Machine learning models to identify patient and microbial genetic factors associated with carbapenem-resistant *Klebsiella pneumoniae* infection

Zena Lapp BA[1], Jennifer Han MD[2], Jenna Wiens PhD[3], Ellie JC Goldstein MD[4,5], Ebbing Lautenbach MD[6], Evan Snitkin PhD[7]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan; 1510A MSRB I, 1150 W. Medical Center Dr., Ann Arbor, MI, 48109-5680

[2]GlaxoSmithKline; 14200 Shady Grove Road, Rockville, MD 20850

[3]Department of Electrical Engineering and Computer Science, University of Michigan; 3765 Beyster, 2260 Hayward Street, Ann Arbor, MI 48109

[4]R M Alden Research Laboratory

[5]David Geffen School of Medicine, University of California, Los Angeles; 2021 Santa Monica Blvd. #640-E, Santa Monica, CA 90404-2208

[6]Department of Medicine (Infectious Diseases), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania; 502A Johnson Pavilion, 3610 Hamilton Walk, Philadelphia, PA 19104-6073

[7]Department of Microbiology and Immunology, Department of Internal Medicine/Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan; 1520D MSRB I, 1150 W. Medical Center Dr., Ann Arbor, MI, 48109-5680

Corresponding author: Evan Snitkin

Address: 1520D MSRB I, 1150 W. Medical Center Dr., Ann Arbor, MI, 48109-5680

Email: esnitkin@med.umich.edu

Telephone: (734) 647-6472

Abstract

Background

Among patients colonized with carbapenem-resistant *Klebsiella pneumoniae* (CRKP), only a subset develop clinical infection. While patient characteristics may influence risk for infection, it remains unclear if the genetic background of CRKP strains contributes to this risk. We applied machine learning to quantify the capacity of patient characteristics and microbial genotypes to discriminate infection and colonization, and identified patient and microbial features associated with infection across multiple healthcare facilities.

Methods

Machine learning models were built using whole-genome sequences and clinical metadata from 331 patients colonized or infected with CRKP across 21 long-term acute care hospitals. To quantify variation in performance, we built models using 100 different train/test splits of the entire dataset, and urinary and respiratory site-specific subsets, and evaluated predictive performance on each test split using the area under the receiver operating characteristics curve (AUROC). Patient and microbial features predictive of infection were identified as those consistently important for predicting infection based on average change in AUROC when included in the model.

Findings

We found that patient and genomic features were only weakly predictive of clinical CRKP infection vs. colonization (AUROC IQRs: patient=0·59-0·68, genomic=0·55-0·61, combined=0·62-0·68), and that one feature set did not consistently outperform the other (genomic vs. patient p=0·4). Comparable model performances were observed for anatomic site-specific models (combined AUROC IQRs: respiratory=0·61-0·71, urinary=0·54-0·64). Strong genomic predictors of infection included the presence of the ICEKp10 mobile genetic element carrying an iron acquisition system (yersiniabactin) and a toxin (colibactin), along with disruption of an O-antigen biosynthetic gene in a sub-lineage of the epidemic ST258 clone. Teasing apart sequential evolutionary steps in the context of clinical metadata indicated that

altered O-antigen biosynthesis increased association with the respiratory tract, and subsequent acquisition of ICEKp10 was associated with increased virulence.

Interpretation

Our results support the need for rigorous machine learning frameworks to gain realistic estimates of the performance of clinical models of infection. Moreover, integrating microbial genomic and clinical data using such a framework can help tease apart the contribution of microbial genetic variation to clinical outcomes.

Funding

Centers for Disease Control and Prevention, National Institutes of Health, National Science Foundation

Research in context

Evidence before this study

We searched PubMed for "crkp" OR "carbapenem resistant klebsiella pneumoniae" AND "infection" AND "machine learning" for papers published up to April 14, 2020 and found no results. Substituting “machine learning” with "bacterial genome-wide association studies" produced one relevant paper investigating pathogenicity-associated loci in *K. pneumoniae* clinical isolates. When we searched for "infection" AND "machine learning" AND "genom*" AND "clinical", there was one relevant result - a study that used clinical and bacterial genomic features in a machine learning model to identify clonal differences related to *Staphylococcus aureus* infection outcome.

Added value of this study

To our knowledge, this is the first study to integrate clinical and genomic data to study anatomic site-specific colonization and infection across multiple healthcare facilities. Using this method, we identified clinical features associated with CRKP infection, as well as a sub-lineage of CRKP with potentially altered niche-specific adaptation and virulence. This method could be used for other organisms and other

clinical outcomes to evaluate performance of predictive models and identify features that are consistently associated with clinical outcomes of interest across facilities or geographic regions.

Implications of all the available evidence

Few studies have combined patient and microbial genomic data to study important clinical outcomes. However, those that have done this, including ours, have identified clinical and/or genomic features associated with the outcome of interest that provide a foundation for future epidemiological, clinical, and biological studies to better understand bacterial infections and clinical outcomes.

Introduction

Infections due to multidrug resistant organisms (MDROs) lead to hundreds of thousands of deaths worldwide each year.[1] Carbapenem-resistant Enterobacterales (CRE) is a critical-priority antibiotic resistance threat that has emerged over the past several decades, spread across the globe, and accumulated resistance to last-line antibiotic agents.[2,3] In the United States (US), CRE infections are primarily caused by the sequence type (ST) 258 strain of carbapenem resistant *Klebsiella pneumoniae* (CRKP), which has become endemic in regional healthcare networks.[3–7] In this background of regional endemicity the risk of patient exposure to CRKP is high, as evidenced by alarmingly high rates of colonization, especially in long-term care settings.[7,8] However, even among critically ill patients residing in long-term care facilities, not all colonized patients develop clinical infections that require antibiotic treatment.[9] Currently, our understanding of the factors that influence whether a colonized patient develops an infection is incomplete.

