UNIVERSITY OF CALIFORNIA SAN DIEGO

Comparative Pangenomics: Finding structure in the endless diversity of microbial life

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Jason Hyun

Committee in charge:

Professor Bernhard Ø. Palsson, Chair
Professor Theresa Gaasterland, Co-Chair
Professor Rob Knight
Professor Victor Nizet
Professor Debashis Sahoo

2023

The Dissertation of Jason Hyun is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To my mother and father

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY DATASETS

ACKNOWLEDGEMENTS

Though most people that know me are familiar with my tendency to work in solitude, it is through the guidance and kindness of those who continuously reached out and supported me that I have been able to overcome the challenges faced in the development of this dissertation.

First and foremost, I am grateful for my mentors, for their continued support through my academic career and the passion they helped instill for tackling the challenges of this work. Jonathan Monk, who introduced me to the wonders of pangenomics as well as the terrors of antimicrobial resistance and guided me through my first real research project, from conception to publication. Daniel Zielinski, for his excellent instruction in applying machine learning to the mess that is biological data; his quote "to a data scientist, every dataset looks like a treasure trove" continues to drive my search for insights from such data. And last but certainly not least, Professor Bernhard Palsson, for both his unparalleled wisdom in turning mathematical analyses into meaningful biological insights and for providing me with all the right opportunities to grow my technical skills and build my confidence as a scientist.

I am also grateful for all the wonderful members of the SBRG with whom I had the fortune of working with. Patrick Phaneuf, one of the first people I met at the SBRG, was a tremendous source of both material and emotional support when navigating each milestone of the Bioinformatics and Systems Biology program, and surely this journey would have been much more harrowing without his help. Erol Kavvas, for the exciting and inspiring discussions on how to make machine learning useful in understanding microbial life. Siddharth Chauhan, for many insightful discussions on how to continue developing pangenomics and for taking the charge in advancing broader pangenome efforts at the SBRG. Richard Szubin and Ying Hefner, for patiently taking on the challenge of experimentally validating my computational results. And Marc Abrams, for somehow managing all the little things and countless tasks necessary for my time at the SBRG to

have been as smooth as it was.

I would also like to thank my funding sources that have supported this work. These include the National Institute of Allergy and Infectious Diseases (U01-AI124316), the National Institutes of Health (T32GM8806), and the Novo Nordisk Foundation (NNF-20CC0035580).

Finally, I'd like to thank my friends from high school, Viju, Thien, and Phoebe, for tolerating my random months-long disappearances from social media and making sure my life had a little bit of color between the dry coding and statistical analyses of this dissertation.

Chapter 2 is a reprint of material published in: **Jason C Hyun**, Jonathan M Monk, Bernhard O Palsson. 2022. "Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity." *BMC Genomics* 23(1):7. The disseration author is the primary author.

Chapter 3 is a reprint of material published in: **Jason C Hyun**, Erol S Kavvas, Jonathan M Monk, Bernhard O Palsson. 2020. "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens." *PLoS computational biology* 16(3):e1007608. The disseration author is the primary author.

Chapter 4 has been submitted for publication in *Nature Communcations*: **Jason C Hyun**, Jonathan M Monk, Richard Szubin, Ying Hefner, Bernhard O Palsson. "Global pathogenomic analysis identifies known and novel genetic antimicrobial resistance determinants in twelve species." The disseration author is the primary author.

Chapter 5 has been submitted for publication in *Proceedings of the National Academy of Sciences*: **Jason C Hyun** and Bernhard O Palsson. "Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome." The disseration author is the primary author.

VITA

2017 Bachelor of Science, Chemical-Biological Engineering and Biology, Massachusetts Institute of Technology

2023 Doctor of Philosophy, Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

Jason C Hyun, Erol S Kavvas, Jonathan M Monk, Bernhard O Palsson. 2020. "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens." *PLoS computational biology* 16(3):e1007608.

Jason C Hyun, Jonathan M Monk, Bernhard O Palsson. 2022. "Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity." *BMC Genomics* 23(1):7.

Jason C Hyun, Jonathan M Monk, Richard Szubin, Ying Hefner, Bernhard O Palsson. "Global pathogenomic analysis identifies known and novel genetic antimicrobial resistance determinants in twelve species." *Under review.*

Jason C Hyun and Bernhard O Palsson. "Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome." *Under review.*

Saugat Poudel, **Jason C Hyun**, Ying Hefner, Victor Nizet, Bernhard O. Palsson. "Interpreting roles of mutations in the emergence of S. aureus USA300 strains with genetics and independent component analysis of gene expression." *Under review.*

Akanksha Rajput, Siddharth M Chauhan, Omkar S Mohite, **Jason C Hyun**, Omid Ardalani, Leonie J Jahn, Morten OA Sommer, Bernhard O Palsson. "Pangenome analysis reveals the genetic basis for taxonomic properties of the Lactobacillaceae family." *In preparation.*

ABSTRACT OF THE DISSERTATION

Comparative Pangenomics: Finding structure in the endless diversity of microbial life

by

Jason Hyun

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Bernhard Ø. Palsson, Chair
Professor Theresa Gaasterland, Co-Chair

Rapid developments in genome sequencing technology have revealed tremendous genetic diversity in the bacterial kingdom. The full genetic repertoire, or pangenome, of any microbial species has been found to be far more expansive than that of any individual organism, and understanding this complexity promises insights into both fundamental and practical questions regarding the diversity of microbial life. This dissertation aims to systematize the analysis of pangenomes to enable comparisons between multiple pangenomes and build generalizable workflows for investigating complex biological phenomena not limited to individual species. First, a robust pipeline for pangenome construction is pre-

sented as the foundation of this work and used to identify constants in intraspecies genetic diversity across twelve microbial species. Second, pangenome construction is combined with machine learning to elucidate global patterns of antimicrobial resistance (AMR) and identify novel AMR-conferring gene candidates. Finally, pangenome analysis is scaled to limits of publicly available data to construct and compare the core genomes of 183 species, and integrated with phylogenetics to reconstruct the core genome of the last bacterial common ancestor and identify implications regarding the minimum genetic requirements for life. These results demonstrate how pangenomes can reveal novel facets of genetic diversity previously invisible at smaller scales, and continued development of pangenomics will be necessary to close the gap between the pace of data collection and the pace of discovery.

Chapter 1

Introduction

One of the primary endeavors of biological science is to identify and explain the diversity of life, whether observed across the myriad environments of the natural world, under the pressures of various human interventions, or in synthetic conditions designed to produce specific traits. Perhaps the most significant recent development to impact these efforts has been the rapid fall in the cost of genome sequencing, enabling truly comprehensive comparisons across closely related and distant organisms alike through the common language of DNA. However, rather than cement our understanding of known species, the consequent explosion of genetic data revealed an unprecedented level of diversity. With each new organism sequenced revealing hundreds of never-before-seen genes within even a single, relatively simple microbial species, novel analytical techniques are necessary to better characterize the full genetic repertoire of a species beyond individual strains and the resulting phenotypic consequences [1].

1.1 The microbial pangenome

One of the concepts developed to help understand this diversity is the "pangenome," or the set of all genes observed in a set of genomes, such as those from a single species. Many tools have been developed for constructing pangenomes [2], which often follow three main steps: 1) identifying open reading frames in all genomes, 2) clustering the

open reading frames by sequence similarity, and 3) treating each sequence cluster as a set of variants for a single gene, which can be further characterized by distribution and function. This workflow takes advantage of modern computational resources for the *de novo* identification of genes and can accommodate any number of newly encountered genes that would be beyond the scope of previous reference genome-dependent approaches [3].

The pangenome concept provides structure to what are initially amorphous collections of genome sequences. This structure has enabled comprehensive empirical assessments of how strongly each gene is conserved within a species, and analyses of microbial species have yielded two consistent properties of pangenomes. First, the relationship between the number of genomes and the number of unique genes encountered follows a power law relationship or "Heaps' Law," formalizing the seemingly endless novel genes observed when sequencing new strains of a known species [4]. This functional form allowed a species' "openness," or tendency for new strains to harbor novel genes, to be quantified and interpreted in the context of lifestyle and horizontal gene transfer [1, 4]. Second, pangenomes divide into three components based on gene frequency: the core genome (genes present in all genomes, representing the most strongly conserved elements of the species), the accessory genome (genes present in some genomes, representing weakly maintained elements), and the unique genome (genes present in only a small fraction of genomes, representing strain-specific elements) [1, 5, 6]. These core observations continue to guide pangenome studies of increasing scale, and analyses at the intersection between gene frequency, function, and conservation have enabled more nuanced characterizations of a species' capabilities and lifestyle [5].

1.2 Machine learning for pangenome-scale discovery

A valuable product of pangenome construction is the enumeration of all genes and variants in a collection of genomes, which allows the totality of coding sequence variation

to be represented as a series of presence/absence calls, suited for further mathematical analysis. When combined with additional genome metadata (such as that for antimicrobial resistance, infection site, or metabolic activity), associations between these pangenomic features and a complex phenotype can be identified by borrowing statistical methods from genome-wide association studies (GWAS) [7]. However, due to the clonal nature of bacteria, microbial pangenome datasets are highly heterogeneous and can confound traditional GWAS methods that are often designed around human datasets [7, 8]. Consequently, machine learning (ML) approaches are emerging as a promising alternative for biological discovery with microbial pangenomes. Though many studies have already demonstrated the power of ML at predicting a phenotype from pangenomic features, more recent efforts have also aimed to discover novel genotype-phenotype associations by dissecting trained ML models and identifying their most strongly predictive features [9].

One such application of ML has been in the analysis of antimicrobial resistance (AMR), the evolutionary phenomenon of microorganisms adapting to resist previously effective drugs for treating infections. AMR remains a persistent and pervasive threat to public health, currently responsible for 700,000 annual deaths globally [10]. Consequently, one strategy developed to help manage AMR has been the large-scale sequencing of infection isolates, yielding tens of thousands of genome sequences paired with experimental AMR metadata describing each strain's susceptibility/resistance to various drugs [11]. While the direct application of ML models to this data has yielded tools for rapidly predicting the resistance profiles of clinical isolates that are beginning to guide treatment strategies for certain pathogens [12, 13], the integration of pangenome methods has increased their interpretability. Pangenome construction allows ML models to be trained on distinct biological features rather than arbitrary DNA sequences, of which the most predictive can be readily linked to genetic and metabolic mechanisms to explain their contribution to AMR and identify potential novel drug targets [14, 15]. These analyses demonstrate the promise of combining pangenomics and ML for not just predicting but also explaining

the rapidly evolving phenomenon of AMR and will likely prove useful for elucidating the genotype-phenotype relationship for other complex phenotypes.

1.3 Beyond single pangenomes

Just as the falling cost of genome sequencing has enabled comparative genomics to evolve from individual reference genomes to pangenomes, pangenomic analysis stands at a similar tipping point in scale. Given the wide variety of microbial pangenomes already examined individually, there is sufficient genetic data available for the systematic analysis and comparison of multiple species at the pangenome scale, i.e. "comparative pangenomics." With the pangenome concept as the foundation of this dissertation, this work sets out demonstrate robust methods for identifying comparable structures of genetic diversity across multiple pangenomes beyond the two previous core observations, and apply them towards the ML-driven elucidation of AMR and reconstruction of ancient bacterial ancestors.

The following chapters describe four studies centered around achieving these aims. Chapter 2 defines a consistent procedure for pangenome construction and identifies similarities in the relationship between gene frequency, function, and sequence conservation across twelve pathogens at multiple resolutions, from the size of the pangenome down to the locations of individual mutations. Chapter 3 combines pangenome construction with ML and AMR metadata for three pathogens to identify novel AMR gene candidates and demonstrate the competitiveness of ML at identifying genetic determinants of AMR when compared to traditional GWAS methods. Chapter 4 refines the methodology of Chapter 3 and expands the analysis to twelve pathogens to characterize the limits of cross-species AMR gene transfer, identify additional AMR gene candidates, and provide experimental validation for two such candidates. Finally, Chapter 5 scales pangenome analysis up to the limits of genomic data publicly available at the time of writing, constructing pangenomes

for 183 species, comparing the content of their respective core genomes, and integrating the core genomes with the bacterial phylogenetic tree to reconstruct the core genome of the last bacterial common ancestor for insights into the lifestyle of ancient bacteria and the minimum genetic requirements for life.

1.4 References

[1] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev.*, 15(6):589–594, December 2005.

[2] G S Vernikos. A review of pangenome tools and recent studies. In *The Pangenome*, pages 89–112. Springer International Publishing, Cham, 2020.

[3] Jason C Hyun, Jonathan M Monk, and Bernhard O Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022.

[4] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11(5):472–477, October 2008.

[5] Sávio Souza Costa, Luís Carlos Guimarães, Artur Silva, Siomar Castro Soares, and Rafael Azevedo Baraúna. First steps in the analysis of prokaryotic pan-genomes. *Bioinform. Biol. Insights*, 14:1177932220938064, August 2020.

[6] Bo Segerman. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front. Cell. Infect. Microbiol.*, 2:116, September 2012.

[7] Robert A Power, Julian Parkhill, and Tulio de Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.*, 18(1):41–50, January 2017.

[8] Peter E Chen and B Jesse Shapiro. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.*, 25:17–24, June 2015.

[9] Hannah L Nicholls, Christopher R John, David S Watson, Patricia B Munroe, Michael R Barnes, and Claudia P Cabrera. Reaching the end-game for GWAS:

Machine learning approaches for the prioritization of complex disease loci. *Front. Genet.*, 11:350, April 2020.

[10] Jim O'Neill and Wellcome Trust. Antimicrobial resistance : Tackling a crisis for the health and wealth of nations, December 2014.

[11] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, Rory M Butler, Philippe Chlenski, Neal Conrad, Allan Dickerman, Emily M Dietrich, Joseph L Gabbard, Svetlana Gerdes, Andrew Guard, Ronald W Kenyon, Dustin Machi, Chunhong Mao, Dan Murphy-Olson, Marcus Nguyen, Eric K Nordberg, Gary J Olsen, Robert D Olson, Jamie C Overbeek, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, Chris Thomas, Margo VanOeffelen, Veronika Vonstein, Andrew S Warren, Fangfang Xia, Dawen Xie, Hyunseung Yoo, and Rick Stevens. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.*, 48(D1):D606–D612, January 2020.

[12] Jee In Kim, Finlay Maguire, Kara K Tsang, Theodore Gouliouris, Sharon J Peacock, Tim A McAllister, Andrew G McArthur, and Robert G Beiko. Machine learning for antimicrobial resistance prediction: Current practice, limitations, and clinical perspective. *Clin. Microbiol. Rev.*, 35(3):e0017921, September 2022.

[13] Melis N Anahtar, Jason H Yang, and Sanjat Kanjilal. Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research. *J. Clin. Microbiol.*, 59(7):e0126020, June 2021.

[14] Erol S Kavvas, Laurence Yang, Jonathan M Monk, David Heckmann, and Bernhard O Palsson. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.*, 11(1):2580, May 2020.

[15] Jason C Hyun, Erol S Kavvas, Jonathan M Monk, and Bernhard O Palsson. Machine learning with random subspace ensembles identifies antimicrobial resistance determi-

nants from pan-genomes of three pathogens. *PLoS Comput. Biol.*, 16(3):e1007608, March 2020.

Chapter 2

Comparative pangenomics: Revealing conserved structures of genetic and functional diversity

2.1 Abstract

With the exponential growth of publicly available genome sequences, pangenome analyses have provided increasingly complete pictures of genetic diversity for many microbial species. However, relatively few studies have scaled beyond single pangenomes to compare global genetic diversity both within and across different species. We present here several methods for "comparative pangenomics" that can be used to contextualize multi-pangenome scale genetic diversity with gene function for multiple species at multiple resolutions: pangenome shape, genes, sequence variants, and positions within variants. Applied to 12,676 genomes across 12 microbial pathogenic species, we observed several shared resolution-specific patterns of genetic diversity: First, pangenome openness is associated with species' phylogenetic placement. Second, relationships between gene function and frequency are conserved across species, with core genomes enriched for metabolic and ribosomal genes and accessory genomes for trafficking, secretion, and defense-associated genes. Third, genes in core genomes with the highest sequence diversity are functionally diverse. Finally, certain protein domains are consistently mutation enriched across multiple

species, especially among aminoacyl-tRNA synthetases where the extent of a domain's mutation enrichment is strongly function-dependent. These results illustrate the value of each resolution at uncovering distinct aspects in the relationship between genetic and functional diversity across multiple species. With the continued growth of the number of sequenced genomes, these methods will reveal additional universal patterns of genetic diversity at the pangenome scale.

2.2 Background

With the falling cost of sequencing spurring exponential growth in publicly available genome sequences, genetic analyses have similarly increased in scale over the past three decades, from the first complete microbial genome assemblies in 1995, to comparisons between reference strains of model organisms, and now to simultaneous analyses of thousands of genomes from samples isolated worldwide for multiple species [1, 2]. These pangenome analyses have provided increasingly complete pictures of genetic diversity for most major microbial pathogens, revealing species-level properties invisible at smaller scales, such as the nature of species-wide conserved core genomes compared to their more variable accessory genomes [3, 4], or the tendency for newly sequenced strains of a species to harbor previously unobserved genes, commonly referred to as pangenome openness [5, 6]. Furthermore, pangenomes have formed the basis of many multi-strain characterizations of clinically relevant phenotypes such antimicrobial resistance [7], virulence [8], or metabolic capabilities [9].

However, while the variety of species studied and increasing automation of pangenome workflows [10] attest to the versatility of the pangenome for large-scale genome analysis, pangenome studies are currently dominated by those that focus on one species at a time or combine multiple related species into a single pangenome. Relatively few studies describe methods for comparing distinct pangenomes beyond the sizes of core or accessory

genomes: Since Tettelin et.al. introduced the bacterial pangenome and Heaps' Law as a model for quantifying and comparing openness [6], other multi-pangenome works have compared pangenome openness estimates using alternate models beyond Heaps' Law [11], level of conservation within core genomes [12, 13], extent of functional characterization in core and pangenomes [14], and functional distributions between core and accessory genomes of different species or environmental isolates [11, 12, 15, 16]. These methods focus primarily on pangenome scaling or the distribution of gene-level functions and are limited in their analysis of finer genetic variation such as individual sequence variants often examined in single pangenome studies. Consequently, existing pangenome studies often present a tradeoff between "scale" (number of species, genomes, or pangenomes analyzed) and "resolution" (smallest unit of genetic diversity analyzed).

To address this gap in pangenome analysis, we present generalizable "comparative pangenomics" methods to examine genetic and functional diversity within and between 12 pangenomes of pathogenic organisms totalling 12,676 genomes. These analyses span several levels of resolution: pangenome shape, individual genes, individual sequence variants, and specific positions within variants. Contextualizing genes against the other three resolutions provides distinct perspectives of diversity at the pangenome scale: 1) gene conservation within the species (core vs. accessory genes), 2) conservation of the gene sequence overall (number and frequency of individual variants), and 3) conservation of specific regions or domains within the gene sequence (positions with high or low diversity within aligned variants). In addition to standard pangenome analyses, we compare functional annotations against these forms of genetic diversity to identify which gene functions are consistently stable against or subjects of major variation across a variety of pathogens.

2.3 Results

2.3.1 Pangenome construction for reference genome-free enumeration of genetic variation

A total of 12,676 genomes across 12 different species were downloaded from the PATRIC database [17] after filtering for assembly quality (see Methods), ranging from 104 to 3183 genomes per species (Fig. 2.1a, Fig. A.1a, Table A.1, Dataset A.1). Each genome was classified *in silico* by multilocus sequence type (MLST) (Fig. A.1b, Dataset A.2) using the mlst tool (https://github.com/tseemann/mlst) based on PubMLST [18]. For each species, a pangenome was constructed by clustering open reading frames by protein coding sequence into putative genes clusters, using CD-HIT [19]. These cluster-derived genes were used to define three other genetic feature types within each pangenome, namely "coding variants" (individual protein sequence variants of the genes, based on members of gene sequence clusters), "5′ IG variants" (DNA variants of the 300nts directly upstream of all observed instances of a given gene), and "3′ IG variants" (DNA variants of the 300nts directly downstream of a gene).

2.3.2 A subtype-based estimate of pangenome size and openness using Heaps' law is more accurate than genome-based estimates

At the broadest resolution of genetic diversity, pangenome openness, or the tendency for new genomes of a given species to introduce new genes, can be used quantify overall gene-level diversity within a species at the pangenome level as well as project pangenome size as additional genomes are sequenced. Openness is most commonly estimated as the power law exponent when fitting Heaps' law to pangenome size versus number of genomes, across many iterations of randomly shuffling genome order [6]. This application of Heaps' Law is based on its original discovery in linguistics as an empirical relationship between the number of unique words encountered and the number of documents reviewed, for

which an analogous relationship between genes encountered and genomes sequenced has been observed for multiple bacterial pangenomes [6, 11]. However, MLST classification revealed that the genomes available for some species were highly biased for one or a few subtypes (i.e. 75% of *E. faecium* genomes are from MLST 80), while others were more diverse (Fig. 2.1b, Fig. A.1b). Consequently, estimating pangenome openness based on new genes discovered per genome may in some cases be more characteristic of a single subtype and underestimate overall species-wide openness and/or extrapolate pangenome size poorly.

To address this, we estimated openness with Heaps' law using two methods to generate 100 random genome orderings per species: 1) the standard approach of randomly shuffling all genomes, and 2) randomly selecting one genome per MLST subtype and shuffling the selected genomes (Fig. 2.1c). MLST-based estimates of openness were greater than genome-based estimates in 10/12 species without any notable increase in the standard deviation of the estimates, with larger differences observed in more strongly subtype biased cases; the openness estimate for the strongly subtype-biased *E. faecium* case is nearly doubled when using the MLST-based estimate (Fig. 2.1d, Table A.2).

To compare accuracy at extrapolating pangenome size, Heaps' Law fits were computed on the first half of genomes and evaluated on the second half for all random genome orderings (i.e. for a species of 200 genomes and 20 MLST types, the genome-based approach would be fit to the first 100 genomes and evaluated on the last 100 genomes, while the MLST-based approach would be fit to the first 10 and evaluated on the last 10) (Fig. A.2a). The median mean absolute error (MAE) for the MLST-based approach was lower in 11/12 species in the fit region and 9/12 species in the extrapolation region, despite having fewer points to fit (Fig. A.2b-c, Table A.3). The two cases where the MLST-based approach underperformed the genome-based approach were *P. aeruginosa* (2.0 times larger MAE) and *S. enterica* (1.5 times larger MAE). As the estimated openness and MLST distribution diversity for these species are not particularly different from that

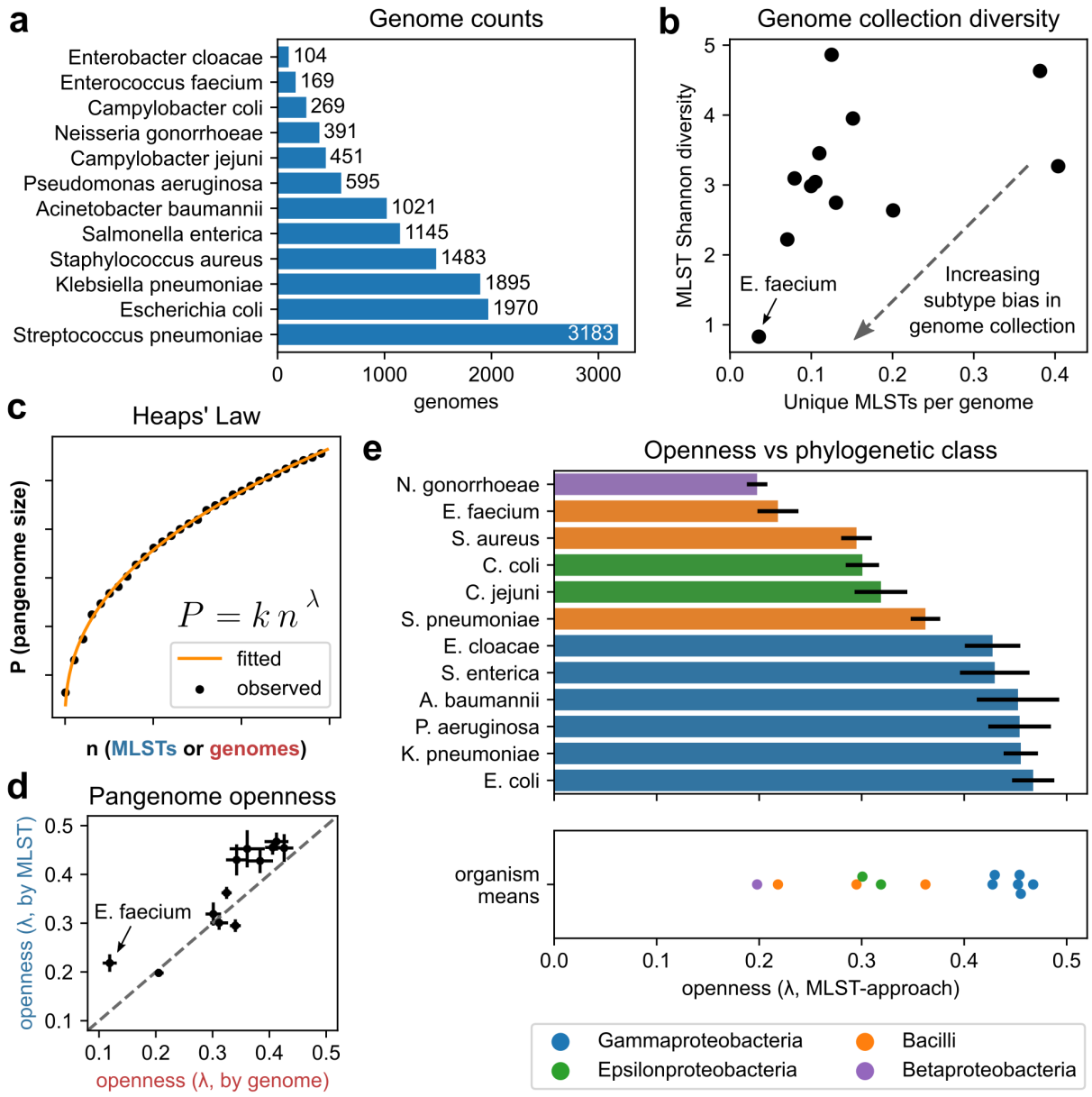


Figure 2.1. Subtype-balanced Heaps' Law estimates of pangenome openness for 12 microbial pathogens. (a) Total number of genomes analyzed per species. (b) MLST subtype diversity per species' genome collection, as quantified by unique MLSTs per genome and Shannon diversity of MLST distributions. Low diversity outlier *E. faecium* is labeled. (c) Example fit of Heaps' Law between the number of genomes or MLSTs versus running pangenome size. (d) Comparison of openness values estimated per species, with or without controlling for MLST (error bars are standard deviations from 100 estimates). (e) Means and standard deviations of openness estimates after controlling for MLST, versus phylogenetic class.

of other species, one possible explanation may be due to these cases resulting in relatively poor fits to Heaps' Law in general, being the 1st and 2nd largest MAE cases by the MLST-based approach and the 3rd and 5th largest MAE cases by the genome-based approach, respectively.

Overall, the MLST-based Heaps' Law approach appears to extrapolate pangenome size more consistently than a full genome-by-genome approach, and may offer a more accurate depiction of the genetic diversity of a given species even when using subtype-biased datasets. The calculated openness values appear to cluster roughly by species' phylogenetic classification (Fig. 2.1e). The top 6 most open pangenomes cluster closely ($\lambda = 0.42$-0.47) and consist of the six Gammaproteobacteria class species examined (*E. cloacae, S. enterica, A. baumannii, P. aeruginosa, K. pneumoniae, E. coli*), followed by a group with intermediate openness ($\lambda = 0.29$-0.36) of two Bacilli class species (*S. aureus, S. pneumoniae*) and the two Campylobacter species (*C. coli, C. jejuni*), and finally the two most closed species ($\lambda = 0.20$-0.22) consisting of *E. faecium* (Bacilli) and *N. gonorrhoeae* (Betaproteobacteria).

2.3.3 Frequency-based division of the pangenome using power functions

Moving from the resolution of overall pangenome shape to individual genes, the distribution of gene frequencies (number of genomes each gene is observed in) was computed for each pangenome to begin exploring sources of genetic diversity in greater detail. Regardless of the number of genomes available or the estimated pangenome openness, all such distributions demonstrate a peak for very rare genes and a smaller peak for highly conserved or core genes (Fig. 2.2a, Fig. A.3). Correspondingly, the cumulative gene frequency distribution takes on an asymmetric, inverse sigmoidal shape, which suggests three intuitive frequency categories by which genes may be classified: the initial asymptotic region consisting of rare, poorly characterized genes representing the "unique" genome,

the opposite asymptotic region consisting of highly conserved genes representing the "core" genome, and the middle linear region consisting of uncommon genes representing the "accessory" genome which captures most of the gene-level diversity in the pangenome (Fig. 2.2a).

Three-part frequency divisions of the pangenome have been previously described, and often achieved through either static thresholds such as having core genes being those in all genomes and unique genes being those in exactly one genome [3], or more scalably through fitting frequency distributions to multiple exponential functions to identify analogous "core-shell-cloud" divisions of the pangenome [20, 21]. Here, we developed an approach based on fitting the distribution to the sum of two power functions and defining the accessory genome relative to the inflection point in the cumulative distribution. This functional form is derived from the observation that gene frequency distributions tend to resemble power laws for very small and very large frequencies (Fig. 2.2b-c), and achieves accurate fits to cumulative frequency distributions with MAE ranging from 25 to 116 genes, less than 0.5% of the total pangenome size for 11/12 species. The fits also achieve $R^2 > 0.99$ for 10 of 12 pangenomes and a minimum of 0.964 (Fig. A.4, Dataset A.3). The frequency thresholds for core-accessory-unique divisions defined from these fits ranged from 95.8 to 98.6% of all genomes for core genes, and 5.8 to 8.6% for unique genes, highlighting the asymmetry present in the original frequency distributions (Table A.4).

This pangenome division approach yielded core genomes ranging from 4.3 to 34.2% of their corresponding pangenomes with a minimum of 1,237 core genes in *C. jejuni* to a maximum of 4,585 in *P. aeruginosa*, while accessory genomes ranged from 5.0 to 38.1% of their corresponding pangenomes with a minimum of 1,046 accessory genes in *C. coli* to 5,046 in *E. coli* (Fig. 2.3a, Table A.4). Core genomes were similar in size to corresponding accessory genomes and larger genomes were associated with more open pangenomes, though there was no relationship between the ratio of core to accessory genome size and openness (Fig. 2.3b). Overall, by creating three frequency categories,

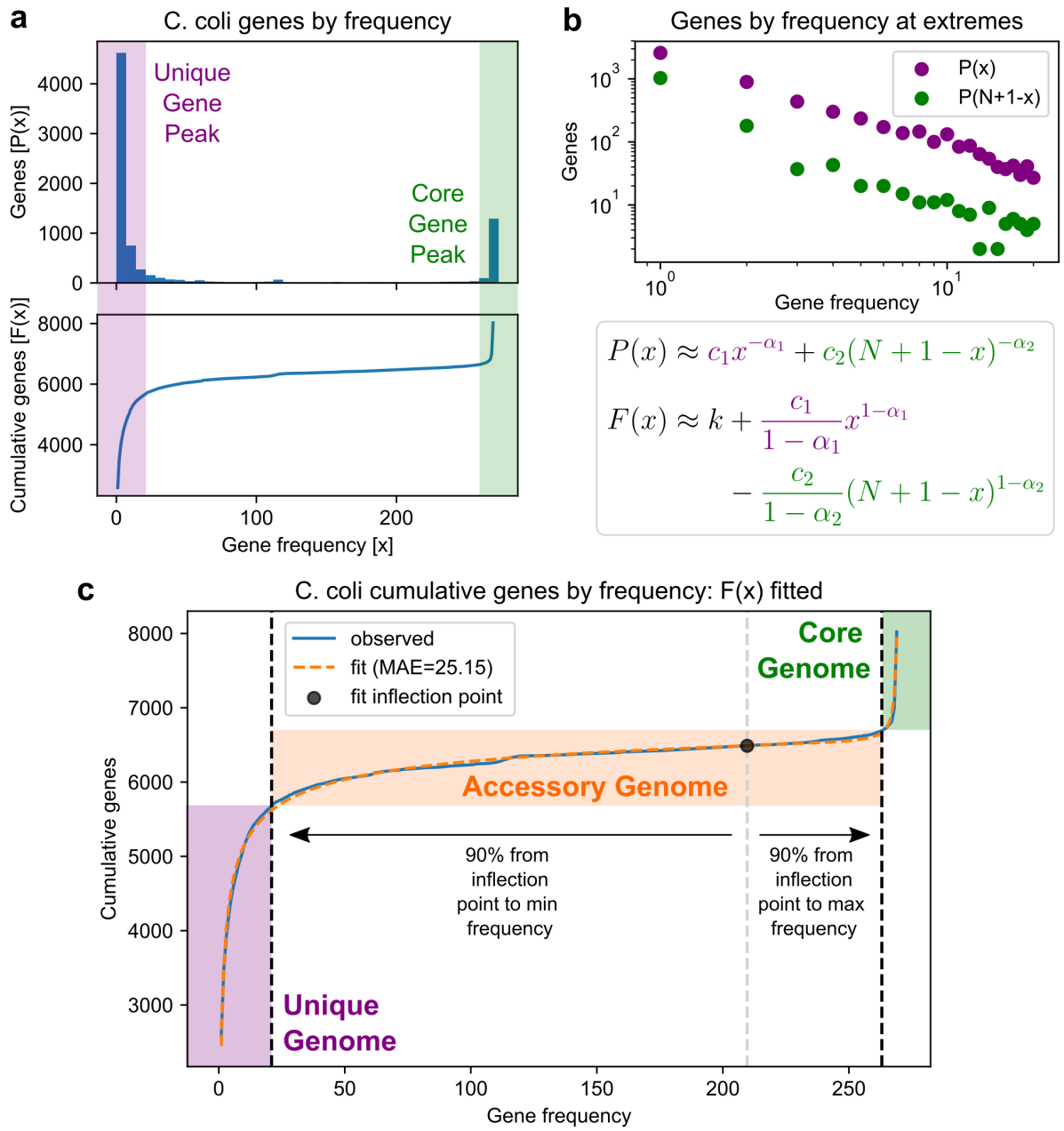


Figure 2.2. Example division of the *Campylobacter coli* pangenome into unique, accessory, and core genomes. (a) Distribution of gene frequencies $P(x)$, or the number of times a gene is observed in a genome, with peaks at very low ("unique") and very high ("core") frequencies. The corresponding cumulative distribution $F(x)$ is shown below. (b) Log-log plots of the frequency distribution at very low and very high frequencies showing approximately linear trends, and the corresponding models of $P(x)$ as the sum of two power functions and $F(x)$ as the integral. N is the total number of genomes. (c) Division of the pangenome into unique, accessory, and core genomes based on the cumulative distribution fit. Frequency thresholds for unique and core genes are defined relative to the fitted inflection point.

this method allows subsequent analyses to focus on a smaller number of genes (relative to full pangenomes) such as highly conserved core genes or accessory genes that constitute most of the gene-level diversity in an species, rather than the more abundant but often under-characterized or erroneous unique genes.

2.3.4 Consistent enrichment of specific gene functions in core and accessory genomes

To identify associations between gene frequency and function, all genes were annotated for Clusters of Orthologous Groups (COG) functional categories and GO terms using eggNOG-emapper [22], and Fisher's exact tests were conducted between each frequency category and COG category within each pangenome (Dataset A.4). This revealed consistent enrichment of several metabolic COGs in the core genome, with COGs C (energy production and conversion), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), and H (coenzyme transport and metabolism), as well as non-metabolic COGs J (translation, ribosomal structure and biogenesis) and O (post-translational modification, protein turnover, and chaperones), significantly enriched in the core genome for at least 11/12 species ($p < 7 * 10^5$, FWER < 0.05 with Bonferroni correction) and mean $\log_2$ odds ratios (LOR) ranging from 1.7 to 2.8 across the species (Fig. 2.3c, Fig. A.5a). Accessory genomes also showed frequent, albeit weaker functional enrichments, with two COGs with mean LOR > 1 across the species: U (intracellular trafficking, secretion, and vesicular transport) was enriched in 9/12 species with a mean LOR of 1.2, and V (defense mechanisms) enriched in 7/12 species also with a mean LOR of 1.2 (Fig. 2.3c, Fig. A.5b). Finally, with the exception of COG S (function unknown), no COGs were found with either frequent significant enrichment or mean LOR > 1 in the unique genomes, owing to their relatively poor characterization (Fig. A.5c).

A similar analysis of GO terms revealed that 9 of the top 10 enriched GO terms in core genomes by mean LOR were associated with ribosomes or RNA processing (LOR = 3.8-

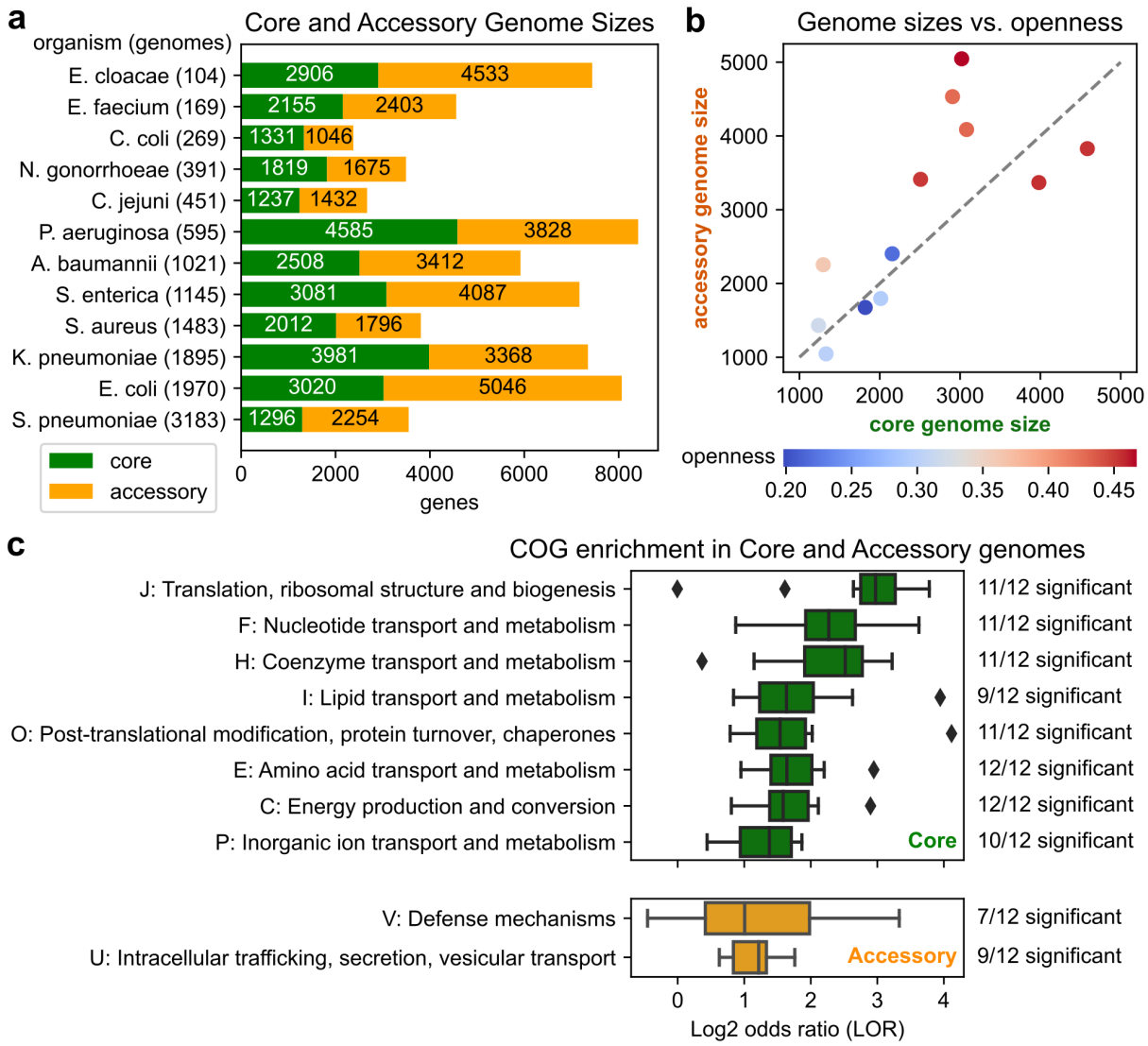


Figure 2.3. Genes and functional enrichments in the core and accessory genomes of 12 species. (a) Number of genes in the core and accessory genomes of each species. (b) Comparison of core genome size, accessory genome size, and pangenome openness. (c) Functional enrichments by COG functional category in the 12 core and accessory genomes. The distribution of $\log_2$ odds ratios (LORs), as well as the number of species with significant enrichment by COG are shown (Fisher's exact test, FWER < 0.05 under Bonferroni correction or $p < 7 * 10^{-5}$, 720 tests). Only COGs showing positive enrichment in over half the species and with mean LOR > 1 are shown. COG "S: Function unknown" is not shown.

6.9). All but one of those terms was also significantly enriched in at least 11/12 species ($p < 3 * 10^{-6}$, FWER < 0.05 with Bonferroni correction), consistent with the J COG previously found enriched in core genomes (Fig. A.6a). In contrast, no GO terms were found to be significantly enriched in a majority of accessory or unique genomes, with the exception of very broad terms such as "cellular process" (Fig. A.6b-c). Overall, this functional analysis suggests that the core genomes of microbial pathogens are likely enriched for metabolic and translational functions, while non-core genes may draw from a wider variety of relatively niche functions.

Finally, an examination of individual orthogroups (OGs) as annotated by eggNOG-emapper reveals specific biosynthetic pathways consistently present in core genomes. 168 OGs were found in all 12 core genomes (Fig. 2.4a), while the most common OG among accessory genomes was found in 11 accessory genomes (Fig. 2.4b). A majority of these conserved core OGs were found to be essential for growth for *E. coli* in LB media (101/168, 60%) [23] (Dataset A.5). Functionally, core OGs were again dominated by translation/ribosomal genes (60/168, 36%) and also included a significant number of metabolic genes (54/168, 32%), many of which share metabolic pathways (Fig. 2.4c). Purine metabolism was strongly represented with 15 OGs conserved in all core genomes: *purD, purE, purF, purH, purM, purN* in IMP biosynthesis; *carA, carB, pyrB, pyrF* in UMP biosynthesis; *guaA, guaB* in GMP biosynthesis; *purA* in AMP biosynthesis; and *adk, gmk, upp* in salvage pathways. Other conserved OGs sharing biosynthetic pathways include *aroA, aroB, aroC, aroE* in chorismate biosynthesis; *coaBC, coaD, coaE* in coenzyme A biosynthesis, and *accB, accD, fabZ* in fatty acid biosynthesis.

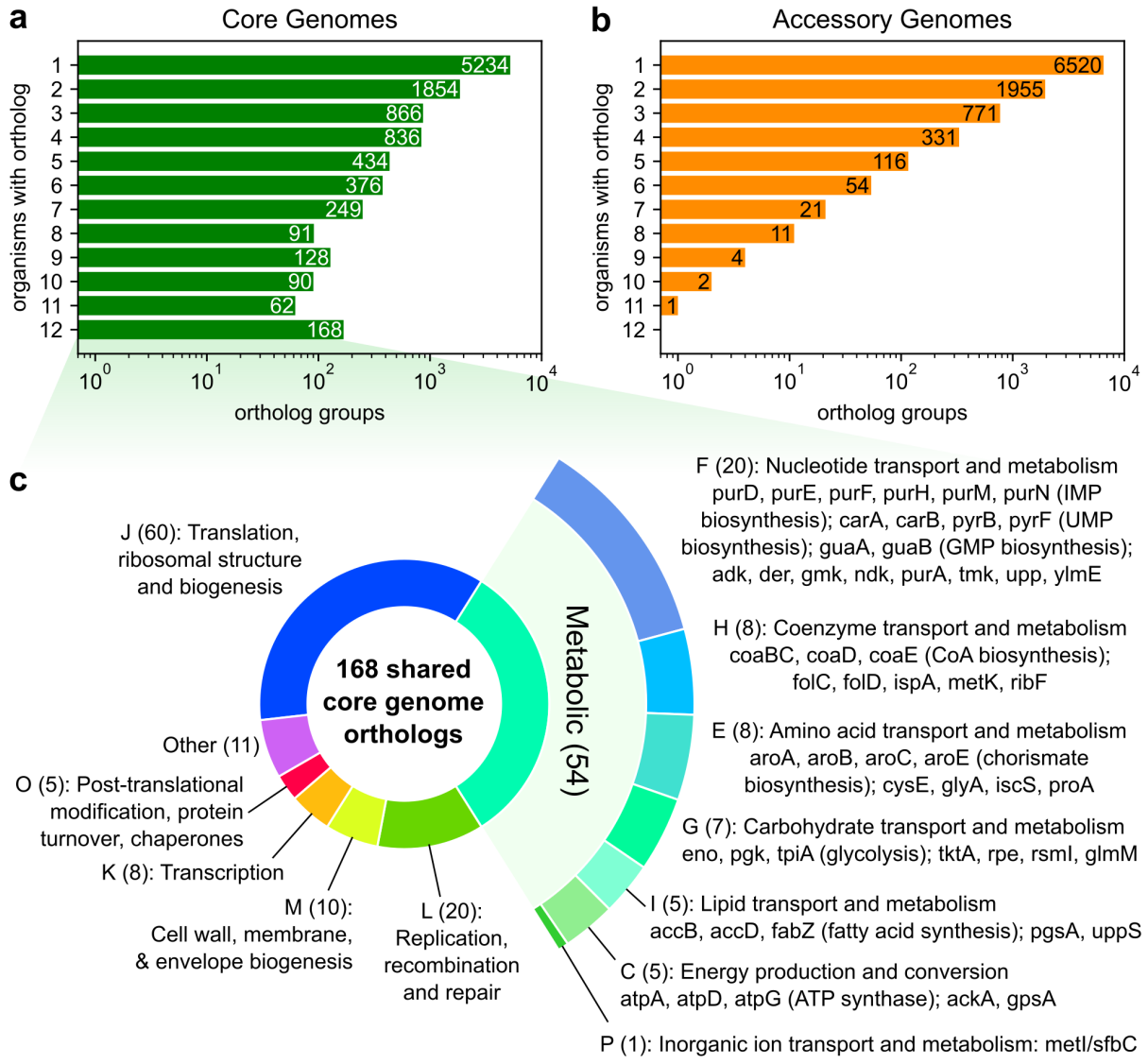


Figure 2.4. Distribution of shared genes in 12 core and accessory genomes. Number of shared genes versus frequency of observation across the (a) core genomes and (b) accessory genomes of 12 species. (c) Functional breakdown of the 168 genes observed in all 12 core genomes. Colors correspond to individual COG functional categories, which are labeled by the number of shared core genes annotated with the COG and COG definition. For metabolic COGs, individual genes and associated pathways are listed.

2.3.5 Genes conserved at the sequence level are enriched for translation-associated genes, while sources of core genome sequence diversity are functionally diverse

Moving to the resolution of individual variants to assess sequence-level genetic diversity, the frequency of each unique protein sequence associated with each core gene within each species' pangenome was computed. The entropies of these variant frequency distributions were computed as a measure of coding sequence diversity for each gene. Similarly, the entropies of analogous frequency distributions for gene-specific 5′ IG and 3′ IG variants were also computed, resulting in three "allelic entropy" measures for each gene (Fig. 2.5a, see Methods for entropy calculations). These measures allow for the quantification of a gene's overall sequence-level diversity, without requiring reference genomes or computationally expensive multiple sequence alignments for each sequence cluster that would be infeasible at the multi-pangenome scale. For each species' core genome, only limited correlation was observed between the level of sequence variability in a gene's coding sequence compared to flanking intergenic sequences; median Spearman correlation across the 12 species was 0.286 between coding and 5′ upstream allelic entropy, and 0.237 between the coding and 3′ downstream allelic entropy (Table A.5).

To identify the most and least sequence-diverse core genes by each feature type (coding, 5′ IG, or 3′ IG), the top and bottom 5% of core genes were identified after sorting by the corresponding allelic entropy measure and classified as "diverse" or "conserved", respectively. In the case of coding sequence diversity, since the metric is sensitive to gene length, the top and bottom 5% of genes were instead identified using quantile regression [24] to estimate the 5 and 95% allelic entropy percentiles as a function of gene length (Fig. A.7, Fig. A.8). Functional enrichment tests between COG functional groups and either the most or least conserved core genes per species revealed generally little association between any of the types of sequence diversity and gene function (Fig. 2.5b). Only COG J (translation, ribosomal structure and biogenesis) exhibited consistent enrichment among

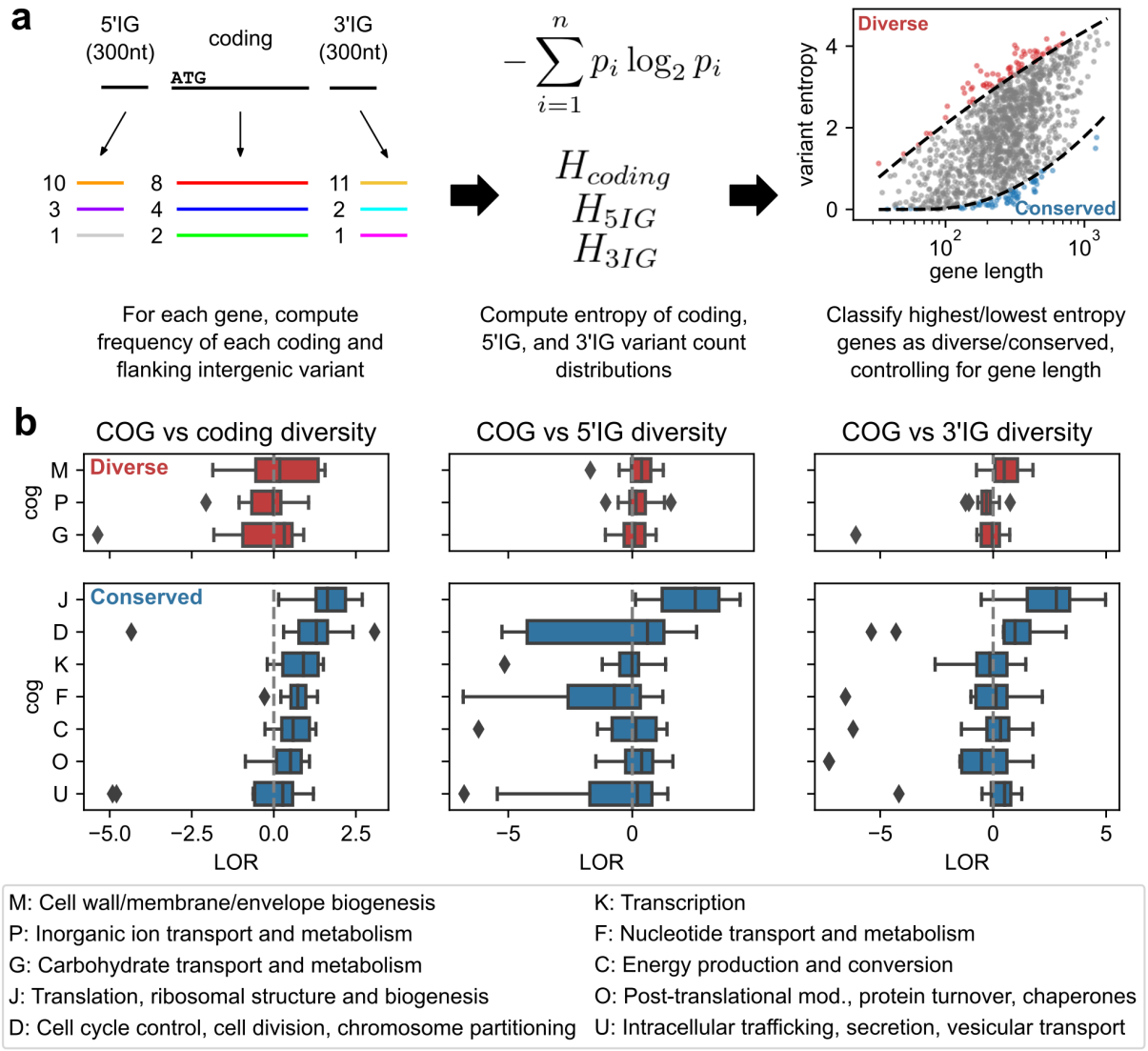


Figure 2.5. Functional enrichment in core genes versus sequence diversity in coding or flanking intergenic sequences. (a) Workflow for identifying genes with high or low sequence diversity. For a given gene and species, frequencies of individual coding, 3′ intergenic (3′ IG), and 5′ intergenic (5′ IG) variants were computed, and entropies of the three variant type-specific frequency distributions were computed as measures of sequence diversity. For a given entropy measure, genes in the top and bottom 5% were classified as "diverse" or "conserved"; in the case of coding sequence entropy, 5 and 95% percentiles as a function of gene length were used instead, estimated through quantile regression. (b) Functional enrichment in genes classified as most diverse or conserved by either coding, 5′ IG, or, 3′ IG sequence entropy. Only COGs with positive mean $\log_2$ odds ratio (LOR) across the 12 species for at least one entropy measure are shown.

the least sequence-diverse core genes, with mean LORs across the 12 species of 1.7, 2.4, and 2.4 among genes conserved by coding, 5′ IG, and 3′ IG allelic entropy, respectively, and statistically significant enrichment in 7/12 species for all three measures ($p < 7 * 10^{-5}$, FWER < 0.05 with Bonferroni correction). Additional weak biases towards other COGs were also observed for conserved or diverse genes by one or more feature types, though none were statistically significant in more than single species (Fig. 2.5b, Dataset A.4).

Finally, genes involved in PubMLST typing schemes for these species ranged from strongly conserved to highly diverse based on coding allelic entropy (Fig. A.9): all MLST genes were classified as core genes, and the most coding diverse gene in each scheme ranged from the 72nd to 99th percentile in the coding allelic entropy distribution of the corresponding species' core genome, while the least coding diverse gene ranged from the 4th to 36th percentile.

2.3.6 Position of variation in conserved core genes is domain-dependent, especially among aminoacyl-tRNA synthetases

Finally, for the highest resolution of pangenomic diversity, the position of sequence-level diversity in the pangenome was examined for 76 of the 168 OGs (here on referred to as just "genes") previously found to be in all 12 core genomes after filtering for those that could be richly annotated for domains (see Methods for gene selection process). For each species-specific set of protein sequence variants of a given gene, a multiple sequence alignment (MSA) was computed using MAFFT [25], from which the consensus sequence was annotated for domains using InterProScan [26]. The entropy at each position of the MSA was computed, and to evaluate a domain's variability relative to its parent gene, the mean MSA entropy across all positions spanned by the domain was computed and compared against the mean MSA entropies of all windows with the same length as the domain in the MSA, yielding the domain's entropy percentile with respect to the given

gene and species (Fig. 2.6a).

Across the 443 gene-domain pairs analyzed, 27 domains were identified to be mutation enriched with significantly elevated entropy consistently across the 12 species analyzed, and 61 domains were identified to be mutation depleted with significantly reduced entropy (Bootstrap test, FDR < 0.05, Benjamini-Hochberg correction) (Fig. 2.6b, Dataset A.6). Both the mutation enriched and mutation depleted domains are functionally diverse and are found in a wide range of genes (Fig. 2.6c, Fig. A.10), though with a bias towards domains related to aminoacyl-tRNA synthetases (AARSs); 26% of mutation enriched domains were related to AARS compared to 14% of the full set of domains analyzed (Fig. 2.7a). A survey of AARS-related domains finds that the extent of a domain's multispecies mutation enrichment is associated with function (Fig. 2.7b-c). Among the 9 AARSs analyzed (*alaS, aspS, cysS, ilesS, metG, pheS, serS, thrS, valS*), domains related to editing, anticodon binding, or tRNA binding were either mutation enriched or mutation neutral on average across the 12 species, while all but two non-editing catalytic domains were mutation depleted. Other domains (structural and/or of unknown function) were distributed over the full range from mutation depletion to enrichment.

Within individual AARS domain functional categories, variability in mutation enrichment was due primarily to gene differences (i.e. differences in catalytic domains between *metG* vs. *ileS*) and lesser so to annotation specificity (i.e. domain vs. superfamily annotation of a similar region) (Fig. 2.7c). Outliers due to gene include the FPG/IleRS-type Zinc finger domain and Rossman-like alpha/beta/alpha sandwich fold domain in *ileS*, the only catalytic domains to be mildly mutation enriched (compared to the catalytic domains of *serS, cysS, thrS,* and *metG* which are strongly mutation depleted); and the putative editing domain of *alaS*, which is mildly mutation depleted compared to the strongly mutation enriched editing domains of *thrS, ileS,* and *valS*.

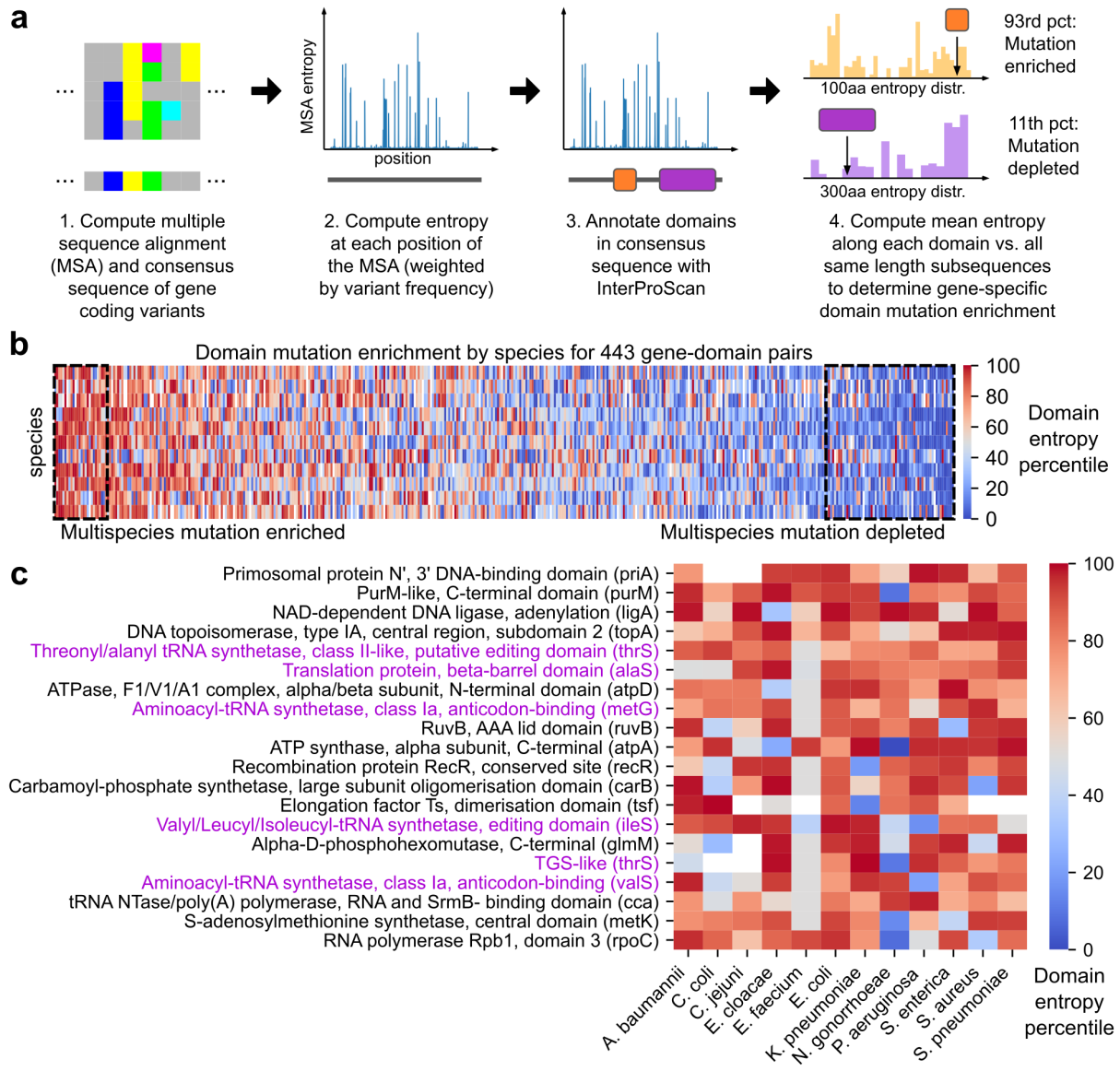
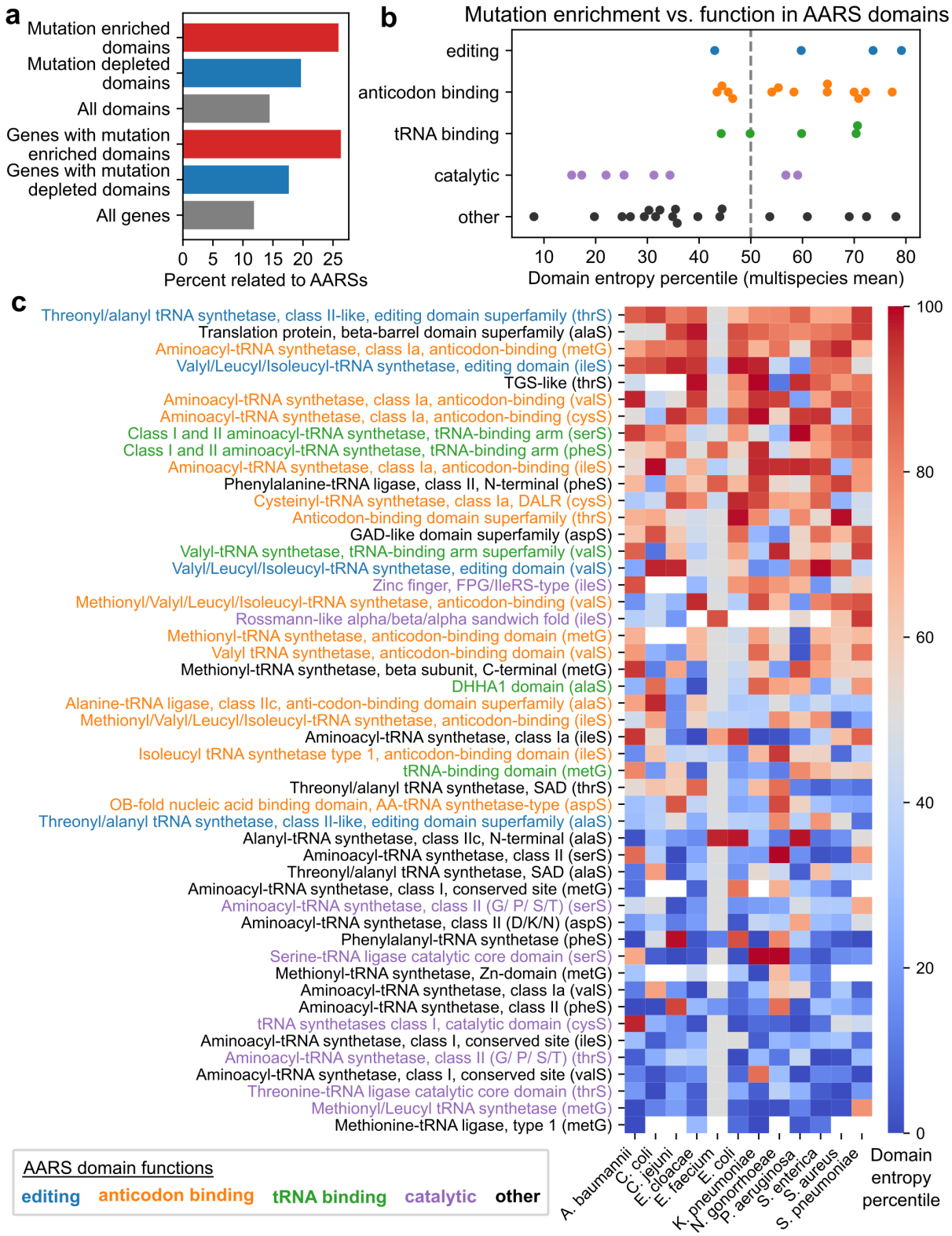


Figure 2.6. Mutation enrichment in protein domains from 76 genes present in 12 species' core genomes. (a) Workflow for computing the extent of mutation enrichment in a domain relative to the full protein from a set of coding variants. (b) Species-specific mutation enrichment for 443 gene-domain pairs, sorted by domain entropy percentile averaged across 12 species. Domains with statistically significant multispecies enrichment or depletion are boxed (Bootstrap test, FDR < 0.05, Benjamini-Hochberg correction). *E. faecium* is not shown, due to low variability attributable to initial subtype imbalances in the genome set. (c) Species-specific mutation enrichment for gene-domain pairs with significant multispecies mutation enrichment. Domains related to aminoacyl-tRNA synthetases are labeled purple. White cells correspond to domains that could not be annotated within the species' consensus sequence for the parent protein.

Figure 2.7. Species-specific mutation enrichment among aminoacyl-tRNA synthetase domains relative to corresponding full proteins. (a) Enrichment of aminoacyl-tRNA synthetase (AARS) related features among all domains with significant mutation enrichment or depletion. (b) Extent of mutation enrichment in AARS domains compared to function across 12 species. Each point corresponds to a single gene-domain pair, categorized by function based on InterPro descriptions. (c) Species-specific mutation enrichment for all AARS-associated gene-domain pairs, sorted by domain entropy percentile averaged across the 12 species. White cells correspond to domains that could not be annotated within the species' consensus sequence of the parent protein.


a
Mutation enriched domains
Mutation depleted domains
All domains
Genes with mutation enriched domains
Genes with mutation depleted domains
All genes
0 5 10 15 20 25
Percent related to AARSs
b
Mutation enrichment vs. function in AARS domains
editing
anticodon binding
tRNA binding
catalytic
other
10 20 30 40 50 60 70 80
Domain entropy percentile (multispecies mean)
c
Threonyl/alanyl tRNA synthetase, class II-like, editing domain superfamily (thrS)
Translation protein, beta-barrel domain superfamily (alaS)
Aminoacyl-tRNA synthetase, class Ia, anticodon-binding (metG)
Valyl/Leucyl/Isoleucyl-tRNA synthetase, editing domain (ileS)
TGS-like (thrS)
Aminoacyl-tRNA synthetase, class Ia, anticodon-binding (valS)
Aminoacyl-tRNA synthetase, class Ia, anticodon-binding (cysS)
Class I and II aminoacyl-tRNA synthetase, tRNA-binding arm (serS)
Class I and II aminoacyl-tRNA synthetase, tRNA-binding arm (pheS)
Aminoacyl-tRNA synthetase, class Ia, anticodon-binding (ileS)
Phenylalanine-tRNA ligase, class II, N-terminal (pheS)
Cysteinyl-tRNA synthetase, class Ia, DALR (cysS)
Anticodon-binding domain superfamily (thrS)
GAD-like domain superfamily (aspS)
Valyl-tRNA synthetase, tRNA-binding arm superfamily (valS)
Valyl/Leucyl/Isoleucyl-tRNA synthetase, editing domain (valS)
Zinc finger, FPG/IleRS-type (ileS)
Methionyl/Valyl/Leucyl/Isoleucyl-tRNA synthetase, anticodon-binding (valS)
Rossmann-like alpha/beta/alpha sandwich fold (ileS)
Methionyl-tRNA synthetase, anticodon-binding domain (metG)
Valyl tRNA synthetase, anticodon-binding domain (valS)
Methionyl-tRNA synthetase, beta subunit, C-terminal (metG)
DHHA1 domain (alaS)
Alanine-tRNA ligase, class IIc, anti-codon-binding domain superfamily (alaS)
Methionyl/Valyl/Leucyl/Isoleucyl-tRNA synthetase, anticodon-binding (ileS)
Aminoacyl-tRNA synthetase, class Ia (ileS)
Isoleucyl tRNA synthetase type 1, anticodon-binding domain (ileS)
tRNA-binding domain (metG)
Threonyl/alanyl tRNA synthetase, SAD (thrS)
OB-fold nucleic acid binding domain, AA-tRNA synthetase-type (aspS)
Threonyl/alanyl tRNA synthetase, class II-like, editing domain superfamily (alaS)
Alanyl-tRNA synthetase, class IIc, N-terminal (alaS)
Aminoacyl-tRNA synthetase, class II (serS)
Threonyl/alanyl tRNA synthetase, SAD (alaS)
Aminoacyl-tRNA synthetase, class I, conserved site (metG)
Aminoacyl-tRNA synthetase, class II (G/ P/ S/T) (serS)
Aminoacyl-tRNA synthetase, class II (D/K/N) (aspS)
Phenylalanyl-tRNA synthetase (pheS)
Serine-tRNA ligase catalytic core domain (serS)
Methionyl-tRNA synthetase, Zn-domain (metG)
Aminoacyl-tRNA synthetase, class Ia (valS)
Aminoacyl-tRNA synthetase, class II (pheS)
tRNA synthetases class I, catalytic domain (cysS)
Aminoacyl-tRNA synthetase, class I, conserved site (ileS)
Aminoacyl-tRNA synthetase, class II (G/ P/ S/T) (thrS)
Aminoacyl-tRNA synthetase, class I, conserved site (valS)
Threonine-tRNA ligase catalytic core domain (thrS)
Methionyl/Leucyl tRNA synthetase (metG)
Methionine-tRNA ligase, type 1 (metG)
100
80
60
40
20
0
Domain entropy percentile
A. baumannii
C. coli
C. jejuni
E. cloacae
E. faecium
E. coli
K. pneumoniae
N. gonorrhoeae
P. aeruginosa
S. enterica
S. aureus
S. pneumoniae
AARS domain functions
editing
anticodon binding
tRNA binding
catalytic
other

2.4 Discussion

Advances in sequencing technologies have rapidly expanded the scale of public genome collections, allowing the scope of analyses to grow from full genomes, to multiple genomes, and now towards multiple pangenomes for global comparisons of genetic diversity between species. However, though numerous studies have examined genetic and functional diversity present in individual pangenomes, relatively few have offered comparisons between multiple distinct pangenomes, especially at the resolution that single pangenome studies often explore. In this study, we present generalizable "comparative pangenomics" methods for contextualizing genetic diversity with function across multiple species and at multiple resolutions, from the shape of the pangenome overall to specific positions within individual genes.

At the overall pangenome level, we find that by balancing representation of MLST subtypes through undersampling, Heaps' Law more accurately predicts pangenome size as new genomes are introduced while yielding pangenome openness estimates with variances similar to that of estimates derived without balancing. The balanced openness estimates were frequently higher than those derived without balancing (9/12 cases), possibly due to the estimates being less biased towards any individual subtype which typically draws from a more limited gene pool than the species overall. Such a trend was observed in clade-specific analyses of the *E. coli* ST131 pangenome, where all but one clade was more closed than the combined population [27]. Furthermore, the estimates suggest that openness roughly follows phylogenetic placement, especially with all six Gammaproteobacteria species analyzed here having very similar openness values that are all higher than that of the other bacterial classes examined. This is mostly corroborated in previous comparative works, though the exact openness values differ from those calculated here. Park et.al. found four Gammaproteobacteria, *A. baumannii, E. coli, S. enterica,* and *P. aeruginosa*, to have similar openness values compared to three other species analyzed from different

phylogenetic classes [11], and Tettelin et.al. classified *E. coli* and *S. pneumoniae* as open and *S. aureus* as relatively closed [6]. It is possible that subtype balancing is responsible for the differences in exact openness values, and ultimately the results suggest that integrating subtype information into models of pangenome size may more accurately reflect the level of genetic diversity within the species at this scale.

Moving from overall pangenome shape to individual genes, an examination of gene frequency distributions reveals that a double power function can closely model such distributions ($R^2 > 99\%$ in 11/12 species) and provides a scalable method for dividing the pangenome into frequency categories core, accessory, and unique. This approach is similar to the core-shell-cloud division based on a triple exponential function described by Koonin and Wolf [20] and implemented in the GET_HOMOLOGUES pangenome pipeline [21], which was similarly derived based on examining functional forms that closely fit empirical distributions, albeit originally for smaller genome collections. Future analyses may examine which functional form offers closer and more stable fits at scales of thousands of genomes, and how they compare to more sophisticated approaches generalizable to more than three partitions, such as PPanGGOLiN's integration of both gene frequency and synteny conservation information [28], or micropan's use of binomial mixture models [29].

An analysis of gene function distributions across these frequency categories finds several functional categories to be consistently associated with frequency across most of the examined species. Translation/ribosomal genes, as well as a number of genes from specific metabolic categories were significantly enriched in nearly all core genomes examined, while those concerning more niche functions such as trafficking/secretion or defense mechanisms were significantly enriched in a majority of accessory genomes. These results are also partially corroborated in Park et.al., where translation genes were among the top 5 overrepresented functional categories in 3/7 core genomes, trafficking/secretion in 2/7 accessory genomes, and various metabolic categories also overrepresented in some core genomes [11]; differences may be attributed to a more restrictive reporting (only top

5 categories are shown rather than all statistically significant cases), as well as a different statistical setup resulting in the reporting of some categories (such as transcription- or replication-associated) as overrepresented in both the core and accessory genomes. Additionally, the enrichments found here, especially that of translation genes in core genomes, were recovered in more focused studies examining 1-3 species or genuses at a time, such as for *A. baumannii* [30], *Campylobacter* [31], *E. coli* [32], *E. faecium* [33], *N. gonorrhoeae* [34], *P. aeruginosa* [35, 36], and *S. aureus* [37]. Finally, an analysis of individual genes identified 168 genes in the core genome of all 12 species, which were predominantly genes essential in *E. coli* (60%) and follow a functional distribution similar to that of core genomes overall, composed primarily of translation (36%) and metabolic (32%) genes (especially in nucleotide metabolism). This functional breakdown strongly resembles that of the "minimal gene set" identified in 1996 by Mushegian and Koonin for three species in one of the earliest characterizations of a bacteria-wide conserved gene set [38]. The repeated observation of specific functional enrichments in both this work and others suggest that core and accessory genomes from a wide array of bacterial species may share a consistent structure regarding functional distribution.

At the level of individual variants, we find less consistency within and between species regarding sequence-level genetic diversity. Using entropy of variant distributions to quantify sequence-level diversity without reference genomes or computationally expensive multiple sequence alignments, we find that the level of variability within a core gene's coding sequence is only weakly correlated with that of its immediate $5'$ or $3'$ flanking intergenic region in all 12 species examined (Spearman correlation between $0.2 - 0.3$). Pangenome-wide disparities in variation between the coding and flanking intergenic regions of a gene have been previously observed at the gene level: at least 11% of E. coli core genes were found to exhibit "regulatory switching" between nonhomologous flanking intergenic regions [39], and 7% of S. aureus core genes were found adjacent to non-core intergenic sequences [40]. Furthermore, while translation/ribosomal genes were consistently overrepresented among

the genes most strongly conserved at the sequence-level, the functional distribution of core genes responsible for the most sequence-level variability differs significantly by species. Whereas the functional distribution of overall gene content may be relatively stable between species, this finer-grained, shorter-term genetic diversity appears to impact a much broader range of functions within and across different species.

At the highest resolution assessment of genetic diversity, applying multiple sequence alignment and domain annotation to shared core genes revealed that specific structural features are disproportionately more conserved or diverse than the remainder of their parent gene, consistently across multiple species. Domains from AARS genes especially tended to exhibit this tendency for multispecies mutation depletion or enrichment, and an AARS-specific analysis revealed that the level of mutation enrichment strongly followed domain function, with non-editing catalytic domains being consistently mutation depleted, while tRNA-binding, anticodon-binding, and editing domains tending to be mutation enriched. This finding of short-term, intraspecies divergence of AARSs being localized away from catalytic domains for multiple species is consistent with previous analyses examining longer-term, interspecies differences in AARSs. Comparisons between representative AARSs of different species have shown significant diversity in overall domain architecture between different species and AARS classes in general [41], but catalytic domains are observed to be most frequently conserved at this level [42].

Additionally, two exceptions were observed in the broader trends between AARS domain function and mutation enrichment. First, the catalytic domains in *ileS* were the only catalytic domains not to be mutation depleted. One potential explanation may be that mutations near the catalytic Rossman fold of *ileS* have been associated with mupirocin resistance in *S. aureus* [43], and we find the Rossman fold domain of *ileS* to be more mutation enriched in *S. pneumoniae* and *S. aureus* compared to naturally mupirocin-resistant *P. aeruginosa* [44]. Second, the editing domain of *alaS* is the only editing domain not to be mutation enriched, while the editing domains of *thrS* and *ileS*

are among the most significantly mutation enriched across all domains examined. This result may be interpreted as possible instances of amino acid-specific misaminoacylation being tolerated or even improving fitness under certain stressful conditions, as previously observed for specific amino acids and environments [45, 46]. For example, editing-deficient *ileS* increases the growth rate of *E. coli* under isoleucine starvation [47] and the loss of *thrS* editing may trigger responses against oxidative stress [48], while the loss of *alaS* fidelity is poorly tolerated in *E. coli* [49]. Altogether, this domain analysis offers a finer-grain contextualization of pangenome-scale genetic diversity, revealing broadly conserved patterns of how mutations are localized in conserved genes as well as exceptions that may be explained by specific environmental stresses.

Finally, we note that pangenome-scale analyses are always limited by the availability of high quality genome assemblies and will continue to improve as more sequences are published. Future development and application of these methods to larger genome collections will provide increasingly complete pictures regarding the full extent of genetic diversity within a species, as well as present new challenges in evaluating the completeness and evenness of represented subtypes. Furthermore, similar analyses of additional species are necessary to determine whether the patterns of genetic diversity observed here are also present more broadly across the bacterial domain beyond major human pathogens.

2.5 Conclusions

Overall, in developing efficient and generalizable methods for pangenome analysis, we find that each resolution of the pangenome reveals distinct aspects of the relationship between genetic and functional diversity across multiple species located across the phylogenetic tree. In increasing resolution, we find across 12 pathogenic species that pangenome openness is associated with phylogenetic placement, the distribution of gene functions in the core and accessory genome is conserved across species, short-term sequence variation

in core genomes impacts a functionally diverse range of genes, and certain protein domains are enriched for mutations consistently across multiple species in a function-dependent manner, especially among AARSs. Many of the conserved patterns of genetic diversity uncovered here are consistent with previous studies focused on individual species, and continued development of multi-scale comparative pangenomic techniques may further elucidate similarities in how different species adapt to their environmental niches and pressures.

2.6 Acknowledgements

All authors contributed to project Conceptualization. J.C.H. and J.M.M. contributed to Methodology and Software. J.C.H. conducted the Data Curation, Investigation, Formal Analysis, Validation, and Visualization of results. J.C.H. prepared the Original Draft and all authors were involved in Review and Editing. J.M.M. and B.O.P. contributed to Funding Acquisition and Project Administration, and provided Supervision and Resources. The authors read and approved the final manuscript.

This research was supported by a grant from the National Institute of Allergy and Infectious Diseases (U01-AI124316, awarded to J.M.M. and B.O.P.). This research was also supported by a grant from the National Institutes of Health (T32GM8806, awarded to J.C.H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 2 is a reprint of material published in: **Jason C Hyun**, Jonathan M Monk, Bernhard O Palsson. 2022. "Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity." *BMC Genomics* 23(1):7. The disseration author is the primary author.

2.7 References

[1] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, Suresh Poudel, and David W Ussery. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, 15(2):141–161, March 2015.

[2] Alice Maria Giani, Guido Roberto Gallo, Luca Gianfranceschi, and Giulio Formenti. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.*, 18:9–19, 2020.

[3] L Rouli, V Merhej, P-E Fournier, and D Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect.*, 7:72–85, September 2015.

[4] Duccio Medini, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. The microbial pan-genome. *Curr. Opin. Genet. Dev.*, 15(6):589–594, December 2005.

[5] Luis Carlos Guimarães, Jolanta Florczak-Wyspianska, Leandro Benevides de Jesus, Marcus Vinícius Canário Viana, Artur Silva, Rommel Thiago Jucá Ramos, Siomar de Castro Soares, and Siomar de Castro Soares. Inside the pan-genome - methods and software overview. *Curr. Genomics*, 16(4):245–252, August 2015.

[6] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11(5):472–477, October 2008.

[7] Stephen Wood, Karen Zhu, Defne Surujon, Federico Rosconi, Juan C Ortiz-Marquez, and Tim van Opijnen. A pangenomic perspective on the emergence, maintenance, and predictability of antibiotic resistance. In *The Pangenome*, pages 169–202. Springer International Publishing, Cham, 2020.

[8] Yeji Kim, Changdai Gu, Hyun Uk Kim, and Sang Yup Lee. Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotechnol.*, 63:54–62, June 2020.

[9] Charles J Norsigian, Xin Fang, Bernhard O Palsson, and Jonathan M Monk. Pangenome flux balance analysis toward panphenomes. In *The Pangenome*, pages 219–232. Springer International Publishing, Cham, 2020.

[10] G S Vernikos. A review of pangenome tools and recent studies. In *The Pangenome*, pages 89–112. Springer International Publishing, Cham, 2020.

[11] Sang-Cheol Park, Kihyun Lee, Yeong Ouk Kim, Sungho Won, and Jongsik Chun. Large-scale genomics reveals the genetic characteristics of seven species and importance of phylogenetic distance for estimating pan-genome size. *Front. Microbiol.*, 10:834, April 2019.

[12] Oleksandr M Maistrenko, Daniel R Mende, Mechthild Luetge, Falk Hildebrand, Thomas S B Schmidt, Simone S Li, João F Matias Rodrigues, Christian von Mering, Luis Pedro Coelho, Jaime Huerta-Cepas, Shinichi Sunagawa, and Peer Bork. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.*, 14(5):1247–1259, May 2020.

[13] Bo Segerman. The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Front. Cell. Infect. Microbiol.*, 2:116, September 2012.

[14] Sávio Souza Costa, Luís Carlos Guimarães, Artur Silva, Siomar Castro Soares, and Rafael Azevedo Baraúna. First steps in the analysis of prokaryotic pan-genomes. *Bioinform. Biol. Insights*, 14:1177932220938064, August 2020.

[15] Narendrakumar M Chaudhari, Anupam Gautam, Vinod Kumar Gupta, Gagneet Kaur, Chitra Dutta, and Sandip Paul. PanGFR-HM: A dynamic web resource for pan-genomic and functional profiling of human microbiome with comparative features. *Front. Microbiol.*, 9:2322, October 2018.

[16] Emanuele Bosi, Marco Fondi, Valerio Orlandini, Elena Perrin, Isabel Maida, Donatella de Pascale, Maria Luisa Tutino, Ermenegilda Parrilli, Angelina Lo Giudice, Alain Filloux, and Renato Fani. The pangenome of (antarctic) pseudoalteromonas bacteria: evolutionary and functional insights. *BMC Genomics*, 18(1), December 2017.

[17] Alice R Wattam, David Abraham, Oral Dalay, Terry L Disz, Timothy Driscoll, Joseph L Gabbard, Joseph J Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, Dustin Machi, Chunhong Mao, Eric K Nordberg, Robert Olson, Ross Overbeek, Gordon D Pusch, Maulik Shukla, Julie Schulman, Rick L Stevens, Daniel E Sullivan, Veronika Vonstein, Andrew Warren, Rebecca Will, Meredith J C Wilson, Hyun Seung Yoo, Chengdong Zhang, Yan Zhang, and Bruno W Sobral. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, 42(Database issue):D581–91, January 2014.

[18] Keith A Jolley and Martin C J Maiden. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11(1):595, December 2010.

[19] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[20] Eugene V Koonin and Yuri I Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, 36(21):6688–6719, December 2008.

[21] Bruno Contreras-Moreira and Pablo Vinuesa. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.*, 79(24):7696–7701, December 2013.

[22] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, August 2017.

[23] Emily C A Goodall, Ashley Robinson, Iain G Johnston, Sara Jabbari, Keith A Turner, Adam F Cunningham, Peter A Lund, Jeffrey A Cole, and Ian R Henderson. The essential genome of *Escherichia coli* K-12. *MBio*, 9(1), March 2018.

[24] Roger Koenker and Kevin F Hallock. Quantile regression. *J. Econ. Perspect.*, 15(4):143–156, November 2001.

[25] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, April 2013.

[26] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014.

[27] Arun Gonzales Decano and Tim Downing. An escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.*, 9(1):17394, November 2019.

[28] Guillaume Gautreau, Adelme Bazin, Mathieu Gachet, Rémi Planel, Laura Burlot, Mathieu Dubois, Amandine Perrin, Claudine Médigue, Alexandra Calteau, Stéphane Cruveiller, Catherine Matias, Christophe Ambroise, Eduardo P C Rocha, and David Vallenet. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Comput. Biol.*, 16(3):e1007732, March 2020.

[29] Lars Snipen and Kristian Hovde Liland. micropan: an r-package for microbial pan-genomics. *BMC Bioinformatics*, 16(1):79, March 2015.

[30] Fei Liu, Yuying Zhu, Yong Yi, Na Lu, Baoli Zhu, and Yongfei Hu. Comparative genomic analysis of acinetobacter baumannii clinical isolates reveals extensive genomic variation and diverse antibiotic resistance determinants. *BMC Genomics*, 15(1):1163, December 2014.

[31] Ying Zhang and Stefan M Sievert. Pan-genome analyses identify lineage- and niche-specific markers of evolution and adaptation in epsilonproteobacteria. *Front. Microbiol.*, 5:110, March 2014.

[32] Hsuan-Lin Her and Yu-Wei Wu. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the escherichia coli strains. *Bioinformatics*, 34(13):i89–i95, July 2018.

[33] Zhi Zhong, Lai-Yu Kwok, Qiangchuan Hou, Yaru Sun, Weicheng Li, Heping Zhang, and Zhihong Sun. Comparative genomic analysis revealed great plasticity and environmental adaptation of the genomes of enterococcus faecium. *BMC Genomics*, 20(1):602, July 2019.

[34] Qun-Feng Lu, De-Min Cao, Li-Li Su, Song-Bo Li, Guang-Bin Ye, Xiao-Ying Zhu, and Ju-Ping Wang. Genus-wide comparative genomics analysis of neisseria to identify new genes associated with pathogenicity and niche adaptation of neisseria pathogens. *Int. J. Genomics*, 2019:6015730, January 2019.

[35] Luca Freschi, Antony T Vincent, Julie Jeukens, Jean-Guillaume Emond-Rheault, Irena Kukavica-Ibrulj, Marie-Josée Dupont, Steve J Charette, Brian Boyle, and Roger C Levesque. The pseudomonas aeruginosa pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biol. Evol.*, 11(1):109–120, January 2019.

[36] Utkarsh Sood, Princy Hira, Roshan Kumar, Abhay Bajaj, Desiraju Lakshmi Narsimha Rao, Rup Lal, and Mallikarjun Shakarad. Comparative genomic analyses reveal core-genome-wide genes under positive selection and major regulatory hubs in outlier strains of pseudomonas aeruginosa. *Front. Microbiol.*, 10:53, February 2019.

[37] Emanuele Bosi, Jonathan M Monk, Ramy K Aziz, Marco Fondi, Victor Nizet, and Bernhard Ø Palsson. Comparative genome-scale modelling of staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.*, 113(26):E3801–9, June 2016.

[38] A R Mushegian and E V Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U. S. A.*, 93(19):10268–10273, September 1996.

[39] Yaara Oren, Mark B Smith, Nathan I Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z Ron, Jukka Corander, Harris H Wang, Eric J Alm, and Tal Pupko. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*, 111(45):16112–16117, November 2014.

[40] Harry A Thorpe, Sion C Bayliss, Samuel K Sheppard, and Edward J Feil. Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria. *Gigascience*, 7(4), April 2018.

[41] Y I Wolf, L Aravind, N V Grishin, and E V Koonin. Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals

a complex history of horizontal gene transfer events. *Genome Res.*, 9(8):689–710, August 1999.

[42] Patrick O'Donoghue and Zaida Luthey-Schulten. On the evolution of structure in aminoacyl-tRNA synthetases. *Microbiol. Mol. Biol. Rev.*, 67(4):550–573, December 2003.

[43] Martin Antonio, Neil McFerran, and Mark J Pallen. Mutations affecting the rossman fold of isoleucyl-tRNA synthetase are correlated with low-level mupirocin resistance in staphylococcus aureus. *Antimicrob. Agents Chemother.*, 46(2):438–442, February 2002.

[44] R Sutherland, R J Boon, K E Griffin, P J Masters, B Slocombe, and A R White. Antibacterial activity of mupirocin (pseudomonic acid), a new antibiotic for topical use. *Antimicrob. Agents Chemother.*, 27(4):495–498, April 1985.

[45] Kyle Mohler and Michael Ibba. Translational fidelity and mistranslation in the cellular response to stress. *Nat. Microbiol.*, 2(9):17117, August 2017.

[46] Tao Pan. Adaptive translation as a mechanism of stress response and adaptation. *Annu. Rev. Genet.*, 47(1):121–137, August 2013.

[47] V Pezo, D Metzgar, T L Hendrickson, W F Waas, S Hazebrouck, V Döring, P Marlière, P Schimmel, and V De Crécy-Lagard. Artificially ambiguous genetic code confers growth yield advantage. *Proc. Natl. Acad. Sci. U. S. A.*, 101(23):8593–8597, June 2004.

[48] Jiang Wu, Yongqiang Fan, and Jiqiang Ling. Mechanism of oxidant-induced mistranslation by threonyl-tRNA synthetase. *Nucleic Acids Res.*, 42(10):6523–6531, June 2014.

[49] Paul Kelly, Nicholas Backes, Kyle Mohler, Christopher Buser, Arundhati Kavoor, Jesse Rinehart, Gregory Phillips, and Michael Ibba. Alanyl-tRNA synthetase quality control prevents global dysregulation of the escherichia coli proteome. *MBio*, 10(6), December 2019.

Chapter 3

A machine learning approach for identifying antimicrobial resistance determinants in pangenomes

3.1 Abstract

The evolution of antimicrobial resistance (AMR) poses a persistent threat to global public health. Sequencing efforts have already yielded genome sequences for thousands of resistant microbial isolates and require robust computational tools to systematically elucidate the genetic basis for AMR. Here, we present a generalizable machine learning workflow for identifying genetic features driving AMR based on constructing reference strain-agnostic pangenomes and training random subspace ensembles (RSEs). This workflow was applied to the resistance profiles of 14 antimicrobials across three urgent threat pathogens encompassing 288 *Staphylococcus aureus*, 456 *Pseudomonas aeruginosa*, and 1,588 *Escherichia coli* genomes. We find that feature selection by RSE detects known AMR associations more reliably than common statistical tests and previous ensemble approaches, identifying a total of 45 known AMR-conferring genes and alleles across the three organisms, as well as 25 candidate associations backed by domain-level annotations. Furthermore, we find that results from the RSE approach are consistent with existing understanding of fluoroquinolone (FQ) resistance due to mutations in the main drug

targets, *gyrA* and *parC*, in all three organisms, and suggest the mutational landscape of those genes with respect to FQ resistance is simple. As larger datasets become available, we expect this approach to more reliably predict AMR determinants for a wider range of microbial pathogens.

3.2 Summary

Antimicrobial resistance remains a persistent threat to global public health, with 700,000 deaths each year attributable to resistant bacterial infections. The falling cost of genome sequencing offers an avenue for rapidly predicting and elucidating the resistance profiles of infectious isolates, which is necessary for the design of more effective antimicrobial therapies from existing drugs. As such, clinical surveillance programs have already yielded sequences for thousands of distinct, resistant strains of most major pathogens. Here, we have developed a workflow for training machine learning models capable of not just predicting resistance profiles from genome sequences, but also identifying the responsible genes. When applied to 14 drugs and three urgent threat pathogens (*Staphylococcus aureus*, *Pseudomonas aeruginosa*, and *Escherichia coli*), our approach outperformed common statistical methods for detecting gene-level associations, identifying a total of 45 known resistance-conferring genes, as well as 25 candidate genes potentially involved in new mechanisms of resistance. These results show that this method can generalize to other drugs and pathogens to predict and explain resistance profiles at the gene level.

3.3 Background

The emergence of antimicrobial resistance (AMR) remains a persistent problem in the treatment of bacterial infections. Since the discovery of penicillin in 1928, pathogens have developed resistance to almost all major antibiotics, often within a few years of their introduction [1, 2]. Advancements in sequencing technology have already yielded

hundreds to thousands of publicly-available genome sequences for each major bacterial pathogen [3], and analyzing this deluge of data will require robust analytic workflows to extract insights on the acquisition of resistance, its genetic basis, and the underlying molecular mechanisms.

AMR prediction models have already been developed from genome sequence collections of many pathogens, such as *Staphylococcus aureus* [3–5], *Mycobacterium tuberculosis* [4, 6, 7], *Salmonella* [8, 9], *Klebsiella pneumoniae* [10, 11], and *Neisseria gonorrhoeae* [12, 13]. However, these approaches are often designed to maximize accuracy in predicting AMR phenotypes, emphasizing their diagnostic capabilities over their capacity to uncover genetic mechanisms for resistance. Many such models are also based on the detection of genes from a curated set of known AMR determinants, rendering them difficult to generalize to different treatments or organisms and unsuitable for discovering novel genes or interactions that drive resistance. Continued reductions in sequencing costs will enable whole genome sequencing (WGS) of these pathogens at an increasing scale, and soon expand the capabilities of statistical approaches beyond the prediction of AMR phenotypes and towards the reliable identification of their genetic determinants. Thus, computational tools developed with both goals of predicting and explaining AMR phenotypes are sorely needed.

The identification of gene-AMR relationships falls under the umbrella of microbial genome-wide association studies (GWAS), which bear many similarities to human GWAS [14]. However, microbial GWAS methods are still under development as traditional human GWAS methods struggle to generalize to highly clonal datasets without complex adjustments for population structure [15–17]. We present here a simple, reference-agnostic, machine learning approach based on pangenomes for identifying AMR-associated genes using random subspace ensembles (RSEs), previously shown to improve the accuracy of support vector machines trained on high-dimensional biological imaging data [18]. In contrast to more commonly used bootstrapping ensembles, RSEs aggregate classifiers

trained on random subsamples of both the sample set (genomes with associated AMR phenotypes) and the feature set (genes and alleles identified in those genomes). We find this method to both accurately predict AMR phenotypes as well as detect known AMR determinants more reliably than well-known association tests or other ensemble strategies, and use this method to predict novel AMR-linked genes for multiple antimicrobials in *S. aureus*, *P. aeruginosa*, and *E. coli*.

3.4 Results

3.4.1 Selection of genetic features through pangenome construction

Sets of 288, 456, and 1,588 publicly-available genomes for *S. aureus*, *P. aeruginosa*, and *E. coli*, respectively, were downloaded from PATRIC after filtering by contig count and availability of experimental AMR phenotype data (Dataset B.1) [19]. To convert these genome assemblies into fixed feature sets amenable to machine learning, we first constructed a pangenome for each species by clustering open reading frames by protein coding sequence into putative genes and classifying each gene as either core (missing in 0-10 genomes), accessory (missing in >10 genomes, present in >10 genomes), or unique (present in 1-10 genomes). This 10-genome threshold was selected by identifying when the core genome size stabilizes as the threshold for core gene was gradually relaxed (Fig. B.1). We find that this reference genome-agnostic strategy for gene identification produces pangenomes consistent with previous pangenome studies in terms of core genome size, pangenome openness, and relationship between gene function and gene frequency (see Supplemental Discussion).

Furthermore, as the causative variation responsible for AMR often exists at the level of individual mutations, we identified and enumerated all observed unique amino acid sequence variants or "alleles" of each gene for each pangenome (Table B.1). Individual

genomes were encoded based on the presence or absence of core gene alleles and the presence or absence of non-core genes, yielding a binary matrix representation of genetic variation for each pangenome that is not biased towards a reference genome and encodes both fine-grained allelic variations in the core genome and broader variations in the dispensable genome.

3.4.2 Support vector machine ensembles identify known AMR genes more reliably than common statistical tests from the *S. aureus* pangenome

We focus initially on the *S. aureus* pangenome to test variations of a recently reported support vector machine (SVM) approach [6], and evaluate their capacity to detect genes from an a priori assembled list of known AMR determinants, compared to traditional statistical association tests. We examined six antibiotic treatments against *S. aureus* from distinct drug classes for which experimentally measured AMR phenotype data was available, binarized as Susceptible versus Resistant (Table B.2): ciprofloxacin (fluoroquinolone), clindamycin (lincosamide), erythromycin (macrolide), gentamicin (aminoglycoside), tetracycline (tetracycline), and trimethoprim/sulfamethoxazole (dihydrofolate reductase inhibitor/sulfonamide). For validation, known AMR genes were compiled from literature and the CARD database [20] (Dataset B.2), then aligned to the alleles in the pangenome using blastp to identify those that were present in our dataset. From an initial query of 915 sequences, we detected 32 unique genes associated with AMR for at least one of the six drugs, spanning 304 distinct alleles in the *S. aureus* pangenome (Table 3.1). For each allele, the log odds ratio (LOR) for resistance against the corresponding drug and its frequency of occurrence was plotted (Fig. B.2). Aside from rare alleles, we find that alleles of genes involved in either active protection of the drug target or inactivation of the drug molecule almost always have large, positive LORs. However, alleles of genes that may confer AMR via a target site mutation or efflux span a wider range of LORs; this may be

Table 3.1. Known AMR genes present in the *S. aureus* pangenome.

Antibiotic	Genes
ciprofloxacin	gyrA [21, 22], gyrB [21, 22], parC [21, 22], parE [21, 22], norA, norB, norC [23], sdrM [23], mdeA [23], qacA [23], mepA [23], mepR [23], mgrA [23], arlR [23], arlS [23]
clindamycin	ermA [24, 25], ermC [24, 25], lmrS [26], linA [24]
erythromycin	ermA [24, 25], ermC [24, 25], lmrS [26], msrA [27], mphC [24]
gentamicin	aph(3')-III [28, 29], ant(4')-I [28], aac(6')/aph(2") [28, 29], ant(6')-Ia [30]
tetracycline	tetK [31], tetM [31], tet38 [32], norB [23], mgrA [23]
trimethoprim	folA [33], dfrA [33], dfrG [34]
sulfamethoxazole	folP [33]

due to some site mutants not having mutations that directly confer AMR (in which case, large, negative LORs were observed), and some efflux pumps being individually insufficient for conferring clinically relevant levels of resistance.

To define a baseline level of performance for identifying AMR genes from phenotype associations, we examined how reliably common association tests can detect known AMR genes when sorting by p-value. Examining each drug individually, Fisher's exact and Cochran-Mantel-Haenszel (CMH) tests were applied between each *S. aureus* genetic feature and the AMR phenotypes for that drug, and features were ranked by p-value with fractional ranking to address ties. For the CMH tests, genomes were stratified into clusters generated by applying hierarchical clustering to the genetic feature matrix; the resulting clusters align closely to known subtypes and share similar AMR profiles (Fig. 3.1).

Using the same feature matrix and AMR phenotypes, two types of SVM ensemble were trained for each drug case to classify genomes as susceptible or resistant, composed of 500 SVMs each trained on either 1) a random sample of 80% of genomes and all features to yield a bootstrap ensemble similar to in [6], or 2) a random sample of 80% of genomes and 50% of features to yield a random subspace ensemble (RSE), an adjustment previously shown to improve the accuracy of SVMs trained on high-dimensional biological data (Fig.

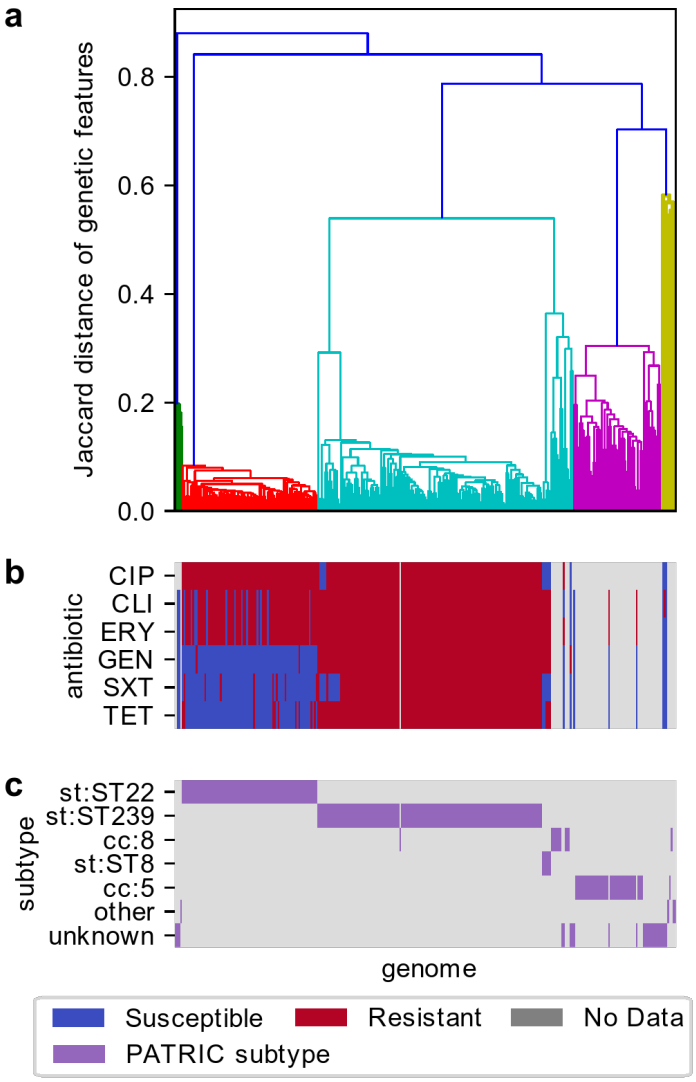


Figure 3.1. *S. aureus* genomes clustered by shared genetic content compared to known subtypes and antibiotic resistance patterns. (a) Genomes clustered using hierarchical clustering with average linkage, based on pairwise Jaccard distances between the sets of genetic features present in each genome. Clusters extracted from this hierarchy align well with (b) experimentally observed resistance patterns and (c) subtype annotations from PATRIC. Antibiotics shown are ciprofloxacin (CIP), clindamycin (CLI), erythromycin (ERY), gentamicin (GEN), sulfamethoxazole/trimethoprim (SXT), and tetracycline (TET).

3.2a) [18]. Analogously, features were ranked by feature weight (Fig. 3.2b).

We find that both SVM methods consistently identified more known AMR features within both the top 10 and top 50 hits than either statistical test (Fig. 3.2b). For instance, *ermC* and *lmrS* for clindamycin and erythromycin were only detectable by SVM methods, and *aac(6')-aph(2")* for gentamicin was detected as ranks 1 and 3 by the two SVM methods, compared to much higher ranks 84.5 and 148 by Fisher's exact and CMH tests, respectively. Additionally, the RSE approach allowed for known AMR genes to be detected at lower ranks compared to bootstrapping in several cases. Notably, *lmrS* for clindamycin and erythromycin was detected more than 70 ranks lower with this adjustment, putting *lmrS* within the top 50 hits in both cases with the random subspace approach. To control for phylogenetic distribution, SVM-RSE was also run with either oversampling (SVM-RSE-O) or undersampling (SVM-RSE-U) of genomes to balance the representation of the clusters used in CMH. However, the impact these controls have on the detection of individual known AMR genes is highly variable and does not suggest an improvement overall (Fig. 3.2b). For instance, SVM-RSE-O is the only approach able to identify *ermA* for clindamycin in the top 10, but loses a *gyrA* allele and two *parC* alleles for ciprofloxacin detected by SVM-RSE. Similarly, SVM-RSE-U improves the ranking of several known AMR genes already in the top 10 when compared to SVM-RSE, but loses *lmrS* from the top 50 for both clindamycin and erythromycin and loses *dfrG* for sulfamethoxazole/trimethoprim entirely. Finally, we note that Fisher's exact test was able to capture two tetracycline resistance genes (*tetM*, *tet38*) albeit at a high ranking of 83.5, while the other three approaches all identified only *tetK* as rank 1 and neither of the other two. However, Fisher's exact test suffered from an extremely high number of significant hits with Bonferroni correction to FWER < 0.05 (Table B.3), most likely due to strong lineage effects driving resistance in which detected features are often markers for a highly resistant subtype rather than true AMR genes [15]. The CMH test with inferred clusters resulted in a more reasonable amount of significant hits, though in the cases of

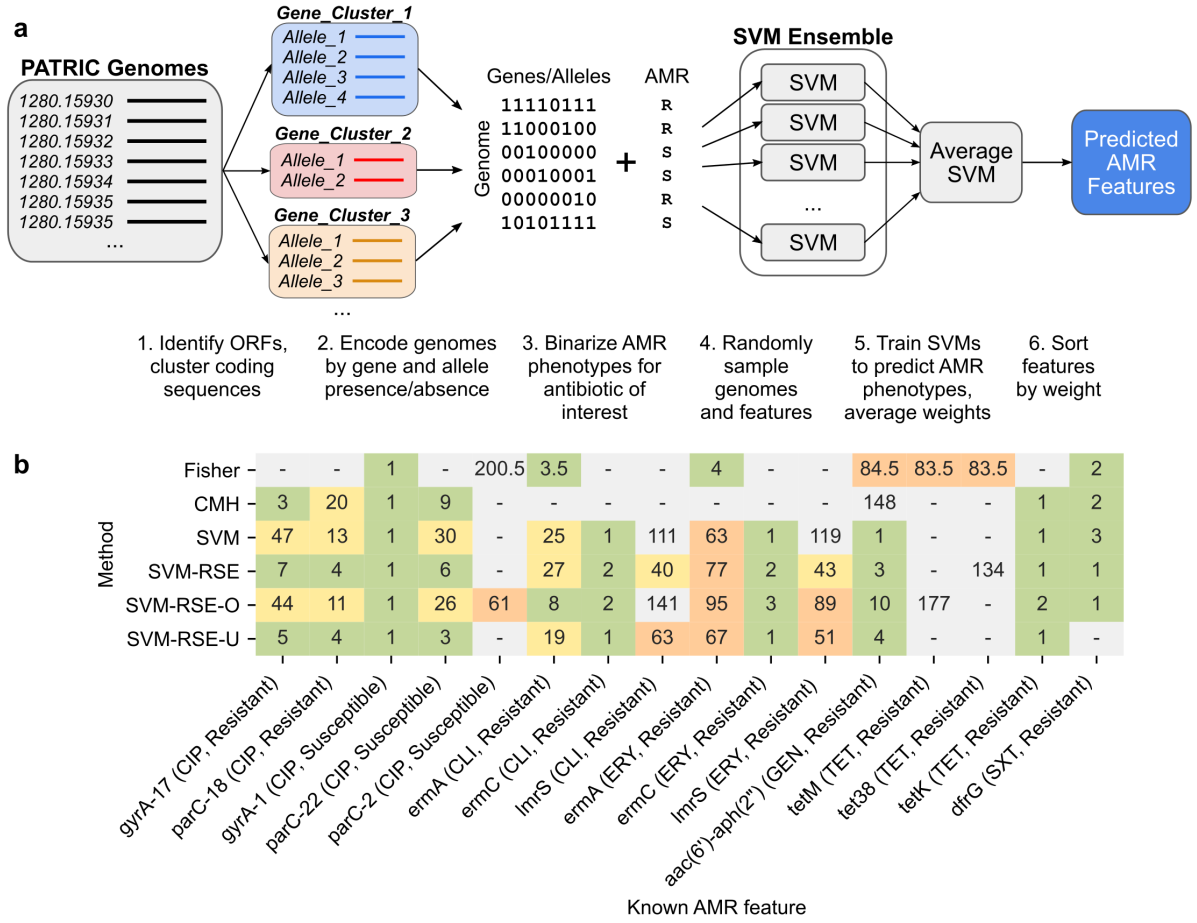


Figure 3.2. Comparison of SVM ensemble approaches and statistical tests for detecting AMR-conferring genes and alleles in *S. aureus*. (a) Workflow for SVM ensemble approaches. Beginning with genomes from PATRIC, open reading frames (ORFs) are identified and clustered by coding sequence to identify putative genes and alleles. Each genome is encoded based on the presence or absence of each gene and allele to capture genomic variation in the pangenome as a sparse binary matrix. Genomes and/or features of this matrix are randomly sampled 500 times and used to train SVMs to predict binary AMR phenotype for a single antibiotic from genotype. Weights for each feature are averaged across all models in the ensemble and used to rank features by association to AMR. (b) Associations between known AMR-conferring genomic features and AMR phenotype, as ranked by Fisher's exact test, Cochran-Mantel-Haenszel test, and four different SVM ensemble types (SVM: ensemble by bootstrapping genomes, SVM-RSE: bootstrapping genomes and features; "random subspace ensemble", SVM-RSE-O: SVM-RSE with oversampling to balance subtypes, SVM-RSE-U: SVM-RSE with undersampling to balance subtypes). Features were ranked either by p-value for statistical tests or by average feature weight for SVM ensembles. Fractional ranking was used for ties. Only features detected by at least one method are shown, colored by rank (green: in top 10, yellow: 11-50, orange: 51-100, gray: >100). Features shown are either genes or individual alleles (denoted as <*gene*>-#).

clindamycin and erythromycin, no genes were found significant even with a less stringent Benjamini-Hochberg correction to FDR < 0.05.

3.4.3 SVM random subspace ensembles identify known AMR genes in *S. aureus*, *P. aeruginosa*, and *E. coli* across multiple antibiotics

We applied our SVM-RSE approach to identify AMR genes in the larger *P. aeruginosa* and *E. coli* pangenomes, using the same core allele/non-core gene encoding of genomes and focusing on features positively associated with resistance. In addition to the six *S. aureus* cases, SVM-RSEs were trained to predict resistance for ten more species-drug cases: for amikacin, ceftazidime, levofloxacin, and meropenem in *P. aeruginosa*, and for amoxicillin/clavulanic acid, ceftazidime, ciprofloxacin, gentamicin, imipenem, and trimethoprim in *E. coli*, for a total of 16 species-drug cases (Table B.2, Fig. 3.3a).

By examining the highest weighted features in each SVM-RSE, this approach was able to identify known AMR genes among the top 50 hits in 15 out of the 16 cases, with more than half of those hits occurring within the top 10 and at least one known AMR gene found among the top 10 in 13 out of the 16 cases (Table 3.2). Only in the case of *P. aeruginosa*-amikacin were no such genes found, in which all aminoglycoside-inactivating enzymes in the pangenome identified by sequence homology had either modest LORs for resistance or were extremely rare (Table B.4). In total, 10, 7, and 28 unique AMR genetic features previously described in literature were detected and associated to the correct antibiotic for *S. aureus*, *P. aeruginosa*, and *E. coli*, respectively.

In terms of AMR phenotype prediction, in all 16 cases the individual SVMs of the corresponding SVM-RSE achieved much higher Matthews correlation coefficients (MCCs) on the test set when trained on the true data compared to data where AMR phenotypes were randomly permuted, suggesting that the associations learned were not due to noise (Fig. B.3). As a whole, the SVM-RSE achieved accuracies ranging from 79.3% to 99.5%,

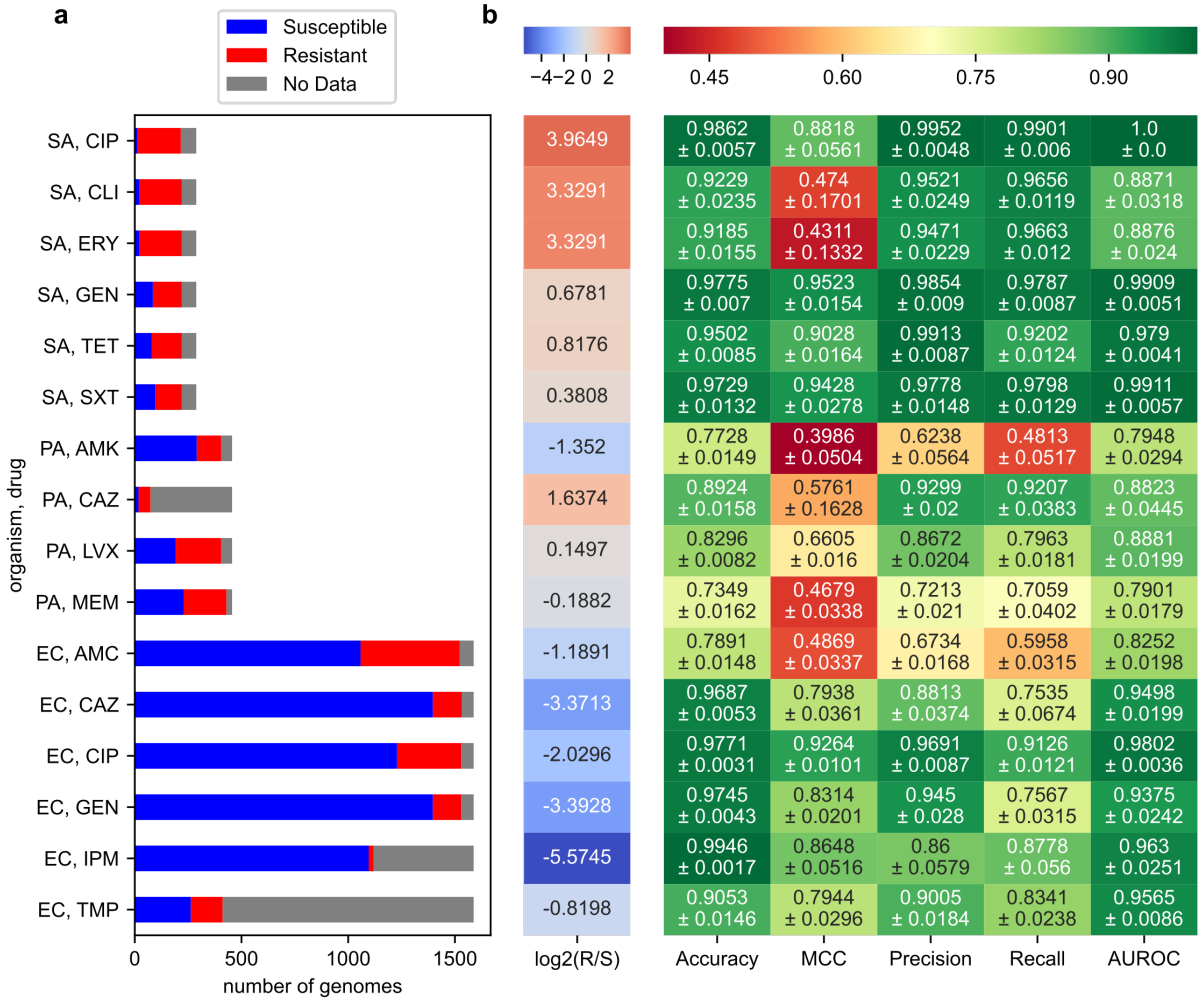


Figure 3.3. Predictive performance of SVM-RSE on 16 species-drug cases. (a) Distribution of AMR phenotypes for each case. Species examined are *S. aureus* (SA), *P. aeruginosa* (PA), and *E. coli* (EC). Antibiotics examined are ciprofloxacin (CIP), clindamycin (CLI), erythromycin (ERY), gentamicin (GEN), tetracycline (TET), sulfamethoxazole/trimethoprim (SXT), amikacin (AMK), ceftazidime (CAZ), levofloxacin (LVX), meropenem (MEM), amoxicillin/clavulanic acid (AMC), imipenem (IPM), and trimethoprim (TMP). (b) SVM-RSE performance metrics from 5-fold cross validation. Performance values shown are averages and standard errors from 5-fold cross validation. The left-most column "log2(R/S)" shows the extent of class imbalance, the $\log_2$ of the number of resistant genomes divided by the number of susceptible genomes.

Table 3.2. Known resistance-conferring genes found by SVM-RSE in *S. aureus*, *P. aeruginosa*, and *E. coli*.

Species	Drug	Feat.	Ranked 1-10	Ranked 11-50
S. aureus	CIP	2	gyrA [21, 22], parC [21, 22]	-
S. aureus	CLI	3	ermC [24, 25]	ermA [24, 25], lmrS [26]
S. aureus	ERY	2	ermC [24, 25]	lmrS [26]
S. aureus	GEN	1	aac(6')-aph(2") [28, 29]	-
S. aureus	SXT	1	dfrG [34]	-
S. aureus	TET	1	tetK [31]	-
P. aeruginosa	AMK	0	-	-
P. aeruginosa	CAZ	1	-	muxC [35]
P. aeruginosa	LVX	4	gyrA **(2)** [36], parC [36], oprD [37]	-
P. aeruginosa	MEM	2	oprD [37], bla_{OXA-2} [38]	-
E. coli	AMC	2	bla_{OXA-1} [39], bla_{TEM} [39]	-
E. coli	CAZ	4	bla_{CTX-M} [39], bla_{SHV} [39], bla_{CMY} [39]	bla_{OXA-1} [39]
E. coli	CIP	8	parC [40], gyrA **(4)** [40]	parC [40], parE [40], mdtA [41]
E. coli	GEN	6	aac(3)-IId/III [42, 43], ant(2")-Ia [43], ant(3")-Ia [42, 43]	aac(3)-VIa [42, 43], aac(6')-Ib [42, 43], ant(3")-Ia [42, 43]
E. coli	IPM	3	-	bla_{CTX-M} [39], mdtA [41], bla_{NDM} [39]
E. coli	TMP	5	dfrA1 [44], dfrA17 [44], dfrA14 [44]	qacE [45], dfrA12 [44]

For each species-drug pair, known AMR genes among the top 50 features detected by SVM-RSE are shown. Features referring to individual alleles of a gene are underlined. In the cases of *P. aeruginosa*-LVX and *E. coli*-CIP, two and four distinct resistant *gyrA* alleles were found in the top 10, respectively. In cases where a gene is mentioned in both the top 10 and rank 11-50 columns, multiple resistant alleles were detected at the different ranks. Antibiotics examined are ciprofloxacin (CIP), clindamycin (CLI), erythromycin (ERY), gentamicin (GEN), sulfamethoxazole/trimethoprim (SXT), tetracycline (TET), amikacin (AMK), ceftazidime (CAZ), levofloxacin (LVX), meropenem (MEM), amoxicillin/clavulanic acid (AMC), imipenem (IPM), and trimethoprim (TMP).

MCCs ranging from 0.394 to 0.952, and area under curves (AUCs) ranging 0.790 to 1.0 on the test set when averaged across 5-fold cross validation experiments (Fig. 3.3b, Fig. B.4). The average precision and recall ranged from 0.624 to 0.995 and 0.481 to 0.990, respectively (Fig. 3.3b). Across these metrics, 6 of 7 problematic cases were either 1) *P. aeruginosa* cases, which involve a notably larger genome than the other two species and thus present more challenging prediction problems, or 2) strongly class-imbalanced cases (*S. aureus*-clindamycin, *S. aureus*-erythromycin), though other strongly class-imbalanced cases performed well (*S. aureus*-ciprofloxacin, most *E. coli* cases). The final problematic case of *E. coli*-AMC is reasonably well balanced and may point to the challenge of predicting resistance for combination therapies of drugs with interacting mechanisms. Nonetheless, the models with the highest predictive performance were not necessarily those with the best detection of known AMR determinants and vice versa, which highlights the need for AMR prediction models to be evaluated both in terms of prediction performance and biological relevance.

Finally, we examined whether these top hits are robust to the core gene threshold used to determine which features of the pangenome are encoded at the gene level and which are encoded at the allele level. Compared to our original threshold of designating all genes missing in no more than 10 genomes as core genes, we also encoded each pangenome using two relative core gene thresholds: genes missing in no more than 2% or 10% of all genomes. After repeating the SVM-RSE workflow with these alternate pangenome representations, the set of the top 50 resistance-associated and top 50 susceptibility-associated was reasonably conserved between all thresholds. Across all species-drug cases, the average Jaccard similarity of selected features was 0.744 when comparing thresholds of 10 vs 10%, and 0.818 when comparing thresholds of 10 vs 2% (Fig. B.5).

3.4.4 Assessment of bias in features selected by SVM random subspace ensembles

We explored two potential biases in the features selected by SVM-RSE: whether there is a preference for genes with low versus high sequence variability, or for chromosomally versus plasmid encoded genes. First, as our approach encodes core genes at the allele level, we examined whether sequence variability impacts the selection of core gene alleles. Within each pangenome, the number of unique alleles ("allele count") for each core gene was computed, and for each species-drug case, the allele count distribution of the genes corresponding to selected core gene alleles was compared to that of all core genes (Fig. B.6a-b). Across all cases, there is a consistent but modest bias towards selecting core genes with higher sequence variability. However, even in the cases with the largest difference in mean allele count, the allele count distribution for selected core features is nearly indistinguishable from that of all core genes (Fig. B.6c-e).

Second, we examined whether SVM-RSE is capable of selecting non-core genes that are plasmid encoded. Contigs from all genome assemblies were identified as plasmid or chromosomal based on similarity to known plasmids on PLSDB [46], and genes with a majority of their alleles located on plasmid contigs were labeled as plasmid encoded genes. For each species-drug case, the number of selected non-core plasmid and chromosomal genes was compared to that of all non-core genes (Table B.5). SVM-RSE selected plasmid genes in 10/16 cases, with eight cases showing enrichment for plasmid genes. The six cases in which plasmid genes were not selected fall into two categories: 1) involving fluoroquinolones (ciprofloxacin, levofloxacin), for which resistance is primarily mediated by mutations in chromosomal genes *gyrA* and *parC*, or 2) involving *P. aeruginosa*, for which a relatively small fraction of non-core genes could be identified as plasmid encoded (1.3%, compared to 4.1% of *S. aureus* and 3.0% for *E. coli*). Overall, the SVM-RSE approach for identifying AMR-associated genetic features appears to be robust to sequence

variability when selecting core gene alleles, as well as sensitive to plasmid genes when selecting non-core genes.

3.4.5 SVM random subspace ensembles specify the space of *gyrA* and *parC* mutations associated with fluoroquinolone resistance

We examined resistance to fluoroquinolones (FQs) to compare AMR patterns in different species against the same drug class. For all three species, the SVM-RSE approach successfully detected at least one allele from both of the two established targets of FQs, *gyrA* and *parC*, within both the top 10 resistance-associated genetic features and the top 10 susceptibility-associated genetic features. All *gyrA* and *parC* alleles that the SVM-RSE associated with resistance bore substitutions previously known to confer resistance to FQs, while those that the model associated with susceptibility had no such known mutations (Table 3.3). Additionally, there were no uncharacterized mutations among the resistance-associated alleles that were not also present in a susceptibility-associated allele, which suggests that FQ resistance attributable to *gyrA* and *parC* may be limited to a narrow space of mutations, even across multiple species. Upon examining all *gyrA* and *parC* alleles, we find that resistance conferred by individual *gyrA* alleles is not dependent on a specific *parC* allele or vice versa; the LOR for resistance of any given *gyrA/parC* allele pair is not larger than that of the corresponding *gyrA* or *parC* alleles individually (Fig. B.7). By this metric, there were also no strong pairwise epistatic effects apparent between any of the top 10 resistance-associated hits in all three species (Fig. B.8).

3.4.6 Characterization of candidate AMR genes

In order to reduce the set of top resistance-associated genetic features to a smaller number of higher confidence AMR gene candidates, we filtered the top 10 hits for each species-drug case based on existing annotations and the level of sequence variability in each hit's assigned gene cluster (Methods). This yielded 25 candidate AMR-associated features

Table 3.3. Alleles of *gyrA* and *parC* associated with fluoroquinolone resistance detected by SVM-RSE.

Species	Feature	Res.	Sus.	Mutations
Alleles associated with fluoroquinolone resistance				
S. aureus	gyrA-18	119	0	**S84L** [47, 48], D402E, T457A, V598I, Δ815, T818E, Δ824, Δ825, E859V, E886D
S. aureus	parC-17	113	0	**S80F** [48], F410Y
P. aeruginosa	gyrA-4	82	2	**T83I** [49, 50]
P. aeruginosa	gyrA-15	18	1	**T83I** [49, 50], Δ909, Δ910
P. aeruginosa	parC-2	78	1	**S87L** [49, 50]
E. coli	gyrA-5	66	1	**S83L, D87N** [22, 40]
E. coli	gyrA-6	15	0	**S83L, D87N** [22, 40], D678E, A828S
E. coli	gyrA-9	157	2	**S83L, D87N** [22, 40], A828S
E. coli	gyrA-14	27	0	**S83L, D87N** [22, 40], D678E
E. coli	parC-6	46	2	**S80I** [22, 40]
Alleles associated with fluoroquinolone resistance				
S. aureus	gyrA-22	2	4	D402E, T457A, V598I, Δ815, T818E, Δ824, Δ825, E859V, E886D
S. aureus	parC-1	0	12	F410Y
P. aeruginosa	gyrA-1	23	115	-
P. aeruginosa	gyrA-6	4	39	Δ909, Δ910
P. aeruginosa	parC-1	52	137	-
E. coli	gyrA-0	3	637	D678E, A828S
E. coli	gyrA-1	1	152	D678E
E. coli	gyrA-22	2	179	-
E. coli	parC-1	1	250	-
E. coli	parC-2	7	475	D475E

Alleles of *gyrA* and *parC* among the top 10 hits associated with either resistance or susceptibility by SVM-RSE were characterized based on mutations relative to the corresponding gene in a reference genome for each species: NC_002745.2 for *S. aureus* (N315), NC_022516.2 for *P. aeruginosa* (PAO1), U00096.3 for *E. coli* (K12 MG1655). Allele-specific mutations are shown, with known resistance-conferring mutations shown in bold. Each allele's frequency among resistant (Res.) and susceptible (Sus.) genomes are shown.

which were further characterized by domain annotations from InterPro [51] (Table 3.4). In 9 out of the 13 core gene allele candidates, only a subset of the mutations present in the predicted AMR-conferring allele were actually enriched for resistance; those mutations were found to be present in known domains of their corresponding core gene and are strong candidates to be AMR-conferring (Fig. 3.4).

We note that a few of the predicted core gene alleles are of genes previously associated to resistance against the corresponding drug, if not necessarily in the target species or mechanistically established. For instance, an HflX-like protein is known to confer resistance in erythromycin in *Listeria monocytogenes* through ribosome recycling [52], and it is possible that the *hflX* gene discovered here may similarly confer resistance in *S. aureus*. In *Helicobacter pylori*, *oppD* was found to be significantly induced by gentamicin exposure [53]. For *ahpF*, overexpression is known to increase the minimum inhibitory concentration (MIC) for streptomycin (another aminoglycoside) [54], and has also been linked to increased multi-drug resistance through increased defense against oxidative stress in *E. coli* [55]. Finally, WP_000664727, probable *repL*, has been associated with the replication of staphylococcal resistance plasmids [56]. Sequences and annotations for these features, as well as for all top 50 hits for all species-drug cases are available in Dataset B.3 and B.4, respectively.

3.5 Discussion

As the number of publicly-available genome sequences for bacterial pathogens continues to grow, there is an increasing need to develop computational methods capable of discerning insights about antimicrobial resistance at scale. To leverage these highly diverse, genomic datasets, we have developed a reference strain-agnostic workflow based on pangenomes for building robust machine learning models capable of predicting AMR phenotypes as well as identifying their genetic determinants. Our SVM-RSE approach

Table 3.4. Novel resistance-conferring gene candidates predicted by SVM-RSE.

Predicted AMR-conferring core gene alleles

Species	Drug	Gene	R/S	LOR	AMR mutations	Mutation location(s)
S. aureus	ERY	hflX	135/0	8.8	Wildtype	-
S. aureus	GEN	SA_RS03845	134/0	13.5	S409N	ABC transporter-like domain
S. aureus	GEN	metS	134/0	13.5	T506N, E541K	Anticodon-binding domain
S. aureus	GEN	oppD	134/0	13.5	S68N, N132K	ABC transporter-like domain
S. aureus	GEN	comGD	134/0	13.5	D126Y	ComG operon protein 4 family (non-cytoplasmic)
S. aureus	GEN	ahpF	134/0	13.5	E38D, S44T, N112K, S422N, K448N	Thioredoxin-like domain superfamily; FAD/NAD(P) binding domain
S. aureus	TET	secE	131/2	8.7	G60R	C-terminus
S. aureus	TET	SA_RS11525	131/2	8.7	H127Y	-
S. aureus	TET	SA_RS10745	130/2	8.5	K641Q	RNA-binding domain S1
S. aureus	TET	kdpB2	130/2	8.5	P26L	P-type ATPase transmembrane domain superfamily
P. aeruginosa	CAZ	PA5359	48/1	6.3	DEL 1-24	N-terminal signal peptide
P. aeruginosa	CAZ	PA1414	48/1	6.3	DEL 1-33	N-terminus
P. aeruginosa	CAZ	PA1942	48/1	6.3	DEL 1-32	N-terminus

Predicted AMR-conferring accessory genes

Species	Drug	Accession	R/S	LOR	Predicted protein/features
S. aureus	CLI	WP_000664727	71/5	0.7	Plasmid replication protein, RepL
S. aureus	GEN	WP_000134308	134/1	11.6	Acyl-CoA N-acyltransferase, GNAT domain
S. aureus	TET	WP_031824444	123/2	7.8	Replication initiation factor
E. coli	AMC	WP_097223430	26/5	3.5	Bacterial toxin RNase RnlA/LsoA
E. coli	AMC	WP_000710826	26/5	3.5	Antitoxin RnlB/LsoB
E. coli	AMC	WP_000774834	25/11	2.4	Plasmid stability protein StbB
E. coli	CAZ	WP_001620093	13/33	2.1	NagB/RpiA transferase-like, DeoR-type HTH domain, DeoR C-terminal sensor domain
E. coli	CAZ	WP_000243817	82/15	7.1	RmlC-like cupin fold metalloprotein, WbuC family
E. coli	CIP	WP_001304218	262/386	3.9	Nucleoside triphosphate hydrolase, AAA domain
E. coli	GEN	WP_001330846	44/1	8.5	Transmembrane protein
E. coli	IMP	WP_001310177	2/25	2.0	PyrBI operon leader peptide
E. coli	TMP	WP_000082530	59/9	4.1	Mercury transport protein MerC

Selected AMR-conferring core gene alleles and accessory genes predicted by SVM-RSE, for *S. aureus*, *P. aeruginosa*, and *E. coli*. For core gene alleles, genes names and mutations are defined relative to the reference genomes N315 (NC_002745.2) for *S. aureus*, PAO1 (NC_002516.2) for *P. aeruginosa*, and K12 MG1655 (U00096.3) for *E. coli*. The number of resistant (R) vs. susceptible (S) genomes are shown for each feature. Log_2 odds ratios (LORs) were computed using weighted pseudocounts to account for zeroes in the contingency table (Methods). Protein features and domains were annotated with either InterPro (for core gene alleles) or InterProScan (for accessory genes).

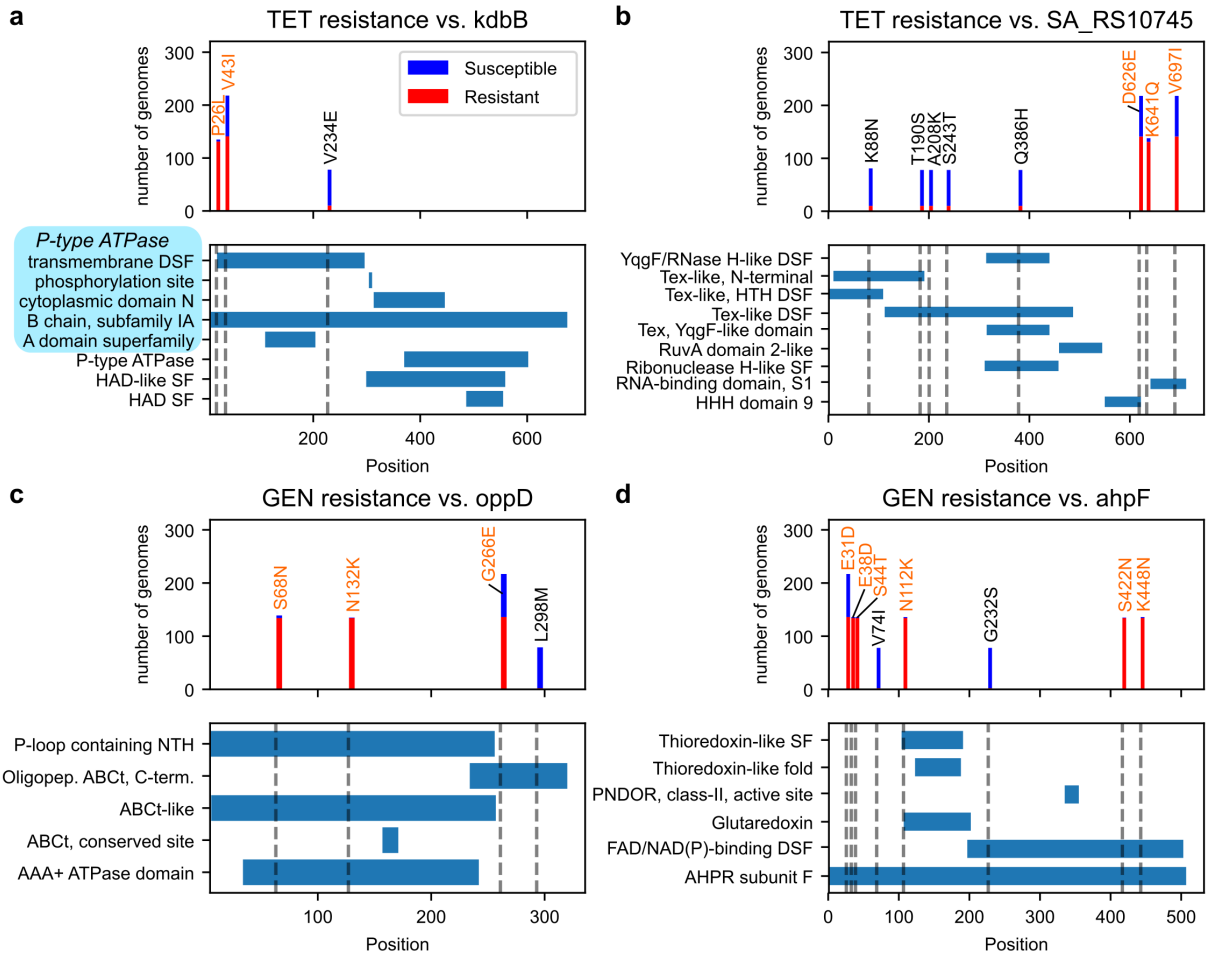


Figure 3.4. Characterization of mutations in four predicted AMR-conferring alleles in *S. aureus*. For each of the predicted AMR-associated genes (a) *kdbB*, (b) SA_RS10745, (c) *oppD* and (d) *ahpF*, the AMR phenotype distributions and locations relative to InterPro structural domains are shown for individual mutations. Mutations in the predicted AMR-associated allele are in orange, while all other mutations observed for that gene are in black (only mutations in at least 5 genomes are shown). For *kdbB*, the first five annotations in light blue are associated with P-type ATPase. Abbreviations include superfamily (SF), domain superfamily (DSF), nucleoside triphosphate hydrolase (NTH), ATP-binding cassette transporter (ABCt), pyridine nucleotide-diphosphate oxidoreductase (PNDOR), and alkyl hydroperoxide reductase (AHPR), in addition to those used in InterPro annotations.

was able to detect known resistance genes in three microbial pathogens (*S. aureus*, *P. aeruginosa*, and *E. coli*) more reliably than common association tests, while achieving prediction accuracies competitive with previous machine learning approaches.

Three pangenomes were constructed from 288 *S. aureus*, 456 *P. aeruginosa*, and 1,588 *E. coli genomes*, and the genetic diversity observed in each species is consistent with what was previously known of each pathogen. Upon integration of AMR profiling data, we found that our SVM-RSE approach effectively identifies established resistance determinants. SVM-RSE detected twice as many known AMR genes than both Fisher's exact and CMH tests for *S. aureus*, and was able to detect at least one known AMR gene in 15 of 16 species-drug cases, spanning a total of 45 known AMR associations identified across all three pathogens. Though none of the methods were comprehensive in their detection of all known AMR genes in the pangenome, the SVM-RSE appears to be the most reliable at detecting those genes for a diverse array of drug classes. We suspect that the success of this approach may be attributable to the following properties: 1) SVMs by design are capable of capturing structure among multiple features, opposed to independent, bivariate association tests, 2) using an ensemble trained on random genome subsets can more robustly determine important features when the feature set is much larger than the sample set, 3) subsampling features introduces training cases where resistance must be learned without the dominant AMR determinant, which often washes out signal from weaker determinants [3, 6], and 4) genes selected by SVM-RSE are neither biased by their extent of sequence variability nor by whether they are plasmid or chromosomally encoded.

The differences in detection rates between cases are partially due to the properties of their corresponding datasets. Generally, more known AMR genes were detected when both a large number of resistant and susceptible genomes were available; the difficult case of *P. aeruginosa*-ceftazidime had only 74 AMR profiles, and cases from the larger *E. coli* dataset typically performed better, with the exception of *E. coli*-imipenem in which only 23 genomes were resistant. In the third problematic case, *P. aeruginosa*-amikacin,

AMR profiles were well balanced, but known AMR-conferring genes were rare and/or had modest LORs for resistance, resulting in a more challenging feature selection problem. We also note that while benchmarking with *S. aureus* genomes, the model performed equally well even with aggressive undersampling to evenly represent different lineages. This suggests that genetically "redundant" genomes in a pangenome may be uninformative with respect to AMR. Finally, in all cases the prediction performances of both individual SVMs and SVM ensembles were high and comparable to previous machine learning approaches, independent of their ability to detect known AMR genes. This result comes as a warning that the raw performance of an AMR-prediction model may have little to do with its capacity to learn real AMR mechanisms.

In a deeper analysis of FQ resistance, we found that the top *gyrA* and *parC* alleles associated with resistance or susceptibility by SVM-RSE segregate perfectly by the presence or absence of known AMR-conferring mutations. The top resistance alleles also bore no uncharacterized mutations that were not also present in susceptible alleles, and no notable epistatic interactions between *gyrA* and *parC* allele pairs or any other pairs of predicted AMR-conferring features could be found. It is possible that the mutational landscape for FQ resistance may be relatively smooth and simple, and FQ resistance may be reliably predicted with simpler techniques; however, such hypotheses will be challenging to validate without more detailed measures of resistance beyond binary AMR phenotypes, such as minimum inhibitory concentrations. Extending this analysis to other predicted hits for all antibiotics identified 25 candidate AMR-conferring genetic features, of which several have evidence in other organisms to be involved in antibiotic-related responses, if not directly contributing to resistance.

Ultimately, by shifting the focus of evaluation from prediction accuracy to biological relevance, our framework more honestly expresses the level of confidence one may have in the generalizability of a machine-learning approach. We find that at the current scale of pathogen sequencing and profiling, our workflow is well-suited for not just predicting AMR

profiles, but also identifying genetic features known to confer resistance. The inherent flexibility of this approach opens it up to many improvements to expand the range of biological phenomena the models may draw upon to explain AMR; the incorporation of non-coding genetic features, integration of annotations into the learning process, or implementation of more sophisticated resampling and model aggregation strategies are just a few potential extensions of this work. The continued development of the techniques developed here may eventually be used to systematically extract confident explanations of resistance from pangenomic datasets to robustly inform responses to the AMR threat.

3.6 Acknowledgements

J.C.H., E.S.K., J.M.M., and B.O.P. conceived and designed the study. J.C.H. conducted all analysis, with contributions from J.M.M.; J.M.M. performed the pangenome construction pipeline. J.C.H., E.S.K., J.M.M., and B.O.P. provided study oversight, wrote the manuscript, and edited the manuscript. J.M.M. and B.O.P. managed the study. All authors reviewed and approved the final manuscript. We also thank Dr. Shankar Subramanian for helpful commentary on the evaluation of known resistance genes.

This research was supported by a grant from the National Institute of Allergy and Infectious Diseases (U01-AI124316, awarded to J.M.M. and B.O.P.). This research was also supported by a grant from the National Institutes of Health (T32GM8806, awarded to J.C.H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 3 is a reprint of material published in: **Jason C Hyun**, Erol S Kavvas, Jonathan M Monk, Bernhard O Palsson. 2020. "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens." *PLoS computational biology* 16(3):e1007608. The disseration author is the primary author.

3.7 References

[1] C Lee Ventola. The antibiotic resistance crisis: part 1: causes and threats. *P T*, 40(4):277–283, April 2015.

[2] Kai Kupferschmidt. Resistance fighters. *Science*, 352(6287):758–761, May 2016.

[3] James J Davis, Sébastien Boisvert, Thomas Brettin, Ronald W Kenyon, Chunhong Mao, Robert Olson, Ross Overbeek, John Santerre, Maulik Shukla, Alice R Wattam, Rebecca Will, Fangfang Xia, and Rick Stevens. Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.*, 6:27930, June 2016.

[4] Phelim Bradley, N Claire Gordon, Timothy M Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J Pankhurst, Luke Anson, Mariateresa de Cesare, Paolo Piazza, Antonina A Votintseva, Tanya Golubchik, Daniel J Wilson, David H Wyllie, Roland Diel, Stefan Niemann, Silke Feuerriegel, Thomas A Kohl, Nazir Ismail, Shaheed V Omar, E Grace Smith, David Buck, Gil McVean, A Sarah Walker, Tim E A Peto, Derrick W Crook, and Zamin Iqbal. Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nat. Commun.*, 6(1), December 2015.

[5] N C Gordon, J R Price, K Cole, R Everitt, M Morgan, J Finney, A M Kearns, B Pichon, B Young, D J Wilson, M J Llewelyn, J Paul, T E A Peto, D W Crook, A S Walker, and T Golubchik. Prediction of staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *J. Clin. Microbiol.*, 52(4):1182–1191, April 2014.

[6] Erol S Kavvas, Edward Catoiu, Nathan Mih, James T Yurkovich, Yara Seif, Nicholas Dillon, David Heckmann, Amitesh Anand, Laurence Yang, Victor Nizet, Jonathan M Monk, and Bernhard O Palsson. Machine learning and structural analysis of mycobac-

terium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.*, 9(1):4306, October 2018.

[7] Alexandre Drouin, Sébastien Giguère, Maxime Déraspe, Mario Marchand, Michael Tyers, Vivian G Loo, Anne-Marie Bourgault, François Laviolette, and Jacques Corbeil. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17(1):754, September 2016.

[8] Marcus Nguyen, S Wesley Long, Patrick F McDermott, Randall J Olsen, Robert Olson, Rick L Stevens, Gregory H Tyson, Shaohua Zhao, and James J Davis. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.*, 57(2), February 2019.

[9] Patrick F McDermott, Gregory H Tyson, Claudine Kabera, Yuansha Chen, Cong Li, Jason P Folster, Sherry L Ayers, Claudia Lam, Heather P Tate, and Shaohua Zhao. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal salmonella. *Antimicrob. Agents Chemother.*, 60(9):5515–5520, September 2016.

[10] Marcus Nguyen, Thomas Brettin, S Wesley Long, James M Musser, Randall J Olsen, Robert Olson, Maulik Shukla, Rick L Stevens, Fangfang Xia, Hyunseung Yoo, and James J Davis. Developing an in silico minimum inhibitory concentration panel test for klebsiella pneumoniae. *Sci. Rep.*, 8(1):421, January 2018.

[11] N Stoesser, E M Batty, D W Eyre, M Morgan, D H Wyllie, C Del Ojo Elias, J R Johnson, A S Walker, T E A Peto, and D W Crook. Predicting antimicrobial susceptibilities for escherichia coli and klebsiella pneumoniae isolates using whole genomic sequence data. *J. Antimicrob. Chemother.*, 68(10):2234–2244, October 2013.

[12] David W Eyre, Dilrini De Silva, Kevin Cole, Joanna Peters, Michelle J Cole, Yonatan H Grad, Walter Demczuk, Irene Martin, Michael R Mulvey, Derrick W Crook, A Sarah

Walker, Tim E A Peto, and John Paul. WGS to predict antibiotic MICs for neisseria gonorrhoeae. *J. Antimicrob. Chemother.*, 72(7):1937–1947, July 2017.

[13] Yonatan H Grad, Simon R Harris, Robert D Kirkcaldy, Anna G Green, Debora S Marks, Stephen D Bentley, David Trees, and Marc Lipsitch. Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the united states, 2000–2013. *J. Infect. Dis.*, 214(10):1579–1587, November 2016.

[14] Robert A Power, Julian Parkhill, and Tulio de Oliveira. Microbial genome-wide association studies: lessons from human GWAS. *Nat. Rev. Genet.*, 18(1):41–50, January 2017.

[15] Sarah G Earle, Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N Claire Gordon, Timothy M Walker, Chris C A Spencer, Zamin Iqbal, David A Clifton, Katie L Hopkins, Neil Woodford, E Grace Smith, Nazir Ismail, Martin J Llewelyn, Tim E Peto, Derrick W Crook, Gil McVean, A Sarah Walker, and Daniel J Wilson. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nat. Microbiol.*, 1(5), May 2016.

[16] Peter E Chen and B Jesse Shapiro. The advent of genome-wide association studies for bacteria. *Curr. Opin. Microbiol.*, 25:17–24, June 2015.

[17] Caitlin Collins and Xavier Didelot. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.*, 14(2):e1005958, February 2018.

[18] Alberto Bertoni, Raffaella Folgieri, and Giorgio Valentini. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing*, 63:535–539, January 2005.

[19] Alice R Wattam, David Abraham, Oral Dalay, Terry L Disz, Timothy Driscoll, Joseph L Gabbard, Joseph J Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, Dustin Machi, Chunhong Mao, Eric K Nordberg, Robert Olson, Ross Overbeek, Gordon D Pusch, Maulik Shukla, Julie Schulman, Rick L Stevens, Daniel E Sullivan, Veronika Vonstein, Andrew Warren, Rebecca Will, Meredith J C Wilson, Hyun Seung Yoo, Chengdong Zhang, Yan Zhang, and Bruno W Sobral. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, 42(Database issue):D581–91, January 2014.

[20] Baofeng Jia, Amogelang R Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K Tsang, Briony A Lago, Biren M Dave, Sheldon Pereira, Arjun N Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E Williams, Jonathan G Frye, Tariq Elsayegh, Daim Sardar, Erin L Westman, Andrew C Pawlowski, Timothy A Johnson, Fiona S L Brinkman, Gerard D Wright, and Andrew G McArthur. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 45(D1):D566–D573, January 2017.

[21] George A Jacoby. Mechanisms of resistance to quinolones. *Clin. Infect. Dis.*, 41 Suppl 2(Supplement_2):S120–6, July 2005.

[22] Anna Fàbrega, Sergi Madurga, Ernest Giralt, and Jordi Vila. Mechanism of action of and resistance to quinolones. *Microb. Biotechnol.*, 2(1):40–61, January 2009.

[23] Sofia Santos Costa, Miguel Viveiros, Leonard Amaral, and Isabel Couto. Multidrug efflux pumps in staphylococcus aureus: an update. *Open Microbiol. J.*, 7(1):59–71, March 2013.

[24] M C Roberts, J Sutcliffe, P Courvalin, L B Jensen, J Rood, and H Seppala. Nomenclature for macrolide and macrolide-lincosamide-streptogramin B resistance determinants. *Antimicrob. Agents Chemother.*, 43(12):2823–2830, December 1999.

[25] Jung-A Lim, Ae-Ran Kwon, Sook-Kyung Kim, Yunsop Chong, Kungwon Lee, and Eung-Chil Choi. Prevalence of resistance to macrolide, lincosamide and streptogramin antibiotics in gram-positive cocci isolated in a korean hospital. *J. Antimicrob. Chemother.*, 49(3):489–495, March 2002.

[26] Jody L Floyd, Kenneth P Smith, Sanath H Kumar, Jared T Floyd, and Manuel F Varela. LmrS is a multidrug efflux pump of the major facilitator superfamily from staphylococcus aureus. *Antimicrob. Agents Chemother.*, 54(12):5406–5412, December 2010.

[27] J I Ross, E A Eady, J H Cove, W J Cunliffe, S Baumberg, and J C Wootton. Inducible erythromycin resistance in staphylococci is encoded by a member of the ATP-binding transport super-gene family. *Mol. Microbiol.*, 4(7):1207–1214, July 1990.

[28] R Kelmani Chandrakanth, S Raju, and S A Patil. Aminoglycoside-resistance mechanisms in multidrug-resistant staphylococcus aureus clinical isolates. *Curr. Microbiol.*, 56(6):558–562, June 2008.

[29] J E Dowding. Mechanisms of gentamicin resistance in staphylococcus aureus. *Antimicrob. Agents Chemother.*, 11(1):47–50, January 1977.

[30] Maria S Ramirez and Marcelo E Tolmasky. Aminoglycoside modifying enzymes. *Drug Resist. Updat.*, 13(6):151–171, December 2010.

[31] K Trzcinski, B S Cooper, W Hryniewicz, and C G Dowson. Expression of resistance to tetracyclines in strains of methicillin-resistant staphylococcus aureus. *J. Antimicrob. Chemother.*, 45(6):763–770, June 2000.

[32] Q C Truong-Bolduc, G R Bolduc, H Medeiros, J M Vyas, Y Wang, and D C Hooper. Role of the tet38 efflux pump in staphylococcus aureus internalization and survival in epithelial cells. *Infect. Immun.*, 83(11):4362–4372, November 2015.

[33] P Huovinen. Resistance to trimethoprim-sulfamethoxazole. *Clin. Infect. Dis.*, 32(11):1608–1614, June 2001.

[34] Jun-Ichiro Sekiguchi, Prasit Tharavichitkul, Tohru Miyoshi-Akiyama, Vena Chupia, Tomoko Fujino, Minako Araake, Atsushi Irie, Koji Morita, Tadatoshi Kuratsuji, and Teruo Kirikae. Cloning and characterization of a novel trimethoprim-resistant dihydrofolate reductase from a nosocomial isolate of staphylococcus aureus CM.S2 (IMCJ1454). *Antimicrob. Agents Chemother.*, 49(9):3948–3951, September 2005.

[35] Takehiko Mima, Naoki Kohira, Yang Li, Hiroshi Sekiya, Wakano Ogawa, Teruo Kuroda, and Tomofusa Tsuchiya. Gene cloning and characteristics of the RND-type multidrug efflux pump MuxABC-OpmB possessing two RND components in pseudomonas aeruginosa. *Microbiology*, 155(Pt 11):3509–3517, November 2009.

[36] S Jalal, O Ciofu, N Hoiby, N Gotoh, and B Wretlind. Molecular mechanisms of fluoroquinolone resistance in pseudomonas aeruginosa isolates from cystic fibrosis patients. *Antimicrob. Agents Chemother.*, 44(3):710–712, March 2000.

[37] Maria Tomás, Michel Doumith, Marina Warner, Jane F Turton, Alejandro Beceiro, German Bou, David M Livermore, and Neil Woodford. Efflux pumps, OprD porin, AmpC beta-lactamase, and multiresistance in pseudomonas aeruginosa isolates from cystic fibrosis patients. *Antimicrob. Agents Chemother.*, 54(5):2219–2224, May 2010.

[38] Benjamin A Evans and Sebastian G B Amyes. OXA β-lactamases. *Clin. Microbiol. Rev.*, 27(2):241–263, April 2014.

[39] Priyanka Bajaj, Nambram S Singh, and Jugsharan S Virdi. Escherichia coli β-Lactamases: What really matters. *Front. Microbiol.*, 7:417, March 2016.

[40] Maria Karczmarczyk, Marta Martins, Teresa Quinn, Nola Leonard, and Séamus Fanning. Mechanisms of fluoroquinolone resistance in escherichia coli isolates from food-producing animals. *Appl. Environ. Microbiol.*, 77(20):7113–7120, October 2011.

[41] João Anes, Matthew P McCusker, Séamus Fanning, and Marta Martins. The ins and outs of RND efflux pumps in escherichia coli. *Front. Microbiol.*, 6:587, June 2015.

[42] Paul Christoffer Lindemann, Kine Risberg, Harald G Wiker, and Haima Mylvaganam. Aminoglycoside resistance in clinical escherichia coli and klebsiella pneumoniae isolates from western norway. *APMIS*, 120(6):495–502, June 2012.

[43] Dominika Ojdana, Anna Sieńko, Paweł Sacha, Piotr Majewski, Piotr Wieczorek, Anna Wieczorek, and Elżbieta Tryniszewska. Genetic basis of enzymatic resistance of e. coli to aminoglycosides. *Adv. Med. Sci.*, 63(1):9–13, March 2018.

[44] Vaida Šeputienė, Justas Povilonis, Modestas Ružauskas, Alvydas Pavilonis, and Edita Sužiedėlienė. Prevalence of trimethoprim resistance genes in escherichia coli isolates of human and animal origin in lithuania. *J. Med. Microbiol.*, 59(Pt 3):315–322, March 2010.

[45] Detmar KÃcken, Heinz-Hubert Feucht, and Paul-Michael Kaulfers. Association ofqacEandqacEÎ1 with multiple resistance to antibiotics and antiseptics in clinical isolates of gram-negative bacteria. *FEMS Microbiol. Lett.*, 183(1):95–98, February 2000.

[46] Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, 47(D1):D195–D202, January 2019.

[47] S Sreedharan, M Oram, B Jensen, L R Peterson, and L M Fisher. DNA gyrase gyra mutations in ciprofloxacin-resistant strains of staphylococcus aureus: close similarity

with quinolone resistance mutations in escherichia coli. *J. Bacteriol.*, 172(12):7260–7262, December 1990.

[48] F J Schmitz, M E Jones, B Hofmann, B Hansen, S Scheuring, M Lückefahr, A Fluit, J Verhoef, U Hadding, H P Heinz, and K Köhrer. Characterization of grla, grlb, gyra, and gyrb mutations in 116 unrelated isolates of staphylococcus aureus and effects of mutations on ciprofloxacin MIC. *Antimicrob. Agents Chemother.*, 42(5):1249–1252, May 1998.

[49] Roghayeh Nouri, Mohammad Ahangarzadeh Rezaee, Alka Hasani, Mohammad Aghazadeh, and Mohammad Asgharzadeh. The role of gyra and parc mutations in fluoroquinolones-resistant pseudomonas aeruginosa isolates from iran. *Braz. J. Microbiol.*, 47(4):925–930, October 2016.

[50] T Akasaka, M Tanaka, A Yamaguchi, and K Sato. Type II topoisomerase mutations in fluoroquinolone-resistant clinical strains of pseudomonas aeruginosa isolated in 1998 and 1999: role of target enzyme in mechanism of fluoroquinolone resistance. *Antimicrob. Agents Chemother.*, 45(8):2263–2268, August 2001.

[51] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014.

[52] Mélodie Duval, Daniel Dar, Filipe Carvalho, Eduardo P C Rocha, Rotem Sorek, and Pascale Cossart. HflXr, a homolog of a ribosome-splitting factor, mediates antibiotic resistance. *Proc. Natl. Acad. Sci. U. S. A.*, 115(52):13359–13364, December 2018.

[53] Marion S Dorer, Jutta Fero, and Nina R Salama. DNA damage triggers genetic exchange in helicobacter pylori. *PLoS Pathog.*, 6(7):e1001026, July 2010.

[54] Jiqiang Ling, Chris Cho, Li-Tao Guo, Hans R Aerni, Jesse Rinehart, and Dieter Söll. Protein aggregation caused by aminoglycoside action is prevented by a hydrogen peroxide scavenger. *Mol. Cell*, 48(5):713–722, December 2012.

[55] Daniel J Dwyer, Peter A Belenky, Jason H Yang, I Cody MacDonald, Jeffrey D Martell, Noriko Takahashi, Clement T Y Chan, Michael A Lobritz, Dana Braff, Eric G Schwarz, Jonathan D Ye, Mekhala Pati, Maarten Vercruysse, Paul S Ralifo, Kyle R Allison, Ahmad S Khalil, Alice Y Ting, Graham C Walker, and James J Collins. Antibiotics induce redox-related physiological alterations as part of their lethality. *Proc. Natl. Acad. Sci. U. S. A.*, 111(20):E2100–9, May 2014.

[56] Stephen M Kwong, Joshua P Ramsay, Slade O Jensen, and Neville Firth. Replication of staphylococcal resistance plasmids. *Front. Microbiol.*, 8:2279, November 2017.

Chapter 4

Global pathogenomic analysis for antimicrobial resistance in twelve species

4.1 Abstract

Surveillance programs for managing antimicrobial resistance (AMR) have yielded thousands of genomes suited for data-driven mechanism discovery. We present a workflow integrating pangenomics, gene annotation, and machine learning to identify known and novel AMR genes at scale. Applied to 12 species spanning 27,155 genomes and 69 drugs, we 1) found AMR gene transfer mostly confined within related species, with 925 genes in multiple species but just eight in multiple classes, 2) demonstrated that discovery-oriented support vector machines outperform traditional methods at recovering known AMR genes, recovering 265 genes compared to 125 by Fisher's exact test, and 3) identified 142 novel AMR gene candidates. Validation of two candidates revealed cases of conditional resistance: $\Delta cycA$ conferred ciprofloxacin resistance in minimal media with D-serine, and *frdD* V111D conferred ampicillin resistance in the presence of *ampC* by modifying the overlapping promoter. We expect this approach to be adaptable to other species and phenotypes.

4.2 Background

Antimicrobial resistance (AMR) remains a persistent problem in the treatment of bacterial infections. With resistance having been observed against nearly all major antibiotics [1], 700,000 annual deaths are currently attributable to AMR globally and is projected to increase to as high as 10 million by 2050 without major interventions [2]. One strategy for managing AMR is the large-scale sequencing of infection isolates [3], which has yielded tens of thousands of publicly-available genome sequences for each major bacterial pathogen that are frequently paired with resistance metadata [4].

This wealth of data has enabled global analyses on the genetics of AMR, many of which employ machine learning (ML) to predict AMR phenotypes directly from genetic variations [5, 6]. Accurate AMR phenotype prediction models trained on thousands of genomes have been developed for many pathogens such as *Escherichia coli* [7,8], *Klebsiella pneumoniae* [9], *Mycobacterium tuberculosis* [10], *Salmonella enterica* [11], or multiple species [12–15], with numerous others developed from smaller datasets. However, many of these studies report a significant fraction of their models' predictive genetic features to have no relationship to known AMR mechanisms. This disconnect between statistically identified and mechanistically established genetic determinants of AMR highlights the current gap in knowledge in AMR genetics and remains a challenge for the real world adoption of ML-based systems for rapidly predicting AMR and informing treatment strategies [16, 17].

However, the predictive features identified by ML naturally provide a source of testable AMR gene candidates for closing this knowledge gap. While these candidates are often treated as byproducts of phenotype prediction models with candidate identification assigned to traditional genome-wide association studies (GWAS), there is a growing effort towards developing ML workflows specifically aimed at mechanism discovery within AMR genetics [17, 18] and beyond [19]. Recent ML-aided GWAS have identified genetic and

metabolic mechanisms behind resistance in *M. tuberculosis* [20], and demonstrated ML's competitiveness compared to typical statistical testing at recovering known AMR genes for multiple pathogens [21]. These successes demonstrate the potential for ML to carry out the role of GWAS for the ever-growing public collection of bacterial genome sequences and AMR data.

We present a machine learning analysis tailored towards AMR gene discovery consisting of three components drawn from previous workflows: 1) pangenome construction by sequence clustering to enumerate biologically-interpretable genetic features [22], 2) systematic annotation of known AMR genes among those features with RGI [23], and 3) training of support vector machine (SVM) ensembles to learn relationships between all genetic features and a given AMR phenotype [21]. This workflow was evaluated for both phenotype prediction accuracy and recovery of known AMR genes among predictive features across 12 pathogenic species spanning 127 species-drug combinations and 27,155 genomes. We find that this approach provides both a comprehensive overview of the distribution of known AMR genes and consistently yields ML models that both accurately predict AMR phenotype and significantly outperform traditional statistical testing at recovering known AMR genes. Functional analysis of strongly predictive features yielded a set of 142 novel AMR gene candidates, of which two were experimentally confirmed to impact resistance in *E. coli*.

4.3 Results

4.3.1 Assembly of twelve bacterial pathogen pangenomes and antimicrobial resistance data

A total of 27,155 genomes across 12 species were downloaded from the PATRIC database [4] after filtering for assembly quality and availability of AMR data (Dataset C.1). Pangenomes were constructed and genetic features were enumerated for each species

using the sequence clustering approach previously described [22] with CD-HIT v4.6 [24], yielding six unique genetic feature types: 1) open reading frame (ORF) clusters or "genes", 2) ORF variants or "alleles", 3) 300bp ORF-flanking upstream variants or "5′ variants", 4) 300bp ORF-flanking downstream variants or "3′ variants", 5) noncoding feature clusters, and 6) noncoding feature variants (Fig. 4.1a-b, Table C.1).

Experimentally derived susceptible-intermediate-resistant (SIR) phenotypes for the genomes were assembled from directly reported SIRs and SIRs inferred from minimum inhibitory concentrations (MICs). MIC breakpoints for SIR inference were determined from genomes with both SIR and MIC values for each species-drug-standard combination (i.e. CLSI, EUCAST) to yield internally-consistent SIR data (see Methods). From 169,693 MICs, 22,772 SIR inferences across 93 species-drug cases were generated for genomes without SIR data. In total, 176,911 SIR phenotypes were assembled across 69 drugs, with 88.3% phenotypes from directly reported SIRs and 11.7% inferred from MICs (Fig. 4.1c-d, Dataset C.2), comprising the largest internally-consistent AMR dataset known to the authors at the time of publication.

Known AMR genes were identified through direct annotation of alleles by RGI v5.2.0 with CARD ontology v3.1.3 [23] and parsing PATRIC text annotations for drug-associated terms. 7,710 AMR genes were identified across all species, spanning 95,491 gene-drug mappings (Fig. 4.1e, Dataset C.3). The fewest number of AMR genes (71) were identified in *Campylobacter coli* and the most in *Escherichia coli* (1,533), with the best annotation occurring for genes related to major drug classes such as beta-lactams, aminoglycosides, and quinolones.

4.3.2 Global analysis of known AMR genes reveals potential phylogenetic limitations on cross-species gene transfer

To examine the distribution of known AMR genes, all AMR gene alleles across all species were re-clustered with CD-HIT, yielding 6,332 unified AMR genes. Rates at

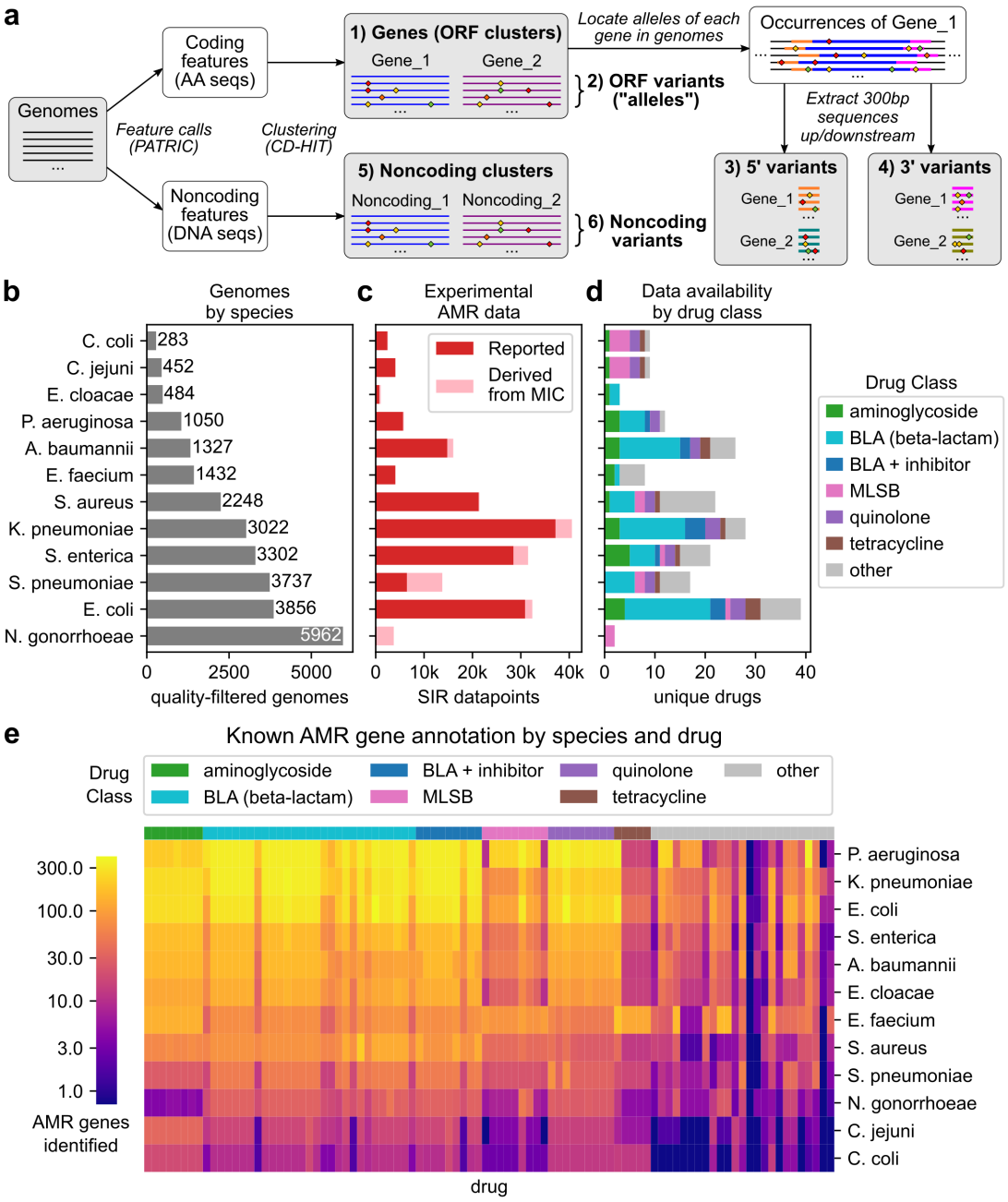


Figure 4.1. Genomic and antimicrobial resistance datasets assembled from the PATRIC database for 12 pathogenic species. (a) Workflow for extracting six types of biological features from a species' genome collection. For each species, the number of (b) high-quality genomes with AMR data, (c) experimentally measured SIR data points (directly reported or inferred from reported MICs) with consistent testing standards across all drugs, and (d) number of drugs for which SIR data is available is shown. Species have been sorted by number of genomes. (e) Number of known AMR genes identified for each species-drug case. Drugs are sorted by drug class, and species are sorted by total number of AMR gene-drug mappings identified.

which these genes were plasmid- or chromosomally-encoded were predicted by labeling contigs containing AMR genes as plasmid or chromosomal with PlasFlow [25] (Dataset C.3). 925 AMR genes were observed in multiple species, with more broadly distributed genes having a greater tendency to be plasmid-encoded (95% of genes in >4 species were plasmid-encoded in a majority of occurrences) (Fig. 4.2a). Similarly, out of 68,324 unique AMR alleles, 830 were observed in multiple species and were primarily plasmid-encoded (Fig. C.1a).

Compared against AMR gene categories, specific functions were significantly enriched among both plasmid-encoded (over chromosomal) and multispecies (over single species) AMR genes (Fig. 4.2b, Fig. C.2, Dataset C.3). Dihydrofolate reductases/dihydropteroate synthases and aminoglycoside modifying enzymes were significantly enriched by both measures (Fisher's exact test, FWER < 0.05, Bonferroni correction, 36 tests), with $\log_2$ odds ratios (LORs) for plasmid over chromosomal genes of 3.0 and 1.8, and LORs for multispecies over single species of 1.5 and 1.3, respectively. Other categories significantly enriched in plasmid genes but with limited enrichment among multispecies genes include chloramphenicol acetyltransferases, ribosomal protection proteins, rRNA methyltransferases, and beta-lactamases (plasmid LOR > 1.0, multispecies LOR < 1.0) (Table C.2). Generally, multispecies AMR genes tended to be plasmid-encoded and vice versa, with the exception of *rpoB* variants (Fig. 4.2b). As *rpoB* is a highly conserved chromosomal bacterial gene [26], this exception may be due to many rarely observed *rpoB* fragments on short contigs being misclassified as plasmid-encoded (Dataset C.3).

The 925 multispecies AMR genes and 830 AMR alleles were predominantly shared within species of the same phylogenetic class, especially within Gammaproteobacteria (Fig. 4.2c, Fig. C.1b). Only 68 (7.4%) multispecies AMR genes and 38 (4.6%) alleles spanned more than one class. Of these, just 8 genes and 5 alleles were observed in at least 10 genomes in each of at least two different phylogenetic classes (Fig. 4.2d, Fig. C.1c). These 8 multi-class genes are functionally varied, including TEM family

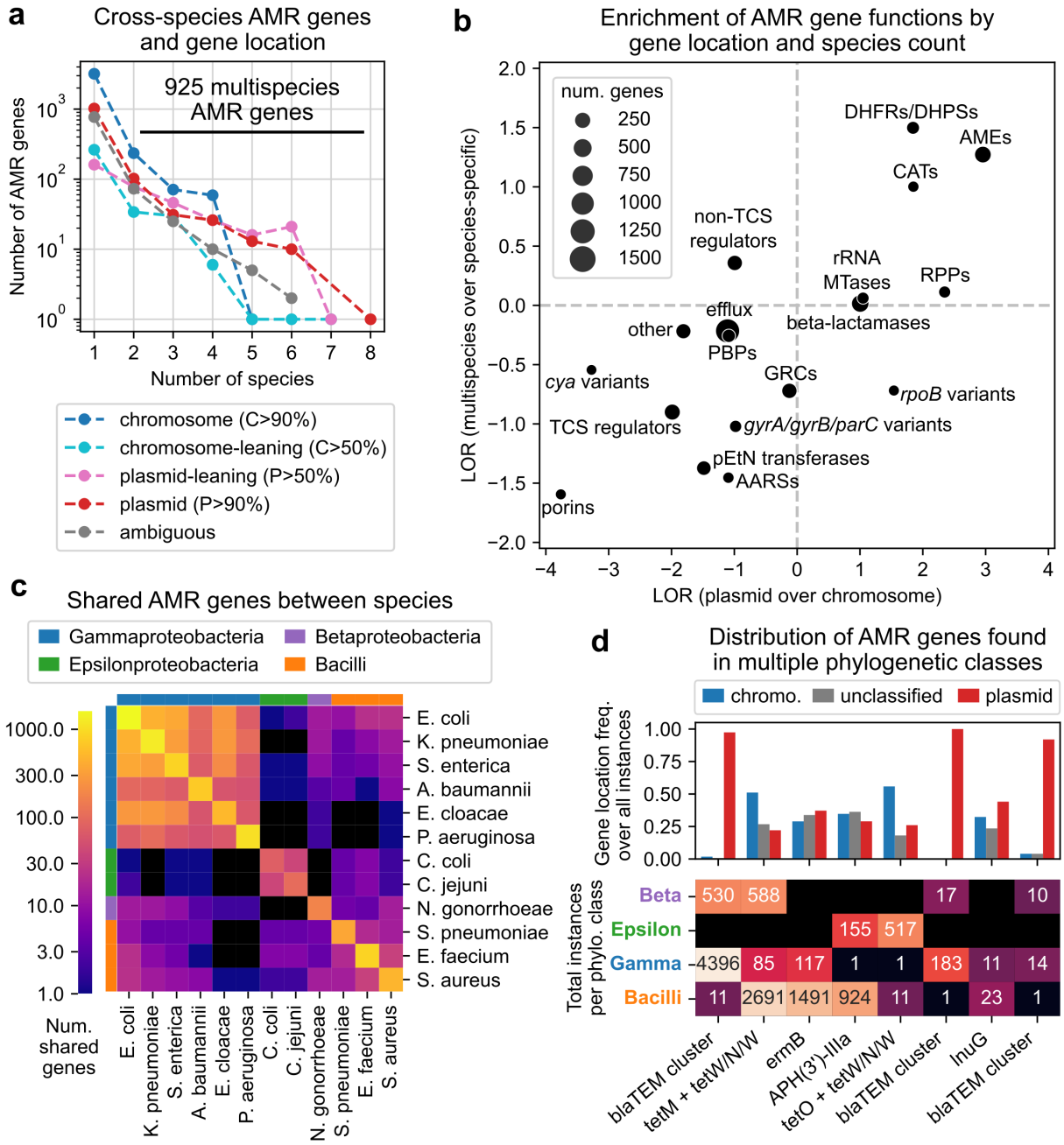


Figure 4.2. Cross-species analysis of 6,332 antimicrobial resistance genes, gene locations, and functions. (a) Relationship between the number of species an AMR gene is observed in and tendency to be plasmid-encoded. (b) Enrichment of AMR gene functional categories in plasmid- over chromosomally-encoded genes and in multispecies over single species genes, based on $\log_2$ odds ratios (LORs). (c) Number of AMR genes shared between each pair of species, compared to species phylogenetic class. (d) Distribution of predicted gene locations and total occurrences per phylogenetic class for AMR genes appearing in at least 10 genomes of multiple classes.

beta-lactamases (blaTEMs), ribosomal protection proteins *tetM*, *tetO*, and *tet(W/N/W)*, 23S rRNA methyltransferase *ermB*, aminoglycoside 3'-phosphotransferase *APH(3')-IIIa*, and lincosamide nucleotidyltransferase *lnuG*. The blaTEMs were observed exclusively on plasmids while all other multi-class AMR genes were observed on both plasmid and chromosomal DNA.

4.3.3 Observation of TEM-family beta-lactamases in both gram-positive and gram-negative strains

Given the prevalence of blaTEMs, all observed blaTEM alleles were mapped to known blaTEM alleles based on RGI annotations, focusing on the 51 "complete" alleles with length within 5% of the TEM-1 allele. Complete blaTEMs were observed in 4,861 genomes spanning 8 species and were dominated by the TEM-1 allele occurring in 4,424 genomes (Fig. C.3a, Table C.3, Dataset C.4). All but five alleles were within two mutations of TEM-1, and only 9 alleles (including TEM-1) were observed in at least 10 genomes. Individual alleles were largely specific to phylogenetic class with 38 alleles limited to Gammaproteobacteria and 10 alleles limited to *N. gonorrhoeae*, compared to two alleles observed in both (TEM-1 and TEM-135). Just one allele was observed in gram-positive strains, TEM-116 (substitutions V82I and A182V relative to TEM-1), occurring in 11 *S. aureus* strains and 17 *S. enterica* strains. Contigs harboring blaTEM were mapped to known plasmids on PLSDB [27] using MASH [28], and the 28 instances of TEM-116 were predicted to be located on one of three plasmids: NZ_AJ437107.1 (14 *S. enterica*, 1 *S. aureus*), NZ_AJ438270.1 (3. *S. enterica*), and NC_019053.1 (10 *S. aureus*) (Fig. C.3b, Dataset C.4).

To assess the impact of blaTEM on beta-lactam resistance in *S. aureus*, the presence of blaTEM and other beta-lactam AMR genes was compared to cefoxitin MIC data, which was available for *S. aureus* genomes harboring blaTEM. Genomes with blaTEM in addition to *mecA* and *blaZ* had significantly elevated MICs compared to those with just *mecA*

and *blaZ* or *mecA* alone ($p = 0.0074$, Mann-Whitney U-test), suggesting blaTEM confers additional protection from beta-lactams in already resistant *S. aureus* (Fig. C.3c, Fig. C.4).

4.3.4 A GWAS-oriented machine learning approach for the identification of AMR-associated genes significantly outperforms conventional statistical testing

To identify novel AMR-associated genetic features, a ML framework was developed to train models for both accuracy at predicting AMR phenotypes and biological relevance, i.e. ability to assign high feature weights to known AMR genes (Fig. 4.3a). For a given species-drug case, we started with the support vector machine (SVM) ensemble design described in our previous study [21]. SVM ensembles were trained to classify genomes as susceptible or non-susceptible based on the presence or absence of the genetic features (grouped into six types described previously). Four hyperparameters (HPs), parameters not learned from the data but fixed in advance to control the learning process and model complexity, were varied to evaluate their impact on model performance. Ensembles under various HP combinations were evaluated over 5-fold cross validation (5CV) for 1) accuracy, as the Matthews correlation coefficient (MCC) on test set genomes to account for class imbalance, and 2) biological relevance, through a "GWAS score" defined as a weighted sum of the rankings of known AMR genetic features after sorting model features by feature weight absolute value. A feature was labeled as a known AMR genetic feature if its corresponding gene cluster was a known AMR gene for the drug of interest as annotated by RGI and PATRIC.

In an analysis of ten representative species-drug cases (Table C.4), we find that ensemble size had little impact on either accuracy or biological relevance, while the other HPs significantly impacted both metrics: SVM regularization term C, fraction of samples per estimator, and fraction of features per estimator (Kruskal-Wallis test, FWER < 0.05,

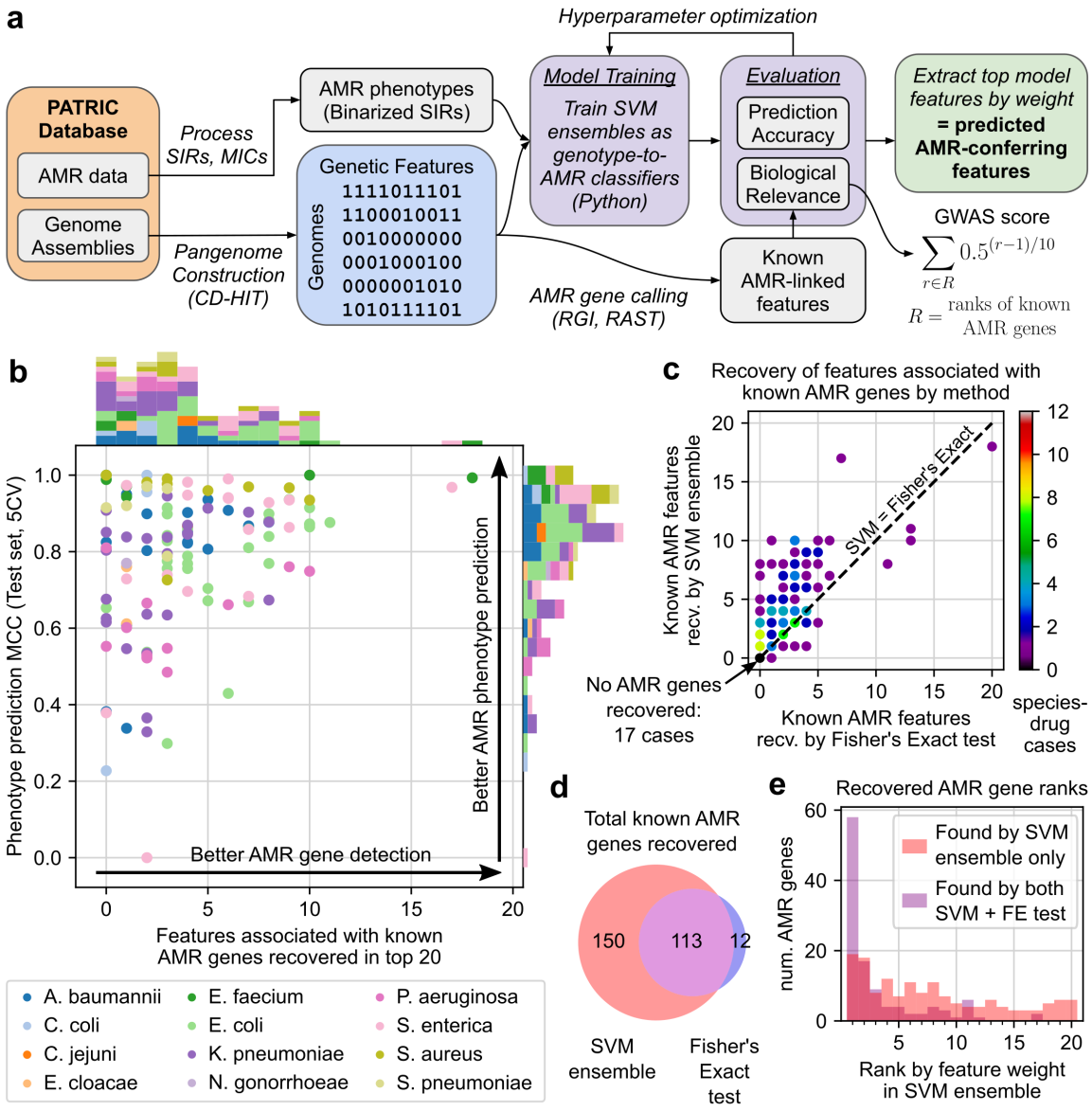


Figure 4.3. Evaluation of a GWAS-oriented machine learning workflow for identifying AMR-associated genetic features in 127 species-drug cases. (a) Workflow integrating SVM ensembles, hyperparameter optimization, and AMR gene annotation to train models suited for both phenotype prediction and AMR gene identification from genome assemblies and SIR phenotype data. (b) Performance of models for 127 species-drug cases, by phenotype prediction accuracy (mean test MCC from 5-fold cross validation) and recovery of known AMR genes. (c) Comparison between SVM ensembles and Fisher's exact tests at recovering known AMR genes. (d) Total known AMR gene-drug mappings recovered by SVM ensembles and Fisher's exact tests across all cases. (e) Rankings of known AMR genes among SVM ensemble top features, grouped by whether or not the gene was also recovered by Fisher's exact test.

Bonferroni correction, 80 tests) (Fig. C.5, Fig. C.6a, Table C.5). More aggressive feature subsampling (smaller fraction of features per estimator) also consistently improved on biological relevance without compromising accuracy (Fig. C.6b). Finally, the smallest subset of HP combinations containing nearly-optimal models (within 90% of maximum MCC and GWAS scores) was identified for more efficient HP optimization of the remaining species-drug cases (Table C.6).

HP optimization was carried out for 127 species-drug cases with at least 100 SIRs, 10 known AMR genes, and minority phenotype >5%. For each case, models under each HP combination were ranked by mean MCC and GWAS score from 5CV, and the HP set with the highest average of the two ranks was selected as optimal. Among the final models, 41 (32%) achieved MCC>90% and 78 (61%) achieved MCC>80% on the test set during 5CV, and 103 models (81%) recovered at least one known AMR genetic feature among the top 20 features (Fig. 4.3b, Dataset C.5). Broadly, a high MCC was necessary but did not guarantee better recovery of known AMR genetic features, i.e. accuracy did not guarantee biological relevance. Certain dataset parameters were weakly but significantly associated with performance: Dataset size was associated with both accuracy and AMR gene recovery, species with accuracy, and class imbalance with AMR gene recovery (Spearman R or Kruskal-Wallis test, FWER < 0.05, Bonferroni correction, 8 tests) (Table C.7, Fig. C.6c-d). Finally, relative to models with fixed HPs closest to those in our previous work [21], optimizing HPs offered modest but consistent improvements to both accuracy and known AMR gene recovery. 5CV experiments showed a mean increase in test MCCs and known AMR genes recovered among the top 20 features of 0.035 and 0.230, respectively (Fig. C.7, Dataset C.5). The top 50 features for each model are available in Dataset C.5, and their sequences in Dataset C.6.

As a baseline level of AMR gene recovery, for each species-drug case, Fisher's exact tests were conducted between each genetic feature and AMR phenotype and features were sorted by p-value. Based on the number of known AMR genetic features recovered

among the top 20 features, the SVM ensemble approach outperformed Fisher's exact test in 78 cases (61%), had equal performance in 38 cases (30%), and underperformed in just 11 cases (9%) (Fig. 4.3c, Dataset C.5). Nearly half of the equal performance cases (17/38, 45%) were instances where neither method could recover any known AMR features and were significantly enriched for cases with fewer available genomes ($p = 0.003$, Mann-Whitney U-test). Across all 127 cases, 113 known AMR gene-drug mappings were recovered by both methods, 150 by SVM ensemble only, and just 12 by Fisher's exact test only (Fig. 4.3d, Dataset C.5). Examining SVM ensemble feature rankings, known AMR genes were distributed throughout the full range of ranks among the top 20 features per model, whereas those that were also recovered by Fisher's exact test were concentrated among the top 3 features (84/113, 74%) with over half being the top weighted feature of the corresponding SVM model (58/113, 51%) (Fig. 4.3e).

An examination of the 12 genes only recovered by Fisher's exact test suggests several possible failure modes of SVM ensemble approach (Table C.8). First, in 4/12 cases, the missed gene is captured slightly outside of the top 20 feature threshold for recovery, with three genes ranked 21 and one ranked 33 by SVM. Second, in another 4/12 cases, many of the top features in the corresponding model were perfectly correlated, saturating the top ranks with these correlates and preventing recovery of additional AMR genes. Third, for 3/12 cases, the corresponding model was not very accurate, with mean test MCC ranging from 0.43 to 0.77. The final remaining case (*arlR* for ciprofloxacin resistance in *S. aureus*) could not be explained by these previous failure modes.

4.3.5 Identification of 142 candidate AMR-conferring genes through cross-drug and functional analysis of AMR-predictive features

We next applied several filters to translate the best performing models to a smaller set of novel, high-confidence, AMR-conferring gene candidates (Fig. 4.4a). Starting with

the 78 models achieving test MCC>80%, the top 10 features from each model were filtered for those that were not already known AMR genes, occurred in at least 10 genomes with SIR data, and had both positive feature weight and LOR for resistant genomes, yielding 347 features predictive of AMR without known associations to AMR. These candidates were scored based on the number of drugs in the same class for which the feature enriches for resistance and the extent of co-occurrence with known AMR genes. Taking the top 10 features by this score for each species-drug class pair yielded 142 AMR gene candidates (see Methods and Dataset C.7).

43 candidates were functionally well-characterized and consisted of four feature types: 16 genes (ORF clusters, Table 4.1), 14 gene coding variants (Table 4.2), and 13 gene

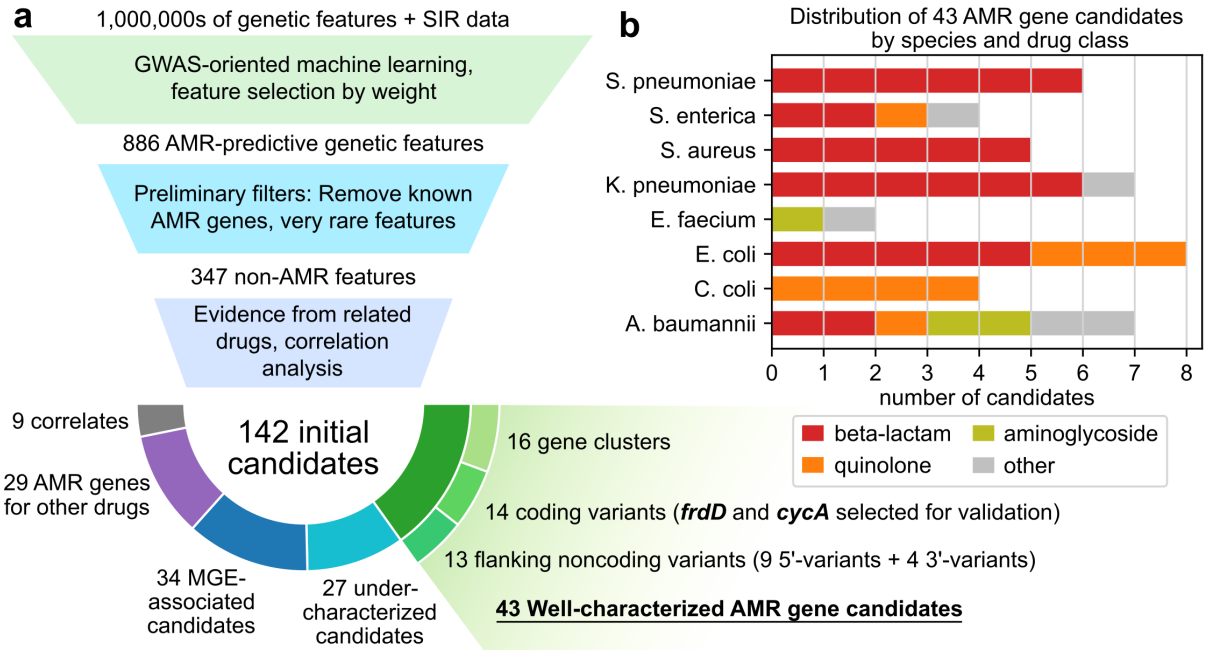


Figure 4.4. Identification of AMR gene candidates from machine learning models through cross-drug and functional analysis. (a) Candidate identification workflow. From each trained AMR-ML model, the top 10 predictive features were identified, filtered for features that are neither known AMR genes nor very rare, ranked based on statistical evidence for resistance in other related drugs, and finally categorized by functional annotation. When multiple features related to the same gene were predictive of resistance, the feature with the strongest evidence was selected and the others were labeled "correlates". Mobile genetic element is abbreviated MGE. (b) Distribution of 43 well-characterized AMR gene candidates by species and drug class.

5′/3′ flanking region variants (Table C.9). The candidates spanned 8 species and 7 drug classes and were majority beta-lactam associated (26/43) due to the relative abundance of beta-lactam AMR data (Fig. 4.4b). The candidates spanned many genetic functions, and two functions occurred more than twice. Candidates related to small multidrug resistance (SMR) efflux transporters (*qacE*, *qacE*Δ*1*, *sugE*) typically associated with resistance to quaternary ammonium compounds [29], were linked to resistance against aminoglycosides, beta-lactams, diaminopyrimidines, and sulfonamides across four species, consistent with previous studies that find SMR transporters associated with resistance against a broad range of antibiotics beyond antiseptics [30, 31]. Additionally, three different formate dehydrogenase genes (*fdhF*, *fdsA*, *fdnG*) were associated with beta-lactam resistance in *K. pneumoniae*, suggesting the importance of formate metabolism in AMR, possibly with respect to stress response [32]. Finally, a majority of the sequence-variant level candidates (16/27), especially those related to flanking regions (10/13), were the most common variant of their respective gene clusters, suggesting that most observed perturbations to these genes may be deleterious with respect to AMR.

We selected the two coding variant candidates related to *E. coli* core genes for experimental validation: Wildtype D-serine/D-alanine/glycine transporter (*cycA*) associated with quinolone resistance, and fumarate reductase subunit D (*frdD*) with a V111D substitution associated with beta-lactam resistance. Wildtype (WT) was defined as the most common variant of the gene among *E. coli* genomes, which in the case of *frdD* and *cycA*, were also the variants present in both the BW25113 and K-12 MG1655 reference genomes.

Table 4.1. 16 gene clusters predicted to be associated with resistance against specific drug classes for individual species.

Species	Drug Class	Accession ID (Gene)	Predicted Gene Product	Resistant/ Susceptible	LORs
Acinetobacter baumannii	AMG	ENU75377.1 (*ydcZ*)	Putative inner membrane exporter	GEN=147/0 AMK=144/3 TOB=114/18	GEN=7.7 AMK=5.8 TOB=1.9
Acinetobacter baumannii	AMG	WP_000679427.1 (*qacEΔ1*)	Small multidrug resistance (SMR) efflux transporter	GEN=453/6 AMK=296/170 TOB=269/163	GEN=4.4 AMK=1.5 TOB=-0.6
Acinetobacter baumannii	beta-lactam	ACC56111.1 (*yfhL*)	Uncharacterized ferredoxin-like protein	AMP=82/0 DOR=13/0 CFZ=45/0	AMP=6.5 DOR=6.1 CFZ=5.6
Enterococcus faecium	AMG	WP_000228166.1	Nucleotidyltransferase domain containing protein	STR=89/0	STR=14.8
Enterococcus faecium	glycopeptide	WP_000754864.1 (*cadC*)	Cadmium resistance transcriptional regulatory protein	TEC=138/0 VAN=498/150	TEC=14.8 VAN=3.4
Escherichia coli	beta-lactam	WP_000243817.1	Tryptophan synthase (indole-salvaging)	CXM=196/1 CTX=236/4 SAM=42/0	CXM=9.1 CTX=7.7 SAM=7.3
Escherichia coli	quinolone	WP_000598813.1 (*dcuC*)	Anaerobic C4-dicarboxylate transporter	CIP=59/14 LVX=31/3	CIP=3.3 LVX=2.1
Klebsiella pneumoniae	beta-lactam	WP_002885150.1 (*fdhF*)	Formate dehydrogenase H	CRO=1508/11 ETP=130/1 CFZ=1487/58	CRO=5.3 ETP=4.0 CFZ=3.7
Klebsiella pneumoniae	beta-lactam	AKE78078.1 (*fdsA*)	Formate dehydrogenase H	CRO=1499/10 CFZ=1486/40 ETP=130/1	CRO=5.4 CFZ=4.4 ETP=4.0
Klebsiella pneumoniae	beta-lactam	EWF54733.1 (*fdnG*)	Formate dehydrogenase N alpha subunit	CRO=1494/10 ETP=128/1 CFZ=1478/49	CRO=5.4 ETP=4.0 CFZ=3.9
Klebsiella pneumoniae	beta-lactam	AHG50656.1 (*ygbK*)	3-oxo-tetronate kinase	CRO=71/0 AMP=62/0 CAZ=107/1	CRO=6.3 AMP=6.0 CAZ=4.0
Klebsiella pneumoniae	diamino-pyrimidine	WP_000679427.1 (*qacEΔ1)*	Small multidrug resistance (SMR) efflux transporter	TMP=10/0 SXT=855/42	TMP=6.2 SXT=3.6
Salmonella enterica	sulfonamide	WP_000800531.1 (*qacE*)	Small multidrug resistance (SMR) efflux transporter	SXT=11/4 SIX=15/0 SMZ=12/4	SXT=6.6 SIX=5.4 SMZ=0.4
Staphylococcus aureus	beta-lactam	WP_000872606.1 (*maoC*)	MaoC domain protein	MET=202/4 FOX=786/0 OXA=29/3	MET=16.9 FOX=16.2 OXA=15.3
Staphylococcus aureus	beta-lactam	WP_000616816.1 (*cadD*)	Cadmium resistance transporter	PEN=461/28 BPG=82/4 FOX=330/165	PEN=2.0 BPG=0.1 FOX=-1.5
Staphylococcus aureus	beta-lactam	WP_001186608.1	Site-specific recombinase	FOX=661/0 MET=198/58 OXA=8/13	FOX=12.3 MET=8.9 OXA=3.3

Accession IDs are provided for the most common sequence variant of each gene cluster (RefSeq when possible, GenBank otherwise), along with gene names when available and gene products. The number of resistant/susceptible genomes and $\log_2$ odds ratios (LORs) for resistance are shown for the top three drugs by LOR when data for more than three related drugs was available. Drug class AMG refers to aminoglycoside and individual drug abbreviations are available in Dataset C.7.

Table 4.2. 14 gene coding variants predicted to be associated with resistance against specific drug classes for individual species.

Species	Drug Class	Accession ID (Gene)	Predicted Gene Product	Muts.*	Resistant/ Susceptible	LORs
Acinetobacter baumannii	quinolone	WP_000586912.1 (*lptF*)	Lipopolysaccharide export system permease protein	-	CIP=795/166 LVX=731/143	CIP=3.0 LVX=2.5
Acinetobacter baumannii	tetracycline	ADX03353.1 (*bccA*)	Biotin carboxylase	-	TET=107/0 MIN=86/20	TET=7.4 MIN=4.0
Campylobacter coli	quinolone	AHK72934.1 (*rny*)**	Ribonuclease Y	d1-86, V87M	NAL=41/16 CIP=39/18	NAL=2.9 CIP=2.6
Escherichia coli	beta-lactam	WP_000811566.1 (*yjfN*)	Putative uncharacterized protein, DUF1471	-	FOX=46/0 CTT=4/15 CXM=246/136	FOX=9.4 CTT=4.5 CXM=3.2
Escherichia coli	beta-lactam	WP_000118520.1 (*sugE*)	Small multidrug resistance (SMR) efflux transporter	T37A, M85A, A88L, A91G, L95A, +13	FOX=40/0 CAZ=101/1 CRO=85/0	FOX=9.0 CAZ=7.7 CRO=7.5
Escherichia coli	beta-lactam	WP_001588947.1 (*frdD*)***	Fumarate reductase subunit D	V111D	AMC=19/1 AMP=14/0 CXM=12/1	AMC=4.7 AMP=4.5 CXM=4.4
Escherichia coli	quinolone	WP_000228346.1 (*cycA*)***	D-serine, D-alanine, glycine transporter	-	LVX=251/35 CIP=573/468 NAL=41/17	LVX=4.0 CIP=3.0 NAL=0.7
Klebsiella pneumoniae	beta-lactam	WP_004183775.1	Cold shock protein of CSP family	V54A, H55L, A57T, Q62P, +2	TIM=16/2 AMP=107/0 CEF=13/0	TIM=7.0 AMP=6.8 CEF=4.1
Salmonella enterica	beta-lactam	WP_001221666.1 (*blc*)	Lipocalin Blc	L49F, S98D, S175P, +10	CRO=302/0 AMC=301/0 CTF=297/3	CRO=13.6 AMC=12.1 CTF=10.9
Salmonella enterica	beta-lactam	WP_000118520.1 (*sugE*)	Small multidrug resistance (SMR) efflux transporter	T37A, A104T, +9	CRO=304/0 AMC=303/0 CTF=299/3	CRO=13.7 AMC=12.1 CTF=11.0
Staphylococcus aureus	beta-lactam	WP_000872606.1 (*maoC*)	MaoC domain protein	-	OXA=29/3 MET=201/4 FOX=713/0	OXA=15.3 MET=14.5 FOX=13.1
Staphylococcus aureus	beta-lactam	WP_000958858.1	Glycerophosphoryl -diester phosphodiesterase	-	OXA=29/3 FOX=775/0 MET=201/3	OXA=15.3 FOX=15.2 MET=14.8
Streptococcus pneumoniae	beta-lactam	WP_000248982.1	Cell wall surface anchor family protein	V66I, D111E, S160G, E172K	AMX=13/2 MEM=15/0 CXM=15/0	AMX=13.1 MEM=10.1 CXM=9.7
Streptococcus pneumoniae	beta-lactam	WP_000203066.1 (*gpo*)	Glutathione peroxidase	A80T, S132G	AMX=13/3 MEM=16/0 CXM=16/0	AMX=12.7 MEM=10.6 CXM=10.3

Accession IDs are provided for the exact sequence variant (RefSeq when possible, GenBank otherwise), along with gene names when available, gene products, and mutations with respect to the most commonly observed variant. The number of resistant/susceptible genomes and $\log_2$ odds ratios (LORs) for resistance are shown for the top three drugs by LOR when data for more than three related drugs was available. Drug abbreviations are available in Dataset C.7.
*When more than four mutations are detected, only mutations with BLOSUM62 score ≤ 0 are shown and the number of additional mutations is listed at the end. Full mutation sets are available in Dataset C.7. If no mutations are shown, the variant of interest is the most commonly observed variant.
**No exact variant was found on GenBank. AHK72934.1 refers to the most common variant of the gene, while the variant of interest contains the 86-amino acid N-terminal truncation "d1-86".
***Selected for experimental validation.

4.3.6 Experimental validation 1: Loss of amino acid transporter CycA confers limited quinolone resistance in minimal media with D-serine

The WT *cycA* variant, the fifth highest weighted feature in the *E. coli* AMR model for levofloxacin, had the highest LOR for resistance among all *cycA* variants in 2/4 quinolone drugs (Fig. 4.5a). To assess the impact of *cycA* on quinolone resistance, maximum cell density was measured for BW25113 (WT) and corresponding $\Delta cycA$ mutant (KO, from the Keio collection [33]) under 60 conditions based on three variables: concentration of ciprofloxacin (CIP), supplementation with known substrates of the D-serine/D-alanine/glycine transporter encoded by *cycA* [34], and choice of rich vs. minimal media (cation-adjusted Mueller-Hinton Broth CA-MHB vs. M9 media with glucose) (Dataset C.8). 6/60 tested conditions resulted in significantly different final densities between WT and KO (Welch t-test, FDR < 0.05, Benjamini-Hochberg correction), three of which involved D-serine and M9 media (Fig. 4.5b). Across all conditions involving D-serine, while increasing CIP concentration reduced final density for both strains, the KO strain achieved higher densities for 16-125μg/L CIP, but only in M9 media and not in CA-MHB (Fig. 4.5c).

One explanation consistent with this conditional increase in CIP resistance by *cycA* KO involves the toxicity of D-serine, its transport by CycA, and interaction with quinolones through the SOS response (Fig. 4.5d). D-serine, which inhibits L-serine and pantothenate biosynthesis, can be bacteriostatic in minimal media [35]. D-serine uptake can be impaired by *cycA* KO but also through competitive inhibition of CycA by other substrates in rich media [34], and its toxicity mitigated by direct uptake of L-serine and pantothenate in rich media. With respect to CIP, both D-serine [36] and fluoroquinolones [37] induce SOS response but to differing extents. As systematic alterations in SOS response induction have been shown to reduce fluoroquinolone resistance [38], the presence of both D-serine and CIP may result in a SOS response that is adapted to neither stress and consequently

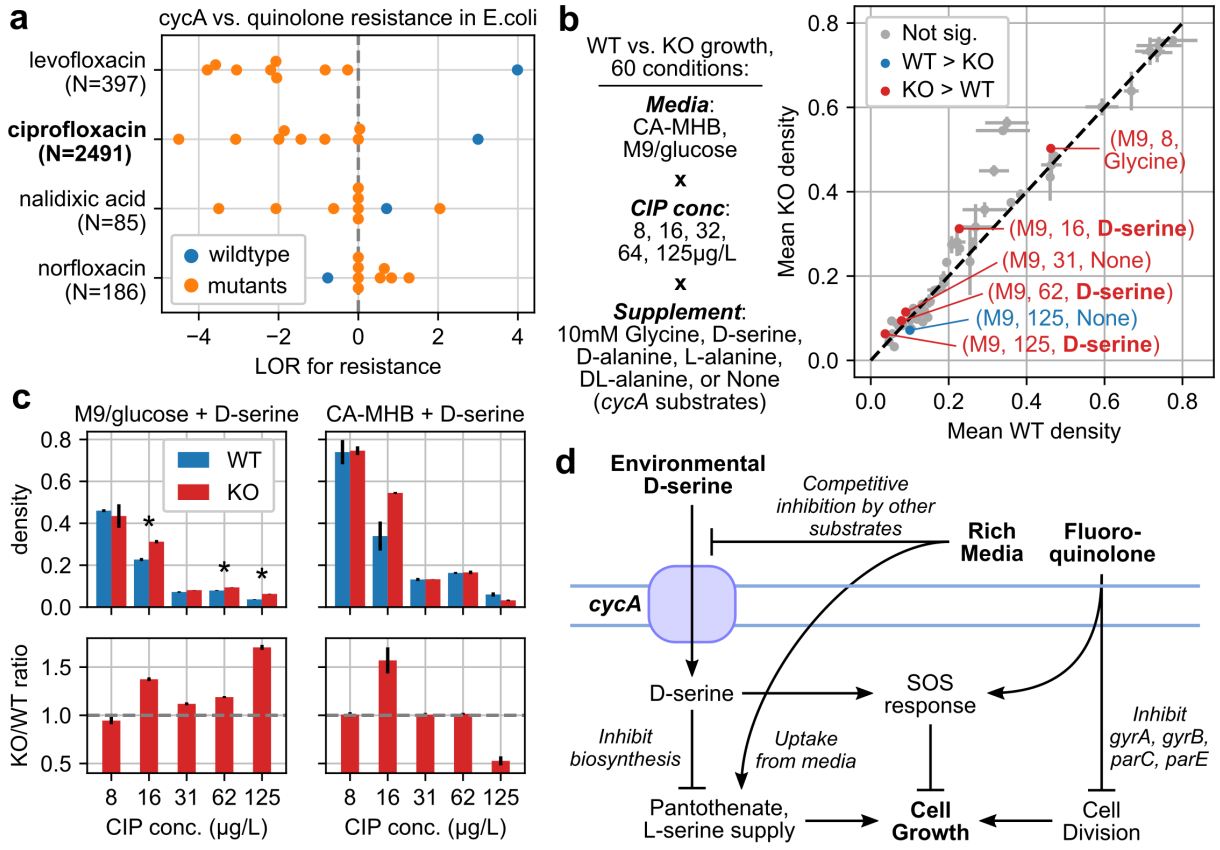


Figure 4.5. D-serine-dependent impact of amino acid transporter CycA on quinolone resistance. (a) Enrichment for resistant over susceptible genomes across observed *cycA* variants for four quinolone drugs, based on $\log_2$ odds ratio (LOR). Number of genomes with AMR data is shown for each drug, and only variants observed in at least 100 genomes are shown. (b) Maximum cell density (OD600) achieved by *E. coli* BW25113 wildtype (WT) vs. *cycA* knockout (KO) across 60 conditions. Error bars indicate standard deviations centered around means from biological triplicates. Conditions where density differs significantly between WT and KO are labeled (FDR < 0.05, Welch t-test, Benjamini-Hochberg correction). (c) Absolute and relative maximum densities between WT and KO in conditions involving D-serine. Error bars indicate standard deviations. Significant differences between WT and KO are starred. (d) Possible model for interactions between D-serine, *cycA*, media, and fluoroquinolones consistent with the observed conditional resistance conferred by *cycA* KO.

results in greater susceptibility to CIP.

4.3.7 Experimental validation 2: The V111D substitution in *frdD* confers beta-lactam resistance solely through altering expression of the overlapping beta-lactamase gene *ampC*

Next, the *frdD* V111D variant was selected for validation as it was the third highest weighted feature in the *E. coli* AMR model for ampicillin, in the top 10 for four *E. coli* beta-lactam models, and enriched for resistant strains (LOR > 3) in 7/14 beta-lactam drugs with AMR data (Fig. 4.6a). Given the proximity of *frdD* to the adjacent beta-lactamase gene *ampC*, this mutation occurs in the -35 box of the *ampC* promoter [39] and coincides with *ampC* overexpression mutations known to increase beta-lactam resistance [40, 41] (Fig. 4.6b). Appropriately, an *ampC* 5′ variant containing the equivalent mutation was ranked 5th in the model for amoxicillin-clavulanate, though no other 5′ variants with the mutation were in the top 50 hits of any other beta-lactam model. To assess whether *frdD* V111D is simply a byproduct of an *ampC* promoter mutation or contributes separately to beta-lactam resistance, we examined two *frdD* mutations both resulting in the V111D substitution but with different effects on *ampC* transcription as predicted by Promoter Calculator [42]: 1) 332T>A, predicted to increase *ampC* transcription 2.6-fold, or 2) 332TC>AT, predicted to have minimal effect on *ampC* (Fig. 4.6b). Six *E. coli* strains were examined, based on *frdD* variants (WT or either mutation) generated in either BW25113 or corresponding Δ*ampC* mutant (from the Keio collection [33]). Maximum cell density achieved by these strains were measured under increasing concentrations of ampicillin in either rich (CA-MHB) or minimal (M9 + glucose) media (Dataset C.8).

Across all strains, only the mutant with both *ampC* and the overexpression mutation was able to grow at ≥2mg/L ampicillin in either media, and final densities were not impacted even at 8mg/L (Fig. 4.6c). The mutant with *ampC* and the non-overexpressing *frdD* mutation grew at 2mg/L in CA-MHB only, but this growth was delayed by at least

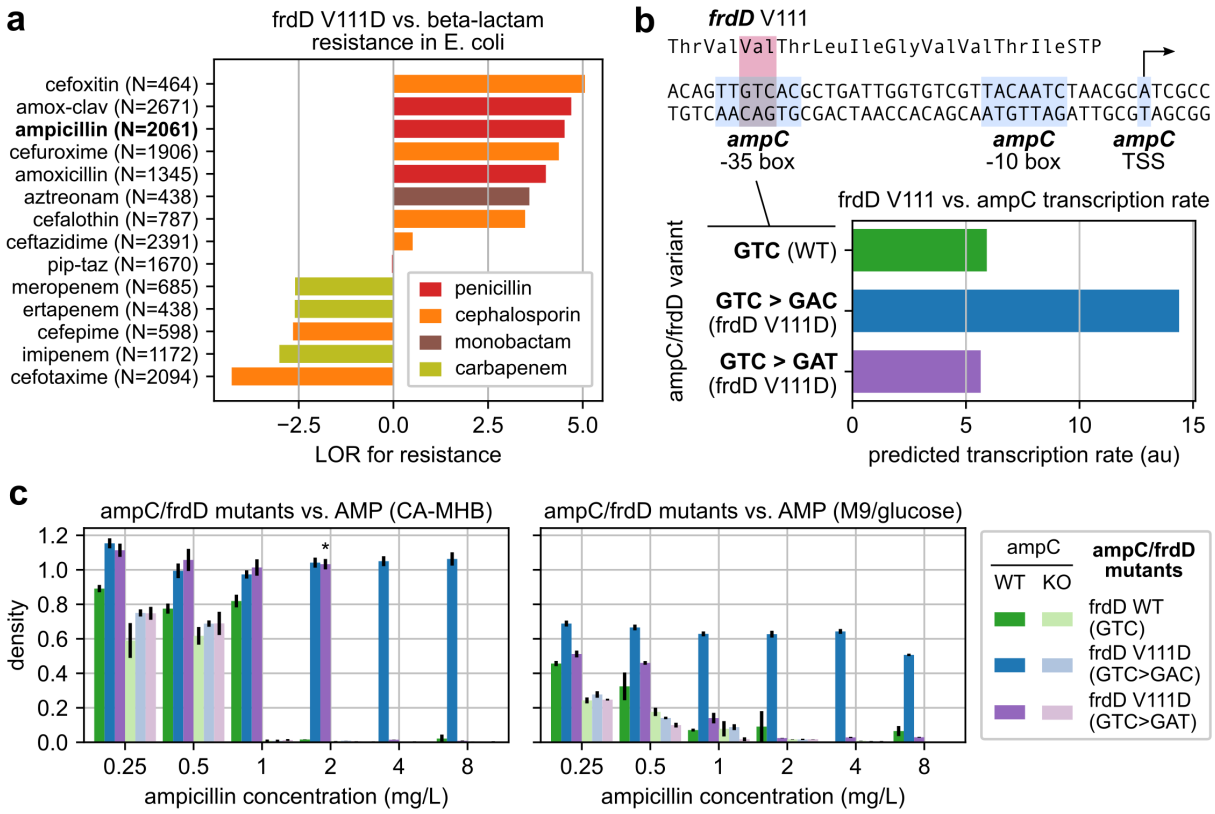


Figure 4.6. *ampC*-dependent beta-lactam resistance conferred by the V111D substitution in fumarate reductase subunit FrdD. (a) Enrichment for resistant over susceptible genomes from the presence of the V111D *frdD* mutation across 14 beta-lactam drugs, based on $\log_2$ odds ratio (LOR). Number of genomes with AMR data is shown for each drug. (b) Overlap between the *frdD* coding region and *ampC* promoter in *E. coli* BW25113 and predicted *ampC* transcription rates for various *frdD* V111 mutations. (c) Maximum cell density (OD600) achieved by *ampC* and *frdD* mutants under increasing ampicillin concentrations in rich (CA-MHB) or minimal (M9) media. Error bars indicate standard deviations centered around means from biological triplicates. Six genotypes were tested, from combinations between three possible codons at *frdD* V111 in either the BW25113 wildtype (WT) or corresponding *ampC* knockout (KO) strain. Starred case indicates growth was not observed until at least 8 hours after inoculation for all replicates.

8 hours in all three replicates, suggesting some degree of susceptibility. Furthermore, all Δ*ampC* mutants failed to grow at $\geq$2mg/L ampicillin regardless of *frdD* status and reached lower densities than their corresponding *ampC* WT strain in 17/18 conditions with <2mg/L ampicillin and significantly so in 15/18 cases (FDR < 0.05, Welch t-test, Benjamini-Hochberg correction, Dataset C.8); the only exception was between strains with WT *frdD* at 1mg/L ampicillin in M9 in which both the *ampC* and *ampC* KO strains exhibited very little growth (OD < 0.1). These results suggest that the *frdD* V111D substitution requires *ampC* to confer beta-lactam resistance and is unlikely to contribute to resistance through any *ampC*-independent mechanism.

4.4 Discussion

The exponential growth of publicly-available bacterial genome sequences and resistance metadata provides valuable opportunities for applying statistical methods to elucidate the genetics of AMR at a global scale. By applying workflows in pangenome construction, systematic AMR gene annotation, and machine learning to 27,155 genomes, 12 species, and 176,911 SIR phenotypes, we have characterized the current interspecies distribution of known AMR genes and demonstrated the broad capability of ML to carry out GWAS analyses of AMR, surpassing traditional statistical tests at recovering known AMR genes. From the most accurate ML models, we have identified 142 novel AMR genetic feature candidates, two of which we have experimentally verified to impact resistance in *E. coli*. Many of these results depend on the reliable recovery of rare biological features and events and were enabled by the scale of this analysis, operating on the largest internally-consistent AMR phenotype dataset known to the authors at the time of publication.

First, an analysis of 6,332 known AMR genes revealed 925 (14.6%) genes to be present in multiple species, which tended to be plasmid-encoded over chromosomally-

encoded in a function-dependent manner. This result is consistent with previous studies finding plasmids frequently responsible for cross-species [43] and cross-genera [44] transfer of AMR genes in clinical environments. However, we also find that AMR gene transfer is much rarer between species differing at higher phylogenetic ranks, with just 8 AMR genes observed in at least 10 genomes outside their main phylogenetic class. It has been suggested that the transfer of AMR genes between unrelated species such as between gram-positive (GP) and gram-negative (GN) species is rare but possible, having been inferred for tetracycline resistance proteins, *ermB*, *APH(3')* [45], and observed for some beta-lactamases [46].

A case study of blaTEM revealed one variant, TEM-116, present in GP strains (*S. aureus*) and significantly increasing beta-lactam resistance in already resistant strains. TEM-116 was observed on three plasmids of which plasmid NZ_AJ437107.1 was observed in both GP and GN strains, suggesting a potential route of transfer. While blaTEM was only recently reported in *S. aureus* [47], the *S. aureus* genomes we identified to harbor TEM-116 were isolated as early as 2009, suggesting that this transfer may be a much older phenomenon. Further analysis of isolation dates may enable the reconstruction of timelines for the spread of multispecies AMR genes, and help identify whether specific species act as reservoirs for enabling interspecies AMR gene transfer. Given these observations, the transfer of AMR genes between GP and GN strains should be treated as a potential significant contributor to the spread of resistance, and large-scale analyses will likely be necessary to continue capturing these rare events of AMR gene transfer between unrelated strains.

Next, we developed a ML workflow for identifying AMR-associated genetic features by optimizing models for both phenotype prediction accuracy and biological relevance, with the latter quantified through a "GWAS score" based on the rankings of known AMR genes among a model's predictive features. In a systematic analysis of four HPs, we found that the optimal ensemble size rarely exceeded 50 estimators, much lower than

previous works and suggesting that smaller, more computationally efficient models are sufficient at this scale. Feature subsampling was also confirmed to improve the recovery of known AMR genes without compromising accuracy in a majority of cases. However, HP optimization offered only modest improvements over using fixed HPs, in contrast to a previous GWAS analysis with neural networks on datasets of comparable scale which found HP optimization to significantly improve performance [48]. Additional analyses are needed to explore the practical limits of HP optimization for SVMs and AMR genetics, especially regarding how the competing metrics of accuracy and biological relevance may limit performance gains.

Applying this approach to 127 species-drug cases yielded models that were both accurate and reliably recovered known AMR genes, while confirming at a larger scale that accuracy is necessary but does not guarantee biological relevance [21]. Compared with Fisher's exact test from traditional GWAS statistical testing, SVM ensembles recovered a near superset of known AMR genes, and recovered more than twice as many known AMR genes in total. AMR genes recovered by both methods were concentrated among the top 3 features of the corresponding SVM ensemble, suggesting that only genes with the strongest statistical signals are reliably recovered by Fisher's exact test. While more sophisticated tests that account for population stratification such as Breslow-Day or Cochran-Mantel-Haenszel may outperform Fisher's exact test [49], ML approaches aid AMR gene recovery without needing to define population structure in advance. These results suggest that even small SVM ensembles with limited HP tuning can reliably carry out GWAS analyses for AMR genetics.

Analysis of the most accurate models yielded 142 novel AMR gene candidates, two of which were selected for experimental validation: amino acid transporter CycA vs. quinolone resistance, and fumarate reductase subunit FrdD vs. beta-lactam resistance. Testing the effects of *cycA* KO vs. WT in *E. coli* in various environments, we found that *cycA* KO confers modest but significant, conditional resistance against CIP, specifically in minimal

media supplemented with the CycA substrate D-serine. One possible explanation of this result involves the 1) toxicity of D-serine [35], which may be mitigated by reducing uptake through loss of CycA or competitive inhibition by other CycA substrates in rich media [34], and 2) the SOS response, which is triggered to differing extents by both D-serine [36] and fluoroquinolones [37] and may result in an SOS response induction no longer optimal to either stress and ultimately greater CIP susceptibility [38]. As neither CycA nor D-serine directly influence CIP's mechanism of action yet ultimately impact growth under CIP exposure, these results provide an example of separate environmental stresses and related genes measurably influencing AMR. Further investigation into D-serine may continue to shed light onto clinically relevant conditional resistance, as D-serine has also been shown to sensitize *S. aureus* to various beta-lactams [50] and concentrations of D-serine similar to those tested here may be encountered in various host environments [36]. CycA substrate glycine has also been shown to sensitize serum-resistant *E. coli* but not $\Delta cycA$ mutants [51], suggesting the critical role of CycA in multiple cases of environment-dependent resistance.

Similar experiments on the effect of the *frdD* V111D mutation on beta-lactam resistance revealed that resistance was conferred only in the presence of beta-lactamase encoding *ampC*, and comparing synonymous codons found this was likely due to overexpression of *ampC* as its promoter overlaps with the *frdD* open reading frame [39]. This substitution was previously associated with ampicillin resistance in *E. coli* induction experiments but not discussed in relation to *ampC* [52], while a similar synonymous mutation V117V was previously shown to be enriched in amoxicillin resistant *E. coli* and attributed to *ampC* overexpression [53]. Given the prevalence of overlapping genes in bacterial genomes [54], future GWAS analyses that identify novel genes associated with a phenotype will benefit from incorporating the genetic context of such genes into their analysis.

Overall, combining pangenomics, systematic gene annotation, and ML provides a workflow for efficiently uncovering patterns of known and novel AMR genes at the scale

of 10,000s of genomes with greater reliability than traditional statistical testing. The flexibility of ML provides numerous opportunities to continue improving this workflow, such as the direct integration of the GWAS score into the loss function, use of different model architectures beyond SVMs, input of additional genetic feature types such as SNPs, or benchmarking against other phenotypes beyond AMR. As the number of genome sequences continues to grow, periodic updates to this analysis with improved techniques will likely be able to capitalize on the benefits of scale and steadily deepen our understanding of AMR across the phylogenetic tree. Continued development is necessary to bring ML into the current GWAS toolbox when mining sequencing data for novel genetic determinants underlying complex phenotypes.

4.5 Acknowledgements

J.C.H., J.M.M., and B.O.P. contributed project Conceptualization. J.C.H. developed the computational Methodology and Software, and conducted the Data Curation, Investigation, Formal Analysis, Validation, and Visualization of results. R.S. developed the Methodology and conducted the Investigation for generating *E. coli* mutants. Y.H. developed the Methodology and conducted the Investigation for measuring growth of *E. coli* mutants in various conditions. J.M.M. and B.O.P. contributed to Funding Acquisition and Project Administration, and provided Supervision and Resources. J.C.H. prepared the Original Draft and all authors were involved in Review and Editing.

This research was supported by a grant from the National Institute of Allergy and Infectious Diseases (U01-AI124316, awarded to J.M.M. and B.O.P.). This research was also supported by a grant from the National Institutes of Health (T32GM8806, awarded to J.C.H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 4 has been submitted for publication in *Nature Communcations*: **Ja-**

son C Hyun, Jonathan M Monk, Richard Szubin, Ying Hefner, Bernhard O Palsson. "Global pathogenomic analysis identifies known and novel genetic antimicrobial resistance determinants in twelve species." The disseration author is the primary author.

4.6 References

[1] C Lee Ventola. The antibiotic resistance crisis: part 1: causes and threats. *P T*, 40(4):277–283, April 2015.

[2] Jim O'Neill and Wellcome Trust. Antimicrobial resistance : Tackling a crisis for the health and wealth of nations, December 2014.

[3] Claire Waddington, Megan E Carey, Christine J Boinett, Ellen Higginson, Balaji Veeraraghavan, and Stephen Baker. Exploiting genomics to mitigate the public health impact of antimicrobial resistance. *Genome Med.*, 14(1):15, February 2022.

[4] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, Rory M Butler, Philippe Chlenski, Neal Conrad, Allan Dickerman, Emily M Dietrich, Joseph L Gabbard, Svetlana Gerdes, Andrew Guard, Ronald W Kenyon, Dustin Machi, Chunhong Mao, Dan Murphy-Olson, Marcus Nguyen, Eric K Nordberg, Gary J Olsen, Robert D Olson, Jamie C Overbeek, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, Chris Thomas, Margo VanOeffelen, Veronika Vonstein, Andrew S Warren, Fangfang Xia, Dawen Xie, Hyunseung Yoo, and Rick Stevens. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.*, 48(D1):D606–D612, January 2020.

[5] Michelle Su, Sarah W Satola, and Timothy D Read. Genome-based prediction of bacterial antibiotic resistance. *J. Clin. Microbiol.*, 57(3), March 2019.

[6] Jonathan M Monk. Predicting antimicrobial resistance and associated genomic features from whole-genome sequencing. *J. Clin. Microbiol.*, 57(2), February 2019.

[7] Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and

Dominik Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2):325–334, October 2021.

[8] Danesh Moradigaravand, Martin Palm, Anne Farewell, Ville Mustonen, Jonas Warringer, and Leopold Parts. Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data. *PLoS Comput. Biol.*, 14(12):e1006258, December 2018.

[9] Marcus Nguyen, Thomas Brettin, S Wesley Long, James M Musser, Randall J Olsen, Robert Olson, Maulik Shukla, Rick L Stevens, Fangfang Xia, Hyunseung Yoo, and James J Davis. Developing an in silico minimum inhibitory concentration panel test for klebsiella pneumoniae. *Sci. Rep.*, 8(1):421, January 2018.

[10] Martin Hunt, Phelim Bradley, Simon Grandjean Lapierre, Simon Heys, Mark Thomsit, Michael B Hall, Kerri M Malone, Penelope Wintringer, Timothy M Walker, Daniela M Cirillo, Iñaki Comas, Maha R Farhat, Phillip Fowler, Jennifer Gardy, Nazir Ismail, Thomas A Kohl, Vanessa Mathys, Matthias Merker, Stefan Niemann, Shaheed Vally Omar, Vitali Sintchenko, Grace Smith, Dick van Soolingen, Philip Supply, Sabira Tahseen, Mark Wilcox, Irena Arandjelovic, Tim E A Peto, Derrick W Crook, and Zamin Iqbal. Antibiotic resistance prediction for mycobacterium tuberculosis from genome sequence data with mykrobe. *Wellcome Open Res.*, 4:191, December 2019.

[11] Marcus Nguyen, S Wesley Long, Patrick F McDermott, Randall J Olsen, Robert Olson, Rick L Stevens, Gregory H Tyson, Shaohua Zhao, and James J Davis. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J. Clin. Microbiol.*, 57(2), February 2019.

[12] D Aytan-Aktug, P T L C Clausen, V Bortolaia, F M Aarestrup, and O Lund. Prediction of acquired antimicrobial resistance for multiple bacterial species using neural networks. *mSystems*, 5(1), January 2020.

[13] Jiwoong Kim, David E Greenberg, Reed Pifer, Shuang Jiang, Guanghua Xiao, Samuel A Shelburne, Andrew Koh, Yang Xie, and Xiaowei Zhan. VAMPr: VAriant mapping and prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput. Biol.*, 16(1):e1007511, January 2020.

[14] Marcus Nguyen, Robert Olson, Maulik Shukla, Margo VanOeffelen, and James J Davis. Predicting antimicrobial resistance using conserved genes. *PLoS Comput. Biol.*, 16(10):e1008319, October 2020.

[15] James J Davis, Sébastien Boisvert, Thomas Brettin, Ronald W Kenyon, Chunhong Mao, Robert Olson, Ross Overbeek, John Santerre, Maulik Shukla, Alice R Wattam, Rebecca Will, Fangfang Xia, and Rick Stevens. Antimicrobial resistance prediction in PATRIC and RAST. *Sci. Rep.*, 6:27930, June 2016.

[16] Jee In Kim, Finlay Maguire, Kara K Tsang, Theodore Gouliouris, Sharon J Peacock, Tim A McAllister, Andrew G McArthur, and Robert G Beiko. Machine learning for antimicrobial resistance prediction: Current practice, limitations, and clinical perspective. *Clin. Microbiol. Rev.*, 35(3):e0017921, September 2022.

[17] Melis N Anahtar, Jason H Yang, and Sanjat Kanjilal. Applications of machine learning to the problem of antimicrobial resistance: An emerging model for translational research. *J. Clin. Microbiol.*, 59(7):e0126020, June 2021.

[18] Janak Sunuwar and Rajeev K Azad. A machine learning framework to predict antibiotic resistance traits and yet unknown genes underlying resistance to specific antibiotics in bacterial strains. *Brief. Bioinform.*, 22(6), November 2021.

[19] Hannah L Nicholls, Christopher R John, David S Watson, Patricia B Munroe, Michael R Barnes, and Claudia P Cabrera. Reaching the end-game for GWAS: Machine learning approaches for the prioritization of complex disease loci. *Front. Genet.*, 11:350, April 2020.

[20] Erol S Kavvas, Laurence Yang, Jonathan M Monk, David Heckmann, and Bernhard O Palsson. A biochemically-interpretable machine learning classifier for microbial GWAS. *Nat. Commun.*, 11(1):2580, May 2020.

[21] Jason C Hyun, Erol S Kavvas, Jonathan M Monk, and Bernhard O Palsson. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput. Biol.*, 16(3):e1007608, March 2020.

[22] Jason C Hyun, Jonathan M Monk, and Bernhard O Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022.

[23] Brian P Alcock, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, Sally Y Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E Werfalli, Jalees A Nasir, Martins Oloni, David J Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N Sharma, Emily Bordeleau, Andrew C Pawlowski, Haley L Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L Winsor, Robert G Beiko, Fiona S L Brinkman, William W L Hsiao, Gary V Domselaar, and Andrew G McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 48(D1):D517–D525, January 2020.

[24] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[25] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, 46(6):e35, April 2018.

[26] Jean-Claude Ogier, Sylvie Pagès, Maxime Galan, Matthieu Barret, and Sophie Gaudriault. rpob, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiol.*, 19(1):171, July 2019.

[27] Georges P Schmartz, Anna Hartung, Pascal Hirsch, Fabian Kern, Tobias Fehlmann, Rolf Müller, and Andreas Keller. PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.*, 50(D1):D273–D278, January 2022.

[28] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, June 2016.

[29] Denice C Bay, Kenton L Rommens, and Raymond J Turner. Small multidrug resistance proteins: a multidrug transporter family that continues to grow. *Biochim. Biophys. Acta*, 1778(9):1814–1838, September 2008.

[30] Mohamed E Enany, Abdelazeem M Algammal, Soad A Nasef, Sara A M Abo-Eillil, May Bin-Jumah, Ayman E Taha, and Ahmed A Allam. The occurrence of the multidrug resistance (MDR) and the prevalence of virulence genes and QACs resistance genes in e. coli isolated from environmental and avian sources. *AMB Express*, 9(1):192, December 2019.

[31] Z Jaglic and D Cervinkova. Genetic basis of resistance to quaternary ammonium compounds – the qac genes and their role: a review. *Vet. Med. (Praha)*, 57(6):275–281, July 2012.

[32] Yumi Iwadate, Noriyuki Funabasama, and Jun-Ichi Kato. Involvement of formate dehydrogenases in stationary phase oxidative stress tolerance in escherichia coli. *FEMS Microbiol. Lett.*, 364(20), November 2017.

[33] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli K-12 in-frame, single-gene knockout mutants: the keio collection. *Mol. Syst. Biol.*, 2(1):2006.0008, February 2006.

[34] J C Robbins and D L Oxender. Transport systems for alanine, serine, and glycine in escherichia coli K-12. *J. Bacteriol.*, 116(1):12–18, October 1973.

[35] S D Cosloy and E McFall. Metabolism of d-serine in escherichia coli k-12: mechanism of growth inhibition. *J. Bacteriol.*, 114(2):685–694, May 1973.

[36] James P R Connolly, Robert J Goldstone, Karl Burgess, Richard J Cogdell, Scott A Beatson, Waldemar Vollmer, David G E Smith, and Andrew J Roe. The host metabolite d-serine contributes to bacterial niche specificity through gene selection. *ISME J.*, 9(4):1039–1051, March 2015.

[37] Ting-Ting Qin, Hai-Quan Kang, Ping Ma, Peng-Peng Li, Lin-Yan Huang, and Bing Gu. SOS response and its regulation on the fluoroquinolone resistance. *Ann. Transl. Med.*, 3(22):358, December 2015.

[38] Charlie Y Mo, Sara A Manning, Manuela Roggiani, Matthew J Culyba, Amanda N Samuels, Paul D Sniegowski, Mark Goulian, and Rahul M Kohli. Systematically altering bacterial SOS activity under stress reveals therapeutic strategies for potentiating antibiotics. *mSphere*, 1(4), July 2016.

[39] T Grundström and B Jaurin. Overlap between ampc and frd operons on the escherichia coli chromosome. *Proc. Natl. Acad. Sci. U. S. A.*, 79(4):1111–1115, February 1982.

[40] Taru Singh, Praveen Kumar Singh, Shukla Das, Sayim Wani, Arshad Jawed, and Sajad Ahmad Dar. Transcriptome analysis of beta-lactamase genes in diarrheagenic escherichia coli. *Sci. Rep.*, 9(1):3626, March 2019.

[41] N Caroff, E Espaze, D Gautreau, H Richet, and A Reynaud. Analysis of the effects of -42 and -32 ampc promoter mutations in clinical isolates of escherichia coli hyperproducing ampc. *J. Antimicrob. Chemother.*, 45(6):783–788, June 2000.

[42] Travis L LaFleur, Ayaan Hossain, and Howard M Salis. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat. Commun.*, 13(1):5159, September 2022.

[43] Nicole A Lerminiaux and Andrew D S Cameron. Horizontal transfer of antibiotic resistance genes in clinical environments. *Can. J. Microbiol.*, 65(1):34–44, January 2019.

[44] Daniel R Evans, Marissa P Griffith, Alexander J Sundermann, Kathleen A Shutt, Melissa I Saul, Mustapha M Mustapha, Jane W Marsh, Vaughn S Cooper, Lee H Harrison, and Daria Van Tyne. Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital. *Elife*, 9, April 2020.

[45] P Courvalin. Transfer of antibiotic resistance genes between gram-positive and gram-negative bacteria. *Antimicrob. Agents Chemother.*, 38(7):1447–1451, July 1994.

[46] Prasanth Manohar, Thamaraiselvan Shanthini, Bulent Bozdogan, Cecilia Stalsby Lundborg, Ashok J Tamhankar, Nades Palaniyar, and Nachimuthu Ramesh. Transfer of antibiotic resistance genes from gram-positive bacterium to gram-negative bacterium. November 2020.

[47] Haoju Wang, Yao Chen, Xiaomei Jia, and Honglei Ding. Prevalence, antimicrobial resistance and staphylococcal toxin gene of blaTEM-1a-producing staphylococcus aureus isolated from animals in chongqing, china. August 2019.

[48] Junjie Han, Cedric Gondro, Kenneth Reid, and Juan P Steibel. Heuristic hyperparameter optimization of deep learning models for genomic prediction. *G3 (Bethesda)*, 11(7), July 2021.

[49] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

[50] Qing Wang, Yuemeng Lv, Jing Pang, Xue Li, Xi Lu, Xiukun Wang, Xinxin Hu, Tongying Nie, Xinyi Yang, Yan Q Xiong, Jiandong Jiang, Congran Li, and Xuefu You. In vitro and in vivo activity of d-serine in combination with β-lactam antibiotics against methicillin-resistant staphylococcus aureus. *Acta Pharm. Sin. B.*, 9(3):496–504, May 2019.

[51] Zhi-Xue Cheng, Chang Guo, Zhuang-Gui Chen, Tian-Ci Yang, Jian-Ying Zhang, Jie Wang, Jia-Xin Zhu, Dan Li, Tian-Tuo Zhang, Hui Li, Bo Peng, and Xuan-Xian Peng. Glycine, serine and threonine metabolism confounds efficacy of complement-mediated killing. *Nat. Commun.*, 10(1):3325, July 2019.

[52] Mengchen Li, Qiaoli Liu, Yanli Teng, Liuyang Ou, Yuanlin Xi, Shuaiyin Chen, and Guangcai Duan. The resistance mechanism of escherichia coli induced by ampicillin in laboratory. *Infect. Drug Resist.*, 12:2853–2863, September 2019.

[53] Nadine Händel, J Merijn Schuurmans, Stanley Brul, and Benno H ter Kuile. Compensation of the metabolic costs of antibiotic resistance by physiological adaptation in escherichia coli. *Antimicrob. Agents Chemother.*, 57(8):3752–3762, August 2013.

[54] Bradley W Wright, Mark P Molloy, and Paul R Jaschke. Overlapping genes in natural and engineered genomes. *Nat. Rev. Genet.*, 23(3):154–168, March 2022.

Chapter 5

Reconstructing the core genome of the last bacterial common ancestor

5.1 Abstract

Cumulative sequencing efforts have yielded enough genomes to construct pangenomes for dozens of bacterial species and elucidate intraspecies gene conservation. Given the diversity of organisms for which this is achievable, similar analyses for ancestral species are feasible through the integration of pangenomics and phylogenetics, promising deeper insights into the nature of ancient life. To this end, we constructed pangenomes for 183 bacterial species from 54,085 genomes and identified their core genomes using a novel statistical model to estimate genome-specific error rates and underlying gene frequencies. The core genomes were then integrated into a phylogenetic tree to reconstruct the core genome of the last bacterial common ancestor (LBCA), yielding three main results: 1) The gene content of modern and ancestral core genomes are diverse at the level of individual genes but are similarly distributed by functional category and share several poorly characterized genes. 2) The LBCA core genome is distinct from any individual modern core genome but has many fundamental biological systems intact, especially those involving translation machinery and biosynthetic pathways to all major nucleotides and amino acids. 3) Despite this metabolic versatility, the LBCA core genome likely requires additional non-core genes for viability, based on comparisons with the minimal organism

JCVI-Syn3A. These results suggest that many cellular systems commonly conserved in modern bacteria were not just present in ancient bacteria but were nearly immutable with respect to short-term intraspecies variation. Extending this analysis to other domains of life will likely provide similar insights into more distant ancestral species.

5.2 Significance

Modern genome sequencing has yielded sufficient data to characterize intraspecies genetic diversity such as the identification of core genes, genes that are conserved across all members of a species. We applied a novel, error-aware statistical model to identify core genes in 183 bacterial species, which were integrated into a phylogenetic tree to reconstruct the core genome of their last common ancestor. We find that this ancestral core genome is distinct from modern core genomes but is functionally versatile with many systems intact, especially those involving translation or biosynthesis of essential metabolites. These results suggest that ancient bacteria closely maintained many biological functions and expanding this analysis to other domains of life will enable similar characterizations of more distant ancestral species.

5.3 Introduction

Since Darwin's sketches of diverse species descending from a single ancestor, the growing volume of phenomenological data and, more recently, genomic data, has enabled increasingly comprehensive reconstructions of such "trees of life" [1, 2] and properties of the common ancestor at their roots [3–6]. These modern phylogenetic analyses can incorporate hundreds of thousands of samples to model the evolutionary history of entire domains of life and reconstruct universal common ancestors. However, the increasing scale of data has also produced competing theories regarding universal ancestry such as 1) due to horizontal gene transfer between distant species, evolutionary histories may not be

well-represented by a tree [7], and 2) that the last universal common ancestor (LUCA) was not a single organism, but rather a community of interdependent primitive cells [8–10]. These concerns highlight limitations in comprehensively modeling evolutionary history as a tree with individual organisms at each node.

A pangenomic approach offers a more flexible interpretation of the universal common ancestor compatible with these competing theories. The term "pangenome" describes the set of all genes present in a collection of genomes, which includes the core genome (genes present in all genomes) and accessory genes (partially conserved genes comprising the gene-level variability across the genomes). While pangenomes are typically defined from genomes of a single species, the perspective is applicable to any collection of related genomes. The LUCA core genome would represent the set of genes present in all individual LUCA organisms, whether they be members of a single ancestral species or a community of primitive cells. Furthermore, as mobile genetic elements are typically found outside the core genome of a species [11], reconstructing the evolutionary history of core genomes would likely be less confounded by horizontal gene transfer events than using genomes of individual organisms, without requiring predefined marker genes that can limit and bias such analyses [2].

To this end, we apply the following methods to 54,085 genomes spanning 183 species to reconstruct the last bacterial common ancestor (LBCA) core genome. 1) Pangenome construction for each species through protein sequence clustering as previously described [12], 2) a novel core gene identification algorithm based on estimating gene frequencies and genome assembly error rates that maximize the likelihood of observed pangenomes, and 3) reconstruction of ancestral core genomes from the 183 modern core genomes by applying asymmetric Wagner parsimony [13] (minimizing the number of gene gain/loss events between species) to the Web of Life phylogeny [2]. We find that the LBCA core genome, while distinct from any one modern core genome, contains many complete or nearly complete biological systems, especially those related to ribosomal

proteins, translation machinery, and metabolic pathways of central carbon metabolism or *de novo* biosynthetic pathways for nucleotides and amino acids.

5.4 Results

5.4.1 Construction of 183 pangenomes across the Web of Life phylogenetic tree

183 bacterial species as defined by GTDB [14] with at least 50 high quality genomes were identified from the Web of Life (WOL) dataset [2], totaling 54,085 genome assemblies (Dataset D.1, see Methods). To integrate species-level properties with the WOL phylogenetic tree which has genomes as leaves, a subtree with species as leaves was extracted as follows. For each species, the most recent common ancestor (MRCA) node of all its genomes was identified, and the edge of the MRCA was extended by the median distance from the MRCA to the species' selected genomes to model a modern representative of that species (Fig. 5.1a). The subtree and representative node assignments are available in Dataset D.2.

Next, open reading frames were annotated in all genomes using Prodigal v2.6.3 [15], and pangenomes were constructed for each species as previously described [12] based on clustering protein sequences with CD-HIT v4.6 [16]. This yielded for each species a sparse binary matrix of presence/absence calls between each genome and gene which could be associated with the species' corresponding node in the phylogenetic tree (Fig. 5.1b). These species represent broad ranges in several genetic properties, computed from genome collections ranging from 50 to 4,930 genomes per species. Species-wide averages in GC content ranged from 28.0 to 73.4%, genome lengths from 0.82 to 8.39 Mb, and number of genes per genome from 902 to 7,515 (Fig. D.1a). Strong correlation was observed between genome length and number of genes (Pearson $r = 0.991$) with a cross-species average and standard deviation of 1080 ± 68 bp per gene, and moderate correlation between

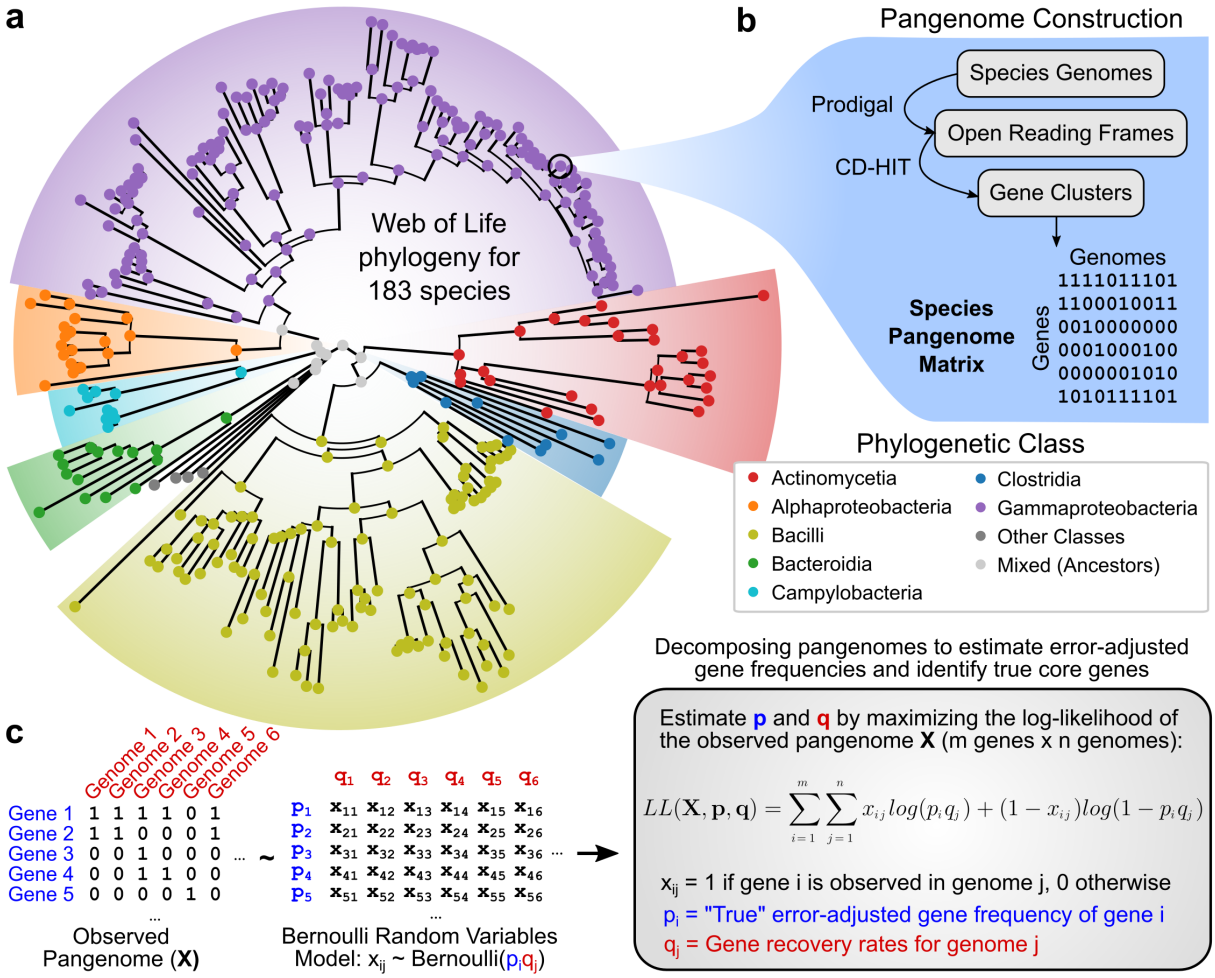


Figure 5.1. Phylogenetic distribution of 54,085 genomes across 183 species selected for core genome analysis. (a) Phylogenetic tree of species' representatives with respect to class, with branch lengths derived from the Web of Life tree. Representatives are based on the most recent common ancestors of selected species' genomes in the original tree. (b) Pangenome construction workflow applied to each species to encode species-wide gene-level variation as binary matrices. (c) Statistical model for estimating each genome's gene recovery rate (fraction of genes recovered in the final genome assembly) and each gene's true frequency from species' pangenome matrices.

either property and GC content ($r = 0.576$ and 0.544, respectively) (Fig. D.1b). Overall, the selected genomes and species capture a substantial degree of genetic diversity in the bacterial domain.

5.4.2 A model for estimating error-adjusted gene frequencies and identifying core genes

In conducting pangenome analyses, a species' core genome is defined as the set of genes observed in all its strains. However, due to artifacts in sequencing, assembly, annotation, or clustering, some genes are invariably lost during pangenome construction and a direct application of this definition is too strict to be meaningful for large genome collections (Fig. D.2a). Consequently, many pangenome studies relax this requirement by allowing core genes to be missing from some fraction of the total number of genomes, such as up to 1% by default in the Roary pipeline [17], or up to 5% for "soft-core" genes in the GET_HOMOLOGUES pipeline [18]. However, this proportional cutoff approach assumes that the rate at which genes are lost per genome is the same for all genomes, regardless of the experimental or analytical techniques used.

To address this issue, we developed a model for estimating genome-specific gene recovery rates (fraction of genes not lost to errors and recovered by the final genome assembly) and "true" error-adjusted gene frequencies (fraction of genomes carrying the gene) from a gene by genome presence/absence matrix. Denoting the true frequency of gene i as p_i and the gene recovery rate of genome j as q_j, we model the presence/absence of gene i in genome j as a Bernoulli random variable with probability $p_i q_j$ for all genes and genomes, assuming for simplicity that the two rates are independent. The p_i and q_j are then estimated simultaneously by maximizing the log-likelihood of the observed presence/absence matrix (Fig. 5.1c).

Applied to the 183 species' pangenomes, we find that this approach yields robust estimates consistent with existing methods for assessing genome quality and current

understanding of core genomes: 1) genome-specific gene recovery rates were correlated with common assembly quality metrics, especially CheckM completeness (Fig. D.2b-d), 2) estimated frequencies were robust to sampling and consistent with the existence of true core genomes (Fig. D.2e, Fig. D.3a-b), 3) estimated frequency distributions had a more consistent functional form than raw observed distributions (Fig. D.3c-e), and 4) core genomes identified using estimated frequencies were of similar size to those defined using earlier approaches but with meaningfully different gene content (Fig. D.4). Details are available in the Supplemental Discussion and Dataset D.3, and the estimated frequencies for each gene and species are available in Dataset D.4.

5.4.3 Core genome content is relatively stable across species at the level of functional categories but not individual orthogroups

Taking the genes with estimated frequency >99.99% as our definition of each species' core genome, functional annotations and orthogroup (OG) assignments were generated for all species' core genes by applying eggNOG-emapper v2.1.6-43 [19] to each gene's most common sequence variant. We find that while core genome size varies significantly across species (214 to 4,766 genes) and relatively independently of phylogenetic placement (Fig. 5.2a), the allocation of these genes to different COG functional categories is stable (Fig. 5.2b-c); principal component analysis (PCA) of these core gene function distributions was unable to distinguish different phylogenetic classes (Fig. D.5). Seven COG categories comprised at least 5% of all core genomes on average: S (18.4%; Function unknown), J (8.9%; Translation, ribosomal structure and biogenesis), E (6.6%; Amino acid transport and metabolism), K (6.6%; Transcription), C (5.7%; Energy production and conversion), P (5.4%; Inorganic ion transport and metabolism), and M (5.3%; Cell wall/membrane/envelope biogenesis).

However, while the distribution of core genes to functional categories was stable

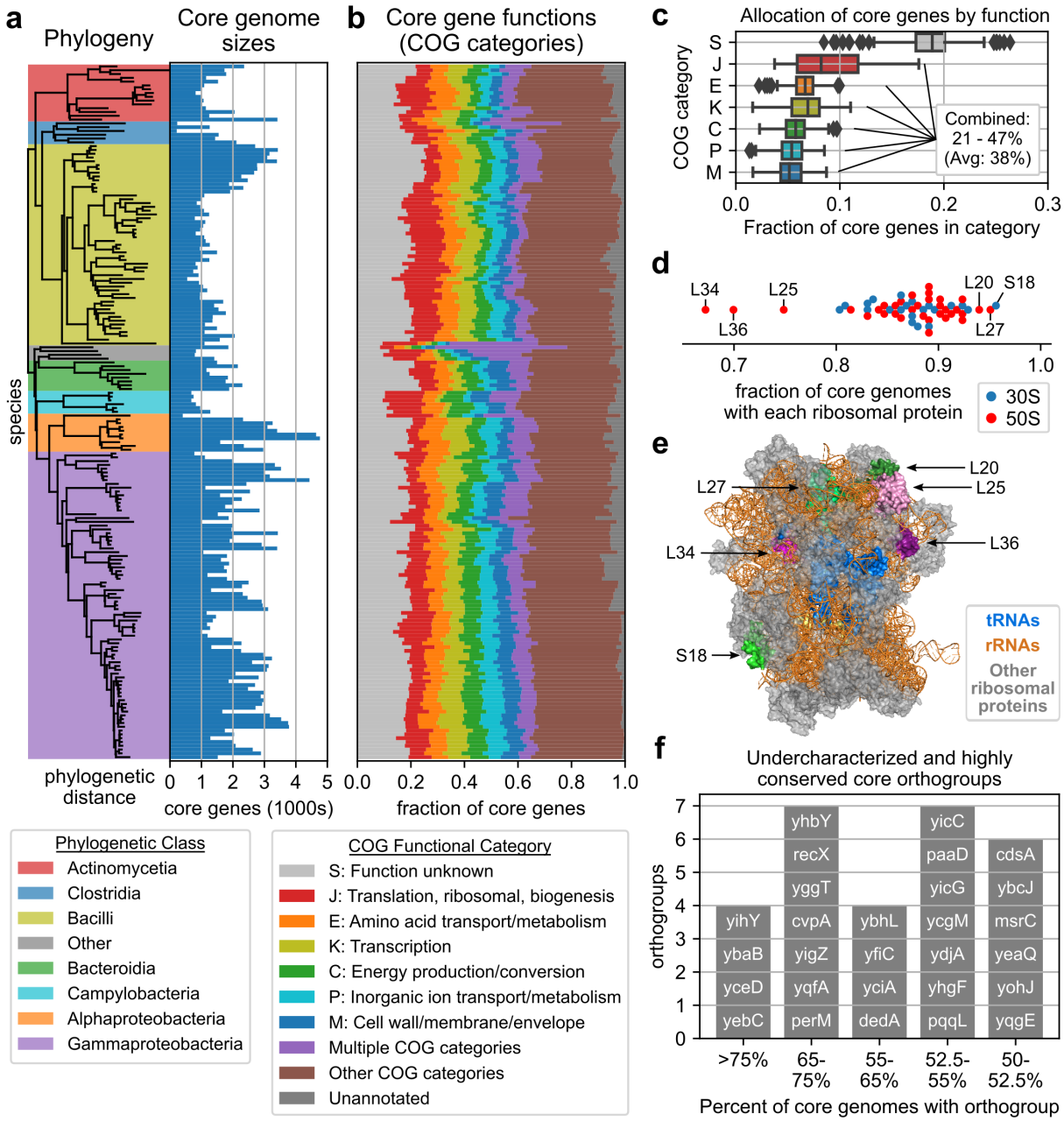


Figure 5.2. Distribution of gene functions and orthogroups across the core genomes of 183 species. (a) Core genome sizes of each species compared to the phylogenetic tree. (b) Distribution of each species' core genes to COG functional categories. COG categories with average frequency >5% are shown, with all others grouped together. (c) Distribution of each species' core genes to most common functional categories. (d) Fraction of core genomes with each bacterial ribosomal protein. (e) Relative positions of the ribosomal proteins appearing in the largest/smallest number of core genomes (structure adapted from PDB:7K00). (f) Under-characterized orthogroups observed in at least 50% of all core genomes.

across species, no individual OG was found universally in all core genomes. The most strongly conserved OG corresponding to *rpsR* (ribosomal protein S18) was found in 175/183 core genomes, just 28 OGs were found in >90% of core genomes, and 156 OGs were found in >80% of core genomes (Fig. D.6a). These strongly conserved core OGs were dominated by ribosomal proteins and other translation-associated genes, with just six in >90% of core genomes not translation-associated: *ruvA* (resolvasome RuvABC subunit), *nusB* (transcription antitermination protein), *gmk* (guanylate kinase), *atpE* (FoF1-type ATP synthase subunit K), *atpH* (FoF1-type ATP synthase delta subunit), and *yajC* (protein translocase subunit). Additionally, this lack of universal core OGs was not sensitive to the 99.99% frequency threshold. No OGs were found in all core genomes even at a 99% threshold, and absences of highly conserved OGs were distributed across a wide range of species (Fig. D.7, see Supplemental Discussion).

Ribosomal proteins were among the most strongly conserved core OGs, with all but three observed in at least 80% of core genomes (L25, L36, L34) (Fig. 5.2d). On the *Escherichia coli* ribosome (PDB: 7K00) [20], L25 and L36 are in close proximity to each other (23Å) and L25 is in contact with both the 5S rRNA and L16, with the latter directly contacting A and P site tRNAs. L34 is embedded within the 50S subunit, interacting primarily with the 23S rRNA and further away from the tRNAs (34Å to A and P site tRNAs, 40Å to E site tRNA) (Fig. 5.2e). The reduced conservation of these ribosomal proteins is consistent with previous experiments suggesting their non-essentiality, with viable deletion mutants having been generated for L25 and L36 in *E. coli* and *Bacillus subtilis* and for L34 in *B. subtilis* only [21].

Finally, a number of highly conserved core OGs were under-characterized and likely comprise an integral but poorly understood component of bacterial genomes. 72 OGs corresponding to y-genes or annotated with the "S: Function unknown" category were found in >50% of core genomes, of which 28 had UniProt annotation level ≤ 3 (Fig. 5.2f, Fig. D.6b-c) and four had entirely no functional annotation: *yihY*, *yicG*, *yicC*, and

yeaQ (Table D.1). Overall, we find that while the allocation of core genes to functional categories is similar across species, core genomes appear highly diverse at the level of individual OGs, with several under-characterized OGs consistently present across most species. Annotations for conserved OGs are available in Dataset D.4.

5.4.4 Reconstruction of the last bacterial common ancestor core genome

We next used these 183 core genomes to reconstruct the core genome of the LBCA based on the species-level Web of Life subtree extracted earlier. Reconstruction was carried out using Count [22] under asymmetric Wagner parsimony [13], using the core OG by species presence/absence table as input. Briefly, given a set of genes for each modern species, this approach assigns genes to each ancestral species such that the number of gene gain/loss events is minimized, where the relative penalty for a gain vs. loss event can be adjusted. As the appropriate gain/loss penalty ratio is not known for core genomes, multiple reconstructions were conducted with different ratios ranging from 0.05 to 2.0. OGs recovered at smaller ratios (and consequently smaller, more stringently defined core genomes) were treated as more likely to be present in ancestral core genomes, similar to a previous interpretation of the gain/loss ratio [23].

The size of the LBCA core genome grew steadily with the gain/loss penalty ratio (g), with the minimum ratio at which an OG was observed in the LBCA core genome correlated with the fraction of modern core genomes with the OG ($p = -0.814$, Fig. D.8a). A large jump in core genome size was observed between $g = 1$ and $g = 1.05$ as the model switches from preferring gain events to loss events (Fig. 5.3a). The first LBCA core OG was observed at $g = 0.25$ (*rplT*, 50S ribosomal protein L20), and the size of the LBCA core genome exceeded 100 OGs at $g = 0.55$ and 1,000 OGs at $g = 1.55$. OGs recovered at the smallest ratios were again mostly translation-associated, which remained the most common functional category for all ratio values (excluding S: Function unknown). Other

functional categories representing at least 5% of the LBCA core genome at $g = 2.0$ were primarily metabolic, involving amino acid (E), coenzyme (H), energy (C), nucleotide (F), or inorganic ion (P) metabolism. Other highly represented categories included "M: Cell wall, membrane, envelope biogenesis" and "L: Replication, recombination and repair".

5.4.5 Functional analysis suggests a versatile LBCA core genome

Annotation of these OGs with biological systems from the 2020 COG database [24] revealed that certain systems were nearly entirely present in the LBCA core genome. Among systems with at least 10 known OGs, 21 had over half of their known OGs present in the LBCA core genome at $g = 2.0$ and at least one at $g = 1.0$, and were again primarily translation-associated or metabolic systems, especially those related to nucleotide or amino acid biosynthesis (Fig. 5.3b, Fig. D.8b). Eight such systems had >90% of their known OGs present in the LBCA core at $g = 2.0$: Mureine biosynthesis, Ribosome 30S subunit, Ribosome 50S subunit, Arginine biosynthesis, Pyrimidine salvage, Ile/Leu/Val biosynthesis, Pyrimidine biosynthesis, and Histidine biosynthesis. Eight systems also retained >80% of their OGs at the more stringent $g = 1.0$ threshold: Mureine biosynthesis, Ribosome 30S subunit, Ribosome 50S subunit, Pyrimidine biosynthesis, FoF1-type ATP synthase, 16S rRNA modification, Translation factors, and Aminoacyl-tRNA synthetases. LBCA OG thresholds and annotations are available in Dataset D.4.

We examined four systems associated with translation with many OGs present in the LBCA core at stringent gain/loss penalty ratios: 50S ribosomal proteins, 30S ribosomal proteins, aminoacyl-tRNA synthetases (AARSs), and translation factors. Most OGs in these systems are present at $g = 1.0$ and highly likely to be in the LBCA core genome: 30/34 50S ribosomal proteins, 21/21 30S ribosomal proteins, 17/20 AARSs, and 9/11 bacterial translation factors (Fig. 5.3c). All but one of the nine OGs missing at $g = 1.0$ were present at $g = 2.0$, and only *rpl7Ae* (ribosomal protein L7Ae) was missing

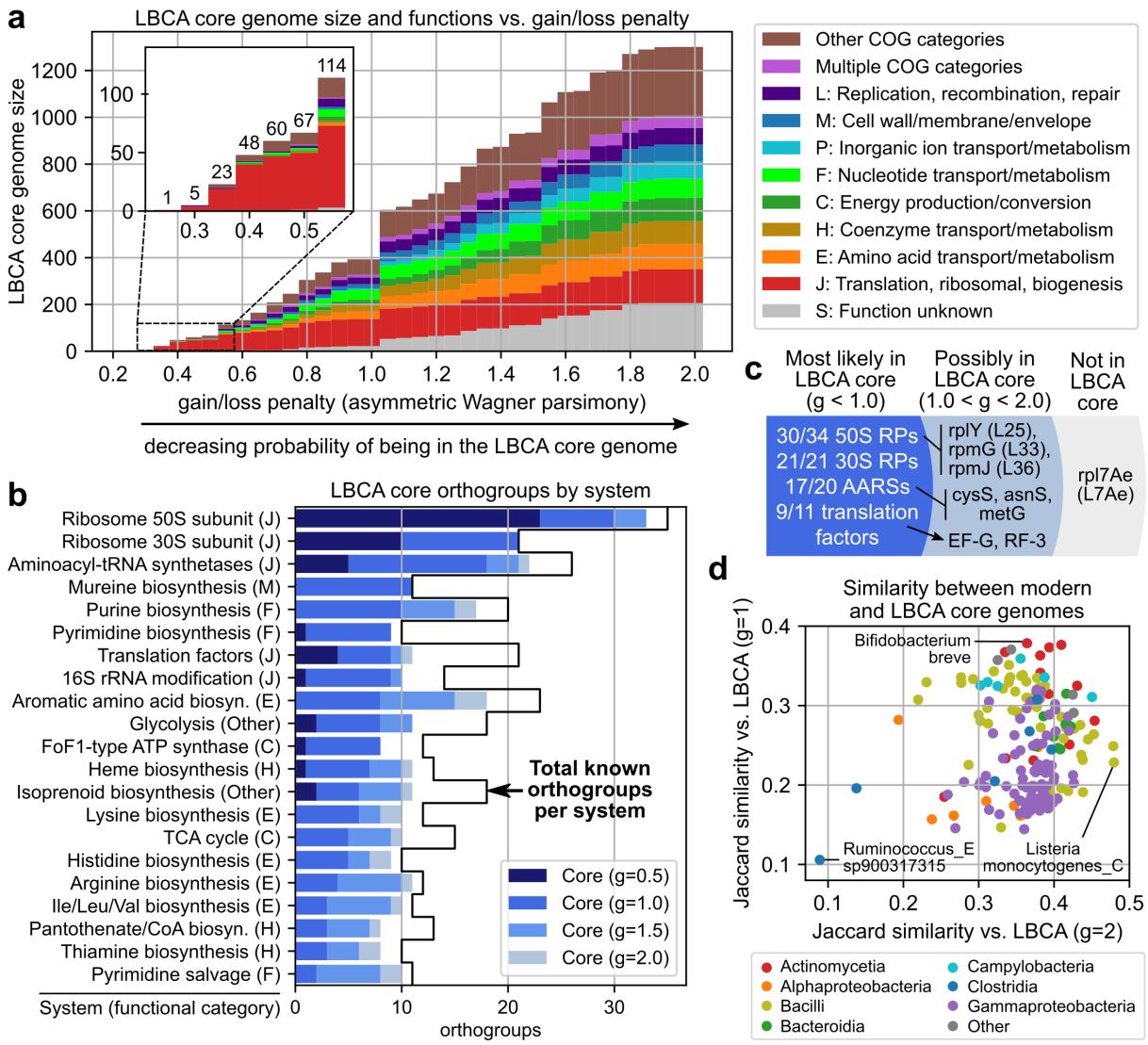


Figure 5.3. Distribution of gene functions in the core genome of the last bacterial common ancestor. (a) Size and functions of the last bacterial common ancestor (LBCA) core genome as the gain/loss penalty ratio (g) is increased during ancestral reconstruction through asymmetric Wagner parsimony. Functional categories comprising at least 5% of the core genome at $g = 2.0$ are shown. (b) Distribution of LBCA core orthogroups by biological system and minimum gain/loss penalty to be observed. Systems with at least 10 orthogroups of which at least 50% are present in the LBCA core genome at $g = 2.0$ and at least one at $g = 1.0$ are shown. (c) Distribution of LBCA core orthogroups related to ribosomal proteins (RPs), aminoacyl-tRNA synthetases (AARSs), and translation factors. (d) Jaccard similarity between modern and LBCA core genomes at $g = 1$ and $g = 2$, by species phylogenetic class. The core genomes least and most similar to LBCA for $g = 1$ and $g = 2$ are labeled.

entirely. The eight marginal core OGs consisted of 50S ribosomal proteins *rplY* (L25), *rpmG* (L33), and *rpmJ* (L36) observed at $g = 1.05$ (Fig. D.9a); AARSs *cysS* (CysRS, observed at $g = 1.1$), *asnS* (AsnRS, $g = 1.3$) and *metG* (MetRS, $g = 1.6$) (Fig. D.9b); and translation factors *fusA* (EF-G, $g = 1.05$) and *prfC* (RF3, $g = 1.6$) (Fig. D.9c). Most of these OGs are known to be non-essential, missing in some bacterial and/or archaeal species, and/or have known compensatory paralogs or pathways (Table D.2). Only *metG* and *fusA* have no previous studies that suggest their absence from LBCA but are known to have complex evolutionary histories resulting in highly diverse sequences [25, 26] and potentially incomplete coverage by a single OG as recoverable by eggNOG. Nonetheless, ancestral reconstruction from core genomes seems to suggest that most, if not all of these translation-related genes traditionally thought of as universal are indeed likely to be present in not just individual LBCA genomes but in the LBCA species-wide core genome.

Analyzing OGs assigned to the "S: Function unknown" category or corresponding to y-genes revealed 108 potentially under-characterized OGs in the LBCA core genome at $g = 2.0$, 10 of which were present at $g = 1.0$ (Fig. D.8c). All but one of these 10 likely core OGs, *yqxC*, was observed in at least 50% of modern core genomes and among the highly conserved under-characterized core OGs reported earlier. Finally, comparisons between LBCA and modern core genomes finds LBCA to be genetically distinct from all 183 modern species in this analysis. Jaccard similarities between the LBCA core genome (at either $g = 1$ or $g = 2$) and modern core genomes did not exceed 0.5, though core genomes from certain phylogenetic classes were more similar to that of LBCA than others (Fig. 5.3d). Campylobacteria had the highest median similarity at $g = 2$, Bacteroidia at $g = 1$, and Actinomycetia was the 2nd most similar for both thresholds.

5.4.6 Comparison against minimal organism JCVI-Syn3A suggests that the LBCA core genome alone is not sufficient for viability

To explore how close the LBCA core genome is to a minimal set of genes required for life, we compared its gene content with the genome of JCVI-Syn3A (Syn3A), a synthetic minimal organism of 493 genes (452 protein-coding) derived from *Mycoplasma mycoides capri* (GenBank: CP016816.2) [27]. Annotation with eggNOG-emapper was able to identify 430 unique OGs in Syn3A. 325 of these OGs (76%) were present in the LBCA core genome at its largest extent at $g = 2.0$, with a majority recovered at $g = 1.0$ (50%, 217/430) (Fig. D.10a, Dataset D.4). Conversely, LBCA core OGs present at the most stringent penalty ratios are nearly all present in Syn3A, but the fraction of LBCA OGs in Syn3A drops to 55% (217/393) by $g = 1.0$ and 25% (325/1301) by $g = 2.0$ (Fig. D.10a). An additional 38 Syn3A OGs were observed in at least one of the 183 modern core genomes but not in the LBCA core genome, and the remaining 67 OGs were not observed in any core genome (Fig. 5.4a).

Comparing the LBCA core genome, the Syn3A genome, and the core genome of related small genome species *Mycoplasmoides pneuomiae* also finds extensive but not complete overlap between the three gene sets (Fig. D.10b). 13 OGs were in both the core genome of *M. pneumoniae* and the Syn3A genome but not in the LBCA core genome, while 92 Syn3