2	500-650	1/22
Forchheim (DEU)	FOR	Special burial	1	1x4	650-700	0/3
Grafendobrach (DEU)	GRA	Settlement burials (*Hofgrablege*)	85	1x3, 1x2+1	850-930	0/3
Kleinlangheim (DEU)	KLH	Early medieval cemetery	244, 56 crem.	8x2, 1x3	470-720	0/5
Leobersdorf (AUS)	LEO	Early medieval cemetery	154	16x2, 4x3, 2x4, 1x5	640-800	0/3
Lunel-Viel Horts (FRA)	LVH	Early medieval cemetery	140	1x2	475-700	0/5
Lunel-Viel Quartier centrale (FRA)	LVC	Demolition trench inhumations	-	6+2 individuals in 2 trenches	400-600	6/16
München-Aubing (DEU)	AUB	Early medieval cemetery	896	4x2	400-700	0/8
Neuburg an der Donau (DEU)	NEU	Late Roman cemetery	130	3x2, 1x3	300-400	0/2
Peigen (DEU)	PEI	Early medieval cemetery	274	3x2	450-700	0/5
Petting (DEU)	PET	Early medieval cemetery	721	min. 1x3, 2x2, 1x2+1	530-730	3/7
Regensburg Fritz-Fend-Str. (DEU)	RFF	Late Roman cemetery	115, 48 crem.	2x2	350-450	0/3
Sindelsdorf (DEU)	SIN	Early medieval cemetery	331	3x2, 1x3+1	500-720	0/5
Straubing Azlburg I/II (DEU)	SAZ	Late Roman cemetery	541, 1 crem.	2x2, 1x3	300-450	0/3
Unterthürheim (DEU)	UNT	Early medieval cemetery	256	14x2, 2x3, 1x4	525-680	5/7
Valencia, Plaça de Almoina (ESP)	VAL	Visigothic intramural cemetery	67	3x2, 3x3, 4x4, 2x5, 15x5+	500-700	1/36
Waging (DEU)	WAG	Early medieval cemetery	239	min. 2x2, 1x2+1	530-700	1/12
Westheim (DEU)	WES	Early medieval cemetery	228	5x2, 1x3	500-650	0/3

934