In addition to clinical characteristics of patients, the genetic background of the colonizing strain may also influence the risk of infection, as there is extensive intra-species variation in antibiotic resistance and virulence determinants harbored by *K. pneumoniae*.[3] To date, most studies of virulence determinants have been carried out in model systems,[10] or examined in human populations without considering patient characteristics or clinical context.[11] One recent study investigated virulence determinants in *K.*

100 *pneumoniae* clinical isolates while controlling for patient characteristics.[12] However, this was a single-site
101 study with a focus on carbapenem-susceptible *K. pneumoniae*, thereby not addressing the impact of
102 genomic variation in antibiotic-resistant lineages that circulate in global healthcare systems.
103 Here, we sought to understand the importance of both patient factors and CRKP genetic background in
104 determining whether a patient is infected (vs. colonized) with CRKP, and identify a set of patient and
105 microbial features that are consistent predictors of CRKP infection across long-term care facilities. To
106 accomplish this, we compared patients with CRKP colonization and infection based upon both clinical
107 characteristics and the genomes of their colonizing or infecting strains. To improve the generalizability of
108 our findings, we employed a rigorous machine learning framework and included patients from 21 long-
109 term acute care hospitals (LTACHs) across the US.
110

111 ## Methods

112 ### Clinical and genomic data

113 We used whole-genome sequences of clinical (non-surveillance) CRKP isolates and associated patient
114 metadata from a prospective observational study performed in 21 LTACHs from across the US over the
115 course of a year (BioProject accession no. PRJNA415194).[13] We included only the first clinical
116 bloodstream, respiratory, or urinary isolate from each patient (n=355; **Figure S1A**), and subset to only
117 ST258 isolates for the majority of analyses (n=331; **Table S1**; see supplementary material for reasoning).
118 Details about the analysis pipeline,[14] genomic data curation,[13,15–21] and phylogenetic tree reconstruction[22–
119 25] are provided in the supplementary material.

120 ### Outcome definition

121 Our outcome of interest was colonization vs. clinical infection (**Figure S1B**). Based on established
122 Centers for Disease Prevention and Control's National Healthcare Safety Network (CDC's NHSN)
123 surveillance definitions, we considered all bloodstream isolates as representative of infection, and used
124 modified definitions to classify urinary and respiratory cultures as representative of infection versus

colonization (**Table S2**).[7,26] Any isolate that did not meet the criteria for infection was classified as colonization.

Feature sets

We studied the association between five different feature sets and infection/colonization in CRKP ST258 (**Figure S1C**), described below. See supplementary methods for details on feature set creation and processing. Counts below are for confident features from the entire dataset prior to processing.

Patient: Clinical features described in Han *et al.*[13] (n=50; **Table S3**).

Uncurated genomic: single nucleotide variants, indels, insertion sequence elements, and pangenome genes (n=2447).

Uncurated grouped genomic: Gene-level variant presence/absence and pangenome genes (n=3159).

Curated genomic: Features identified by Kleborate,[15] a tool designed to identify the presence of various genes and mutations known to be associated with either CRKP virulence or antibiotic resistance (n=91).

Patient & curated genomic: Patient features and curated genomic features (n=141).

Machine learning & model selection

We aimed to classify clinical infection (vs. colonization) using each of the different feature sets (see above); we built classifiers using the first clinical isolate from each patient for all isolates, only respiratory isolates, and only urinary isolates. We performed L2 regularized logistic regression using a modified version of the machine learning pipeline presented in Topçuoğlu *et al.*[27] using caret version 6.0-85[28] in R version 3.6.2[29] (**Figure S1D1**). We randomly split the data into 100 unique ~80/20 train/test splits, keeping all isolates from each LTACH grouped in either the training set or the held-out test set to control for facility-level differences among the isolates (e.g., background of circulating strains within each facility, patient population, and clinician test ordering frequency). For valid comparison, the train/test splits were identical across models generated with different feature sets. Hyperparameters were selected via cross-validation on the training set to maximize the average AUROC across cross-validation folds. See supplementary methods for more details.

151 **Model performance**

152 We measured model performance using the median test area under the receiver operating characteristic
153 curve (AUROC) and area under the precision recall curve (AUPRC), as well as the interquartile range,
154 across all 100 train/test splits (**Figure S1D2**).

155 **Features consistently associated with colonization or infection**

156 To determine the importance of each feature in predicting colonization vs. infection, we measured how
157 much each feature influenced model performance by calculating a permutation importance (**Figure**
158 **S1D3**). For each combination of feature and data split, we randomly permuted the feature and calculated
159 the 'permuted test AUROC' using the model generated with the training data. Features with a correlation
160 of 1 were permuted together. We performed this permutation test 100 times for each feature/data split
161 pair, and obtained a mean permutation importance for each data split. A mean permutation importance
162 above zero indicates that that feature improved model performance for that data split. We highlight
163 features where the mean test AUROC was above zero in at least 75% of the data splits. In this way, the
164 permutation importance method allows us to take into account the variation we observe across the 100
165 models, which is not possible with standard parametric statistical tests or odds ratios.

166 **Data analysis & visualization**

167 See supplementary material for details on data analysis and visualization in R version 3.6.2.[29–35] All code
168 and data that is not protected health information is on GitHub (https://github.com/Snitkin-Lab-Umich/ml-
169 crkp-infection-manuscript).

170 **Role of the funding source**

171 The funding source had no role in study design; data collection, analysis, and interpretation; or report
172 writing. All authors had full access to all data in the study and final responsibility for the decision to
173 submit for publication.

174

175

176 **Results**

177 Of the 355 clinical CRKP isolates from 21 LTACHs across the US,[13] we classified 149 (42%) of the
178 isolates as representing infection based on modified NHSN criteria (**Figure S2**, **Tables S1-3**). Stratified
179 by anatomic site, we classified 29/29 (100%) blood isolates as infection, 69/196 (35%) respiratory isolates
180 as infection, and 51/130 (39%) urinary isolates as infection (**Table S3**). More than 90% of patient isolates
181 were from the epidemic CRKP lineage ST258 (**Tables S1**). Patients harboring different sequence types of
182 CRKP showed no significant differences in infection/colonization status or anatomic site of isolation, and
183 no substantive differences in clinical characteristics (see supplementary material). Thus, we decided to
184 limit our analysis to ST258 to improve our ability to discern whether genetic variation in this dominant
185 strain is associated with infection.

186 **The CRKP epidemic lineage ST258 shows evidence of sub-lineage variation in virulence and**
187 **anatomic site of isolation**

188 We next evaluated if there exist sub-lineages of ST258 with altered virulence properties by looking for
189 clustering of isolates by infection on the whole-genome phylogeny (**Figure 1**; see supplementary
190 methods).[36] Infection status was non-randomly distributed on the phylogeny (p=0·002), supporting our
191 hypothesis that the genetic background of CRKP influences infection. We performed a similar clustering
192 analysis to look at potential niche-specific adaptation to certain anatomic sites (**Figure 1**), and found that
193 respiratory (p=0·001) and urinary (p=0·013) isolates cluster on the phylogeny, but blood isolates do not
194 (p=0·21). This analysis indicates that, in addition to patient features, intra-strain variation in virulence and
195 adaptation to the urinary and respiratory tract might influence whether patients develop an infection.

196 **Both patient and CRKP genetic characteristics are weakly predictive of infection, with relative**
197 **performance being highly facility-dependent**

198 We next performed machine learning to quantify the ability of patient and microbial genetic
199 characteristics to predict CRKP infection (**Figure S1**). To prevent over- or under-fitting and control for
200 facility-level biases, we generated 100 train/test data splits, wherein a given LTACH was only included
201 either in the train or test set. Each LTACH occurred a median of 24 times (range 13-32) in the test data

202 split. In this way, we were able to identify patient and CRKP strain characteristics consistently associated
203 with infection or colonization across data splits, and thus across patient populations in different healthcare
204 facilities.
205 First, we sought to understand if patient and genomic features were individually predictive of CRKP
206 infection. To this end, we independently evaluated patient characteristics as well as three different
207 genomic feature sets for their ability to classify colonization and infection (see methods). Across the 100
208 different train/test splits, we observed that the average predictive performance was weak, with each of the
209 genomic and patient feature sets predictive of infection to a similar degree (all 1st quartile AUROCs >
210 0·5; median range=0·55-0·68; **Figure 2A**; AUPRC: **Figure S3A**). Across the 100 different data splits, no
211 one feature set was consistently the most predictive (e.g. **Figures 2B**, **2C**; all comparisons p > 0·30, see
212 supplementary methods for p-value calculation). Furthermore, for each feature set AUROCs were
213 distributed such that the test AUROC ranged from below 0.5 to over 0·7, depending on how the data were
214 split (i.e., which facilities appear in the train/test sets). This variation in model performance across
215 different train/test sets suggests that the association of CRKP strain and patient characteristics with
216 infection or colonization varies across facilities.
217 **Integration of patient and CRKP strain features does not improve discriminative performance of**
218 **overall or anatomic site-specific models**
219 To determine if the predictive power of patient and genomic features is additive, and if combining these
220 disparate feature sets improved validation on held-out facilities, we built models including both patient
221 and curated genomic features. The discriminative performance of the models based on the combined
222 feature set was not significantly greater than that of the individual feature sets (**Figure 2A**, p ≥ 0·20).
223 Thus, despite variation in the predictive capacity of genomic and patient features across facilities (**Figure**
224 **2C**), combining the two sets did not improve overall performance. Focusing on anatomic site-specific
225 models revealed similar trends, where classification performances were similar for respiratory and urinary
226 specific models, and the relative predictive capacity of patient and CRKP strain features varied across
227 facility subsets (**Figure S4**; AUPRC: **Figure S3B**).

228 **Some patient and genomic features consistently discriminate colonization and infection**

229 After evaluating the predictive capacity of models, we next sought to identify patient and CRKP strain
230 characteristics that are most associated with CRKP infection or colonization. To this end, we identified
231 those patient and genomic features that consistently improved model performance across the 100 different
232 data splits (see methods). Evaluating the importance of features in this way provides insight into those
233 characteristics that generalize across different facility subsets. This approach was taken for both overall
234 and anatomic site-specific models to identify features predictive of different anatomic sites of infection
235 (**Figure 3**, **Figures S5-7**).

236 Several patient features were consistently associated with infection in the overall analysis, including
237 presence of a gastrostomy tube, presence of a central venous catheter, acute kidney injury, and severe
238 chronic kidney disease (**Figure 3**), all markers of critically ill patients. Only a small number of genomic
239 features were consistently associated with infection (**Figure 3**). No known virulence factors were
240 positively associated with colonization; all of the genomic features positively associated with colonization
241 are antibiotic resistance elements. Conversely, all but one of the genomic features positively associated
242 with infection (3/4) are related to virulence. The ICEKp10 element is positively associated with infection
243 and carries colibactin and two different types of yersiniabactin (**Figure S8**). Additionally, insertion
244 sequence-mediated disruption of the O-antigen biosynthetic gene *kfoC* (see supplementary methods and
245 **Figure S9** for insertion sequence identification[20,22,37–39]) was associated with respiratory infection.
246 Colibactin is a toxin,[3] and yersiniabactin is an iron scavenging system that has been identified in previous
247 animal and human studies as being associated with virulence.[9,10] The O-antigen of lipopolysaccharide
248 (LPS) is a known antigenic marker, although association with a specific anatomic site has not been
249 noted.[40]

250 **A sub-lineage of ST258 clade II appears to have sequentially evolved enhanced adaptation for the**
251 **respiratory tract and increased virulence**

252 We noted that *kfoC* disruption is largely confined to a sub-lineage of ST258 (**Figures 4**, **S10**, **S11**).
253 Consistent with this feature being associated with respiratory infection, the disrupted *kfoC* lineage is

254 enriched in respiratory isolates (82/118, 69% of isolates in the disrupted *kfoC* lineage are respiratory
255 isolates vs. 101/213, 47% in all other isolates; Fisher's exact p=0·00011), suggesting that this lineage is
256 associated with increased capacity for respiratory colonization. Furthermore, a subset of isolates in the
257 disrupted *kfoC* sub-lineage contain the ICEKp10 element. Examination of these genetic events in the
258 context of the whole-genome phylogeny revealed that disruption of *kfoC* occurred first, followed by at
259 least two different acquisitions of ICEKp10 (**Figures 4**, **S10**). Within the disrupted *kfoC* lineage, isolates
260 with ICEKp10 are enriched in infection (31/55, 56% of isolates with ICEKp10 are infection isolates vs.
261 16/63, 25% of isolates without ICEKp10, Fisher's exact p = 0·00065), supporting an increase in virulence
262 after acquisition of ICEKp10. It is important to note that the observed clinical associations with ICEKp10
263 and kfoC disruption do not demonstrate causality, as we cannot rule out the role of correlated genetic
264 variation.
265

266 **Discussion**

267 There have been numerous studies aimed at identifying risk factors for healthcare-associated infections
268 caused by prominent antibiotic-resistance threats. For the most part, these studies have found the
269 dominant risk factors to be linked to the magnitude of exposure (e.g. length of stay or colonization
270 pressure), use of antibiotics, and overall comorbidity.[40] Here we found similar results, where length of
271 stay and having certain comorbidities were positively associated with infection. What remains unclear is
272 whether, in the critically ill populations heavily exposed to antibiotics that are at greatest risk, if genetic
273 variation in circulating resistant lineages influences patient infection status. Here, we addressed this
274 question by focusing on CRKP infection in a cohort of patients from 21 LTACHs across the US. To gain
275 a realistic assessment of the predictive capacities of patient and CRKP genetic features, we employed a
276 machine learning framework using multiple facility-level train/test splits. Overall, we found that, while
277 neither patient nor CRKP genetic features have high predictive accuracy on held-out test data, both
278 feature sets were independently associated with infection, with one or the other being more predictive on
279 different facility subsets. Moreover, the integration of clinical and genomic data led to the discovery of an

280 emergent sub-lineage of the epidemic ST258 clone that may have increased adaptation for the respiratory
281 tract, and is more strongly associated with infection.
282 One strength of our machine learning approach is that we were able to measure the variation in
283 discriminative performance across 100 train/test iterations that differed in which facilities were included
284 in train and test sets. We found that performance varied greatly depending on how facilities were
285 allocated to train and test sets, highlighting how smaller studies could overestimate or underestimate the
286 discriminative ability of both their model and individual features. One potential explanation for variation
287 in model performance is that there is facility-level heterogeneity depending on their characteristics (e.g.
288 size, geography, etc.), in which case building sub-models for relevant facility subsets may improve
289 performance. Another possible explanation for variation in model performance may be that the critically-
290 ill nature of LTACH patients is such that most patients are actually highly susceptible to infection (i.e.
291 many patients colonized with CRKP may ultimately develop an infection). However, it's noteworthy that
292 despite these potential challenges in creating generalizable models, our analysis did yield predictors of
293 infection and colonization consistent across test sets, and thus across LTACHs.
294 We built classifiers including all genomic features as well as a curated subset of features, and found that
295 both are similarly weakly predictive of infection. However, while the uncurated feature set presented
296 challenges with downstream interpretation, our analyses on the curated genomic features[15] facilitated
297 novel insights into potential evolutionary trajectories of anatomic site-specific adaptation and virulence.
298 For example, we observed that disruption of the O-antigen biosynthetic gene, *kfoC*, is associated with
299 isolation from the respiratory tract. While we cannot determine from our machine learning analysis if
300 disruption of *kfoC* is directly causal, the biological plausibility of an altered O-antigen structure mediating
301 evasion of innate immunity and/or other beneficial interactions with the host makes this a strong
302 candidate for followup experiments. Supporting this hypothesis, a previous study found that absence of
303 O-antigen is associated with decreased virulence, but not decreased intrapulmonary proliferation, in a
304 murine model.[41] In addition, we noted that a number of antibiotic resistance determinants were associated
305 with colonization. We hypothesize that this observation could be a consequence of longer duration of

306 residence being associated with increased exposure to off-target antibiotics.[42] Finally, we also saw
307 evidence that, after acquiring yersiniabactin and colibactin on the ICEKp10 element, the disrupted *kfoC*
308 subclade became more strongly associated with infection, supporting the idea that circulating ST258 sub-
309 lineages can evolve to become both hypervirulent and multi-drug resistant.[18,43–45]
310 Our study has several important limitations. Specifically, CRKP colonization vs. infection for non-
311 bloodstream isolates may be difficult to discriminate based on surveillance criteria and the clinical data
312 that were available. However, we based our definitions on established CDC criteria with modifications
313 used previously.[7] Encouragingly, we were still able to identify consistent predictors of infection, even
314 with potential misclassifications. A second limitation is that we were limited in the patient data included
315 in our model. It is likely that important differences in underlying patient conditions were not captured by
316 the coarse clinical variables we included, and we also did not account for differences in genetic variation
317 in the host.[46] Other limitations include that our study was restricted to LTACH patients, and had non-
318 random geographic sampling. While LTACHs have unique structural features, based on prior studies, we
319 expect that the types of patient risk factors considered are likely to generalize to other patient populations.
320 Moreover, our restriction to LTACHs in endemic geographic regions has the benefit of focusing on
321 populations at disproportionate risk for CRKP infection.[8] Finally, while the employed machine learning
322 approach allowed for meaningful assessment of discriminative power using a large number of features, by
323 nature it does not yield estimates of attributable risk. However, features identified as consistently
324 associated with colonization or infection on held-out test data can be evaluated by epidemiologists,
325 clinicians, and biologists to identify potential targets for follow-up epidemiologic or laboratory studies.
326

327 **Conclusion**

328 We employed a machine learning approach to quantify our ability to discriminate between CRKP
329 colonization and infection using patient and microbial genomic features. This approach highlighted the
330 high degree of variation in predictive accuracy across different facility subsets. Furthermore, despite
331 modest predictive power, we identified several genomic features consistently associated with infection,

332 indicating that variation in circulating CRKP strains contributes to infection, even in the context of the
333 critically-ill patient populations residing in LTACHs. Future work should aim to corroborate our findings
334 with larger cohorts, and follow up on strong associations to determine whether they are indeed risk factors
335 for infection. This could ultimately help identify patients at high risk for infection and devise targeted
336 strategies for infection prevention.
337

338 **Acknowledgments**

339 We gratefully acknowledge Kindred Healthcare for their assistance in collecting data and isolates used in
340 this study. We also thank Sean Muldoon, MD for his support and guidance throughout the study. Finally,
341 we thank the patients and staff of the long-term acute-care hospitals (LTACHs) for their gracious
342 participation in this study, Begüm Topçuoğlu for help with the machine learning pipeline, Ali Pirani for
343 bioinformatics support, and members of the Snitkin lab, Mike Bachman, and Robert Weinstein for critical
344 review of the manuscript.
345

346 **Author contributions**

347 ESS, JH, EL, and ZL conceptualized the study and acquired funding to support the project. JH and ZL
348 performed data curation. ZL performed formal analysis, investigation, and visualization. ESS, JW, and
349 ZL developed methodology. ESS supervised the project. ZL and ESS wrote the original draft. All authors
350 reviewed and edited the manuscript.
351

352 **Funding**

353 This research was supported by a CDC Cooperative Agreement FOA #CK16-004-Epicenters for the
354 Prevention of Healthcare Associated Infections, and the National Institutes of Health R01 AI139240-01.
355 ZL received support from the National Science Foundation Graduate Research Fellowship Program under
356 Grant No. DGE 1256260. Any opinions, findings, and conclusions or recommendations expressed in this

357 material are those of the authors and do not necessarily reflect the views of the National Science
358 Foundation.
359

360 **Conflicts of interest**

361 JHH was employed at the University of Pennsylvania during the conduct of this study. She is currently an
362 employee of, and holds shares in, the GSK group of companies.
363

364 **References**

365 1. Organization WH. Antimicrobial Resistance: Global Report on Surveillance. Geneva: World Health
366 Organization; 2014. 232 p.
367 2. Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, et al. Clinical
368 epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. Lancet Infect Dis.
369 2013 Sep;13(9):785–96.
370 3. Wyres KL, Lam MMC, Holt KE. Population genomics of Klebsiella pneumoniae. Nature Reviews
371 Microbiology. 2020 Feb 13;1–16.
372 4. Ansari U, Lawsin A, Campbell D, Albrecht V, McAllister G, Bulens S, et al. Molecular
373 Characterization of Carbapenem-Resistant Enterobacteriaceae in the USA, 2011–2015. Open Forum
374 Infect Dis. 2017 Oct 1;4(suppl_1):S179–S179.
375 5. Logan LK, Weinstein RA. The Epidemiology of Carbapenem-Resistant Enterobacteriaceae: The
376 Impact and Evolution of a Global Menace. J Infect Dis. 2017 Feb 15;215(Suppl 1):S28–36.
377 6. Lee C-R, Lee JH, Park KS, Kim YB, Jeong BC, Lee SH. Global Dissemination of Carbapenemase-
378 Producing Klebsiella pneumoniae: Epidemiology, Genetic Context, Treatment Options, and
379 Detection Methods. Front Microbiol [Internet]. 2016 [cited 2018 Feb 10];7. Available from:
380 https://www.frontiersin.org/articles/10.3389/fmicb.2016.00895/full
381 7. Han JH, Goldstein EJC, Wise J, Bilker WB, Tolomeo P, Lautenbach E. Epidemiology of
382 Carbapenem-Resistant Klebsiella pneumoniae in a Network of Long-Term Acute Care Hospitals.

383 Clin Infect Dis. 2017 Apr 1;64(7):839–44.
384 8. Lin MY, Lyles-Banks RD, Lolans K, Hines DW, Spear JB, Petrak R, et al. The Importance of
385 Long-term Acute Care Hospitals in the Regional Epidemiology of Klebsiella pneumoniae
386 Carbapenemase–Producing Enterobacteriaceae. Clin Infect Dis. 2013 Nov 1;57(9):1246–52.
387 9. Lee BY, Bartsch SM, Wong KF, Kim DS, Cao C, Mueller LE, et al. Tracking the spread of
388 carbapenem-resistant Enterobacteriaceae (CRE) through clinical cultures alone underestimates the
389 spread of CRE even more than anticipated. Infection Control & Hospital Epidemiology. 2019
390 Jun;40(6):731–4.
391 10. Bachman MA, Oyler JE, Burns SH, Caza M, Lépine F, Dozois CM, et al. Klebsiella pneumoniae
392 Yersiniabactin Promotes Respiratory Tract Infection through Evasion of Lipocalin 2. Infection and
393 Immunity. 2011 Aug 1;79(8):3309–16.
394 11. Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, et al. Genomic analysis of
395 diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an
396 urgent threat to public health. PNAS. 2015 Jul 7;112(27):E3574–81.
397 12. Martin RM, Cao J, Wu W, Zhao L, Manthei DM, Pirani A, et al. Identification of Pathogenicity-
398 Associated Loci in Klebsiella pneumoniae from Hospitalized Patients. mSystems [Internet]. 2018
399 Jun 26 [cited 2020 Apr 16];3(3). Available from: https://msystems.asm.org/content/3/3/e00015-18
400 13. Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJC, Mattei L, et al. Whole-Genome
401 Sequencing To Identify Drivers of Carbapenem-Resistant Klebsiella pneumoniae Transmission
402 within and between Regional Long-Term Acute-Care Hospitals. Antimicrobial Agents and
403 Chemotherapy [Internet]. 2019 Nov 1 [cited 2019 Dec 18];63(11). Available from:
404 https://aac.asm.org/content/63/11/e01622-19
405 14. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics.
406 2012 Oct 1;28(19):2520–2.
407 15. Holt K. katholt/Kleborate [Internet]. 2020 [cited 2020 Apr 15]. Available from:
408 https://github.com/katholt/Kleborate

409 16. Wyres KL, Wick RR, Gorrie C, Jenney A, Follador R, Thomson NR, et al. Identification of
410 Klebsiella capsule synthesis loci from whole genome data. Microbial Genomics,.
411 2016;2(12):e000102.
412 17. Wick RR, Heinz E, Holt KE, Wyres KL. Kaptive Web: User-Friendly Capsule and
413 Lipopolysaccharide Serotype Prediction for Klebsiella Genomes. Journal of Clinical Microbiology
414 [Internet]. 2018 Jun 1 [cited 2020 Feb 20];56(6). Available from:
415 https://jcm.asm.org/content/56/6/e00197-18
416 18. Lam MMC, Wick RR, Wyres KL, Gorrie CL, Judd LM, Jenney AWJ, et al. Genetic diversity,
417 mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella
418 pneumoniae populations. Microbial Genomics,. 2018;4(9):e000196.
419 19. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and
420 predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012 Apr 1;6(2):80–92.
421 20. Treepong P, Guyeux C, Meunier A, Couchoud C, Hocquet D, Valot B. panISa: ab initio detection of
422 insertion sequences in bacterial genomes from short read sequence data. Bioinformatics. 2018 Nov
423 15;34(22):3795–800.
424 21. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale
425 prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.
426 22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map
427 format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
428 23. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic
429 analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic
430 Acids Res. 2015 Feb 18;43(3):e15–e15.
431 24. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic
432 Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–
433 74.
434 25. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Mol

Biol Evol. 2013 May 1;30(5):1188–95.

26. 2020 NHSN Patient Safety Component Manual. 2020;434.

27. Topçuoğlu BD, Lesniak NA, Ruffin MT, Wiens J, Schloss PD. A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. mBio [Internet]. 2020 Jun 30 [cited 2020 Jun 23];11(3). Available from: https://mbio.asm.org/content/11/3/e00434-20

28. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008 Nov 10;28(1):1–26.

29. R: The R Project for Statistical Computing [Internet]. [cited 2020 Apr 15]. Available from: https://www.r-project.org/

30. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019 Nov 21;4(43):1686.

31. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2” [Internet]. 2019 [cited 2020 Apr 15]. Available from: https://CRAN.R-project.org/package=cowplot

32. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28–36.

33. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol Biol Evol. 2018 Dec;35(12):3041–3.

34. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019 Feb 1;35(3):526–8.

35. grid package | R Documentation [Internet]. [cited 2020 Apr 15]. Available from: https://www.rdocumentation.org/packages/grid/versions/3.6.2

36. Popovich KJ, Snitkin ES, Hota B, Green SJ, Pirani A, Aroutcheva A, et al. Genomic and Epidemiological Evidence for Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus Bloodstream Infections. J Infect Dis. 2017 01;215(11):1640–7.

461 37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
462 architecture and applications. BMC Bioinformatics. 2009 Dec 15;10(1):421.
463 38. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the
464 all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017
465 04;45(D1):D535–42.
466 39. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
467 Bioinformatics. 2009 Jul 15;25(14):1754–60.
468 40. Liu P, Li X, Luo M, Xu X, Su K, Chen S, et al. Risk Factors for Carbapenem-Resistant Klebsiella
469 pneumoniae Infection: A Meta-Analysis. Microbial Drug Resistance. 2017 Jul 27;24(2):190–8.
470 41. Shankar-Sinha S, Valencia GA, Janes BK, Rosenberg JK, Whitfield C, Bender RA, et al. The
471 Klebsiella pneumoniae O Antigen Contributes to Bacteremia and Lethality during Murine
472 Pneumonia. Infection and Immunity. 2004 Mar 1;72(3):1423–30.
473 42. Tedijanto C, Olesen SW, Grad YH, Lipsitch M. Estimating the proportion of bystander selection for
474 antibiotic resistance among potentially pathogenic bacterial flora. Proceedings of the National
475 Academy of Sciences. 2018 Dec 18;115(51):E11988–95.
476 43. Marsh JW, Mustapha MM, Griffith MP, Evans DR, Ezeonwuka C, Pasculle AW, et al. Evolution of
477 Outbreak-Causing Carbapenem-Resistant Klebsiella pneumoniae ST258 at a Tertiary Care Hospital
478 over 8 Years. mBio. 2019 Oct 29;10(5):e01945-19.
479 44. Zhou K, Xiao T, David S, Wang Q, Zhou Y, Guo L, et al. Novel Subclone of Carbapenem-Resistant
480 Klebsiella pneumoniae Sequence Type 11 with Enhanced Virulence and Transmissibility, China -
481 Volume 26, Number 2—February 2020 - Emerging Infectious Diseases journal - CDC. [cited 2020
482 Apr 16]; Available from: https://wwwnc.cdc.gov/eid/article/26/2/19-0594_article
483 45. Gu D, Dong N, Zheng Z, Lin D, Huang M, Wang L, et al. A fatal outbreak of ST11 carbapenem-
484 resistant hypervirulent Klebsiella pneumoniae in a Chinese hospital: a molecular epidemiological
485 study. The Lancet Infectious Diseases. 2018 Jan 1;18(1):37–46.
486 46. Chapman SJ, Hill AVS. Human genetic susceptibility to infectious disease. Nat Rev Genet. 2012

Mar;13(3):175–88.

Figures

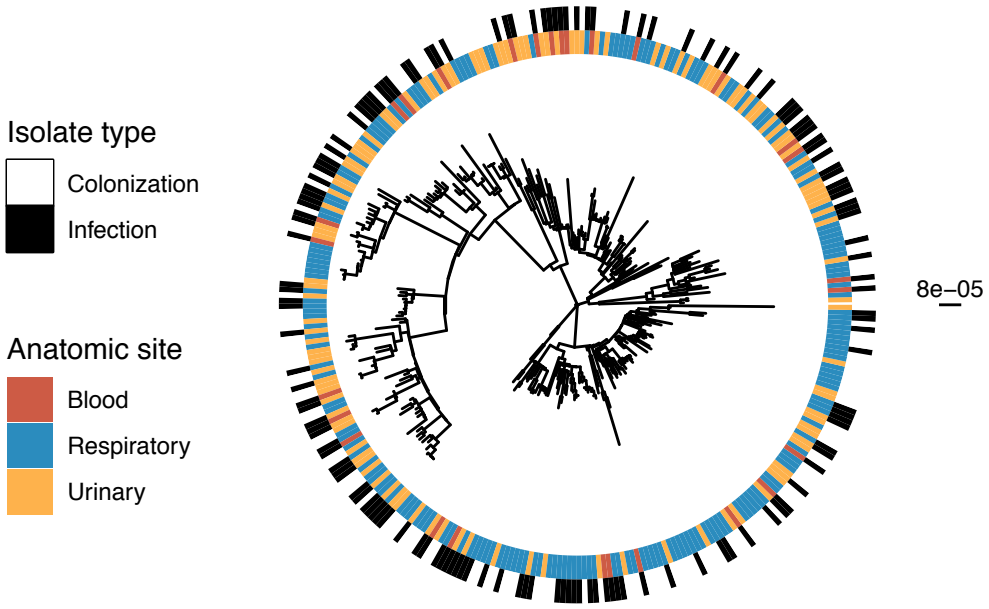


Figure 1: Infection and anatomic site cluster on the phylogeny. Maximum likelihood phylogenetic tree of all isolates including infection or colonization classification for each isolate and anatomic site of isolation. The scale bar to the right of the phylogeny shows the branch length in substitutions per site. Testing for non-random distribution of isolates on the phylogeny (see supplementary methods) revealed clustering of infection, respiratory, and urinary isolates on the phylogeny, respectively.

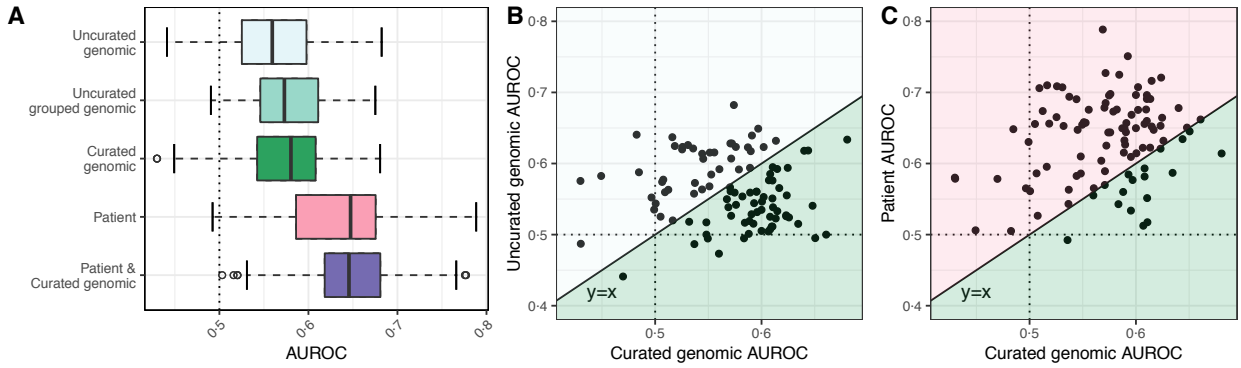


Figure 2: Test AUROCs for various classifiers identifying CRKP colonization vs. infection vary substantially across data splits. (A) Test AUROCs for 100 train/test splits used to build models with L2 regularized logistic regression. All isolates from a given LTACH were included in either the training split or the testing split for each data split. We built models using five different feature sets, keeping the same 100 data splits. AUROCs of different feature sets were not significantly different. In the right two panels, the curated genomic feature set AUROCs are compared to: (B) the uncurated genomic feature set AUROCs, and (C) the patient feature set AUROCs. Each point is the resulting pair of AUROCs for models built with the same data split, but the two respective feature sets. The dotted lines in all 3 panels indicate the AUROC for choosing an outcome randomly (0·5); anything below the line is worse than random, and anything above the line is better than random. The solid diagonal line in the right two panels is the line y=x; points below the line correspond to a higher curated genomic AUROC for that data split, and points above the line correspond to a higher uncurated genomic AUROC (B), or patient AUROC (C), respectively. The colors in panels (B) and (C) correspond to the colors in panel (A); the points in a given colored area indicate that that feature set had the higher AUROC for that data split. In both cases, one feature set does not consistently outperform the other (p=0.4; see supplementary methods for p-value calculation). AUROC=area under the receiver operating characteristic curve.

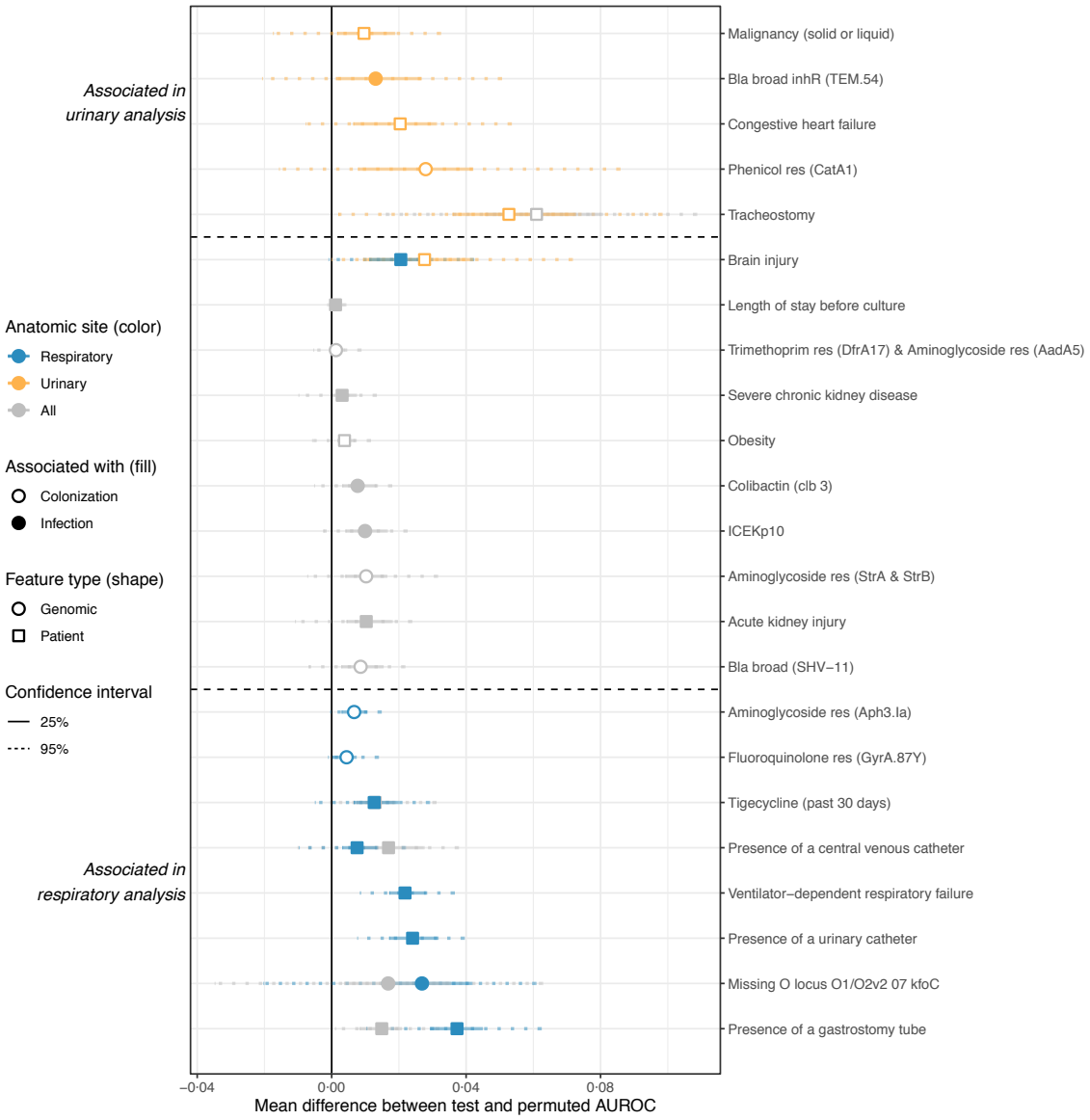
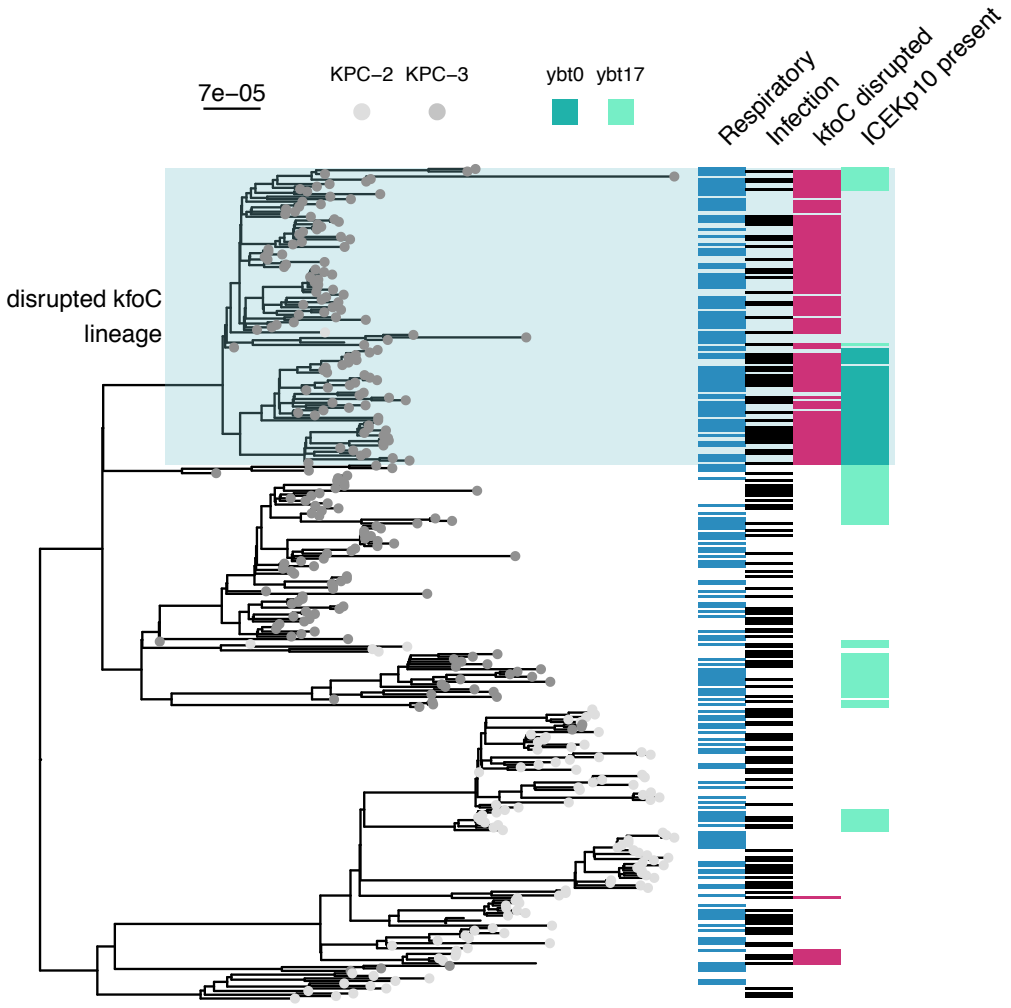


Figure 3: Features consistently associated with colonization or infection sometimes differ between the overall, respiratory, and urinary models. Feature-specific improvement in model performance, measured as the mean difference between test and permuted AUROC (see methods), of features found to be consistently associated with colonization or infection in at least one of the following analyses: overall, respiratory-specific, urinary-specific. We consider features to be associated with infection/colonization if the AUROC difference was greater than zero in over 75% of the 100 data splits. The vertical solid black line indicates a difference of zero (i.e. the feature provides no improvement to model performance). Horizontal dotted lines separate features associated with urinary but not respiratory isolates (top), both urinary and respiratory (or all) isolates (middle), or respiratory but not urinary isolates (bottom). Bla=Beta lactamase, res=confers resistance to that antibiotic class.



524

525 **Figure 4: Select epidemiologic and genomic features visualized on the phylogeny indicate that a sub-clade of**
526 **ST258 clade II may exhibit enhanced niche-specific adaptation and virulence.** ST258 maximum likelihood
527 phylogeny with the tip labels colored by KPC gene. The blue box indicates the sub-lineage with apparent altered
528 niche-specific adaptation that acquires an additional virulence locus. The heatmap beside the tree indicates
529 information about the isolate. From left to right: if it is a respiratory isolate, if it is an infection isolate, if *kfoC* is
530 disrupted, and if it contains ICEKp10. Disrupted *kfoC* was associated with infection in the overall and respiratory
531 machine learning analyses and ICEKp10 presence was associated with infection in the overall analysis. The scale
532 bar to the top left of the phylogeny shows the branch length in substitutions per site. ybt=Yersiniabactin; ybt0 and
533 ybt17 are two ybtSTs defined by Kleborate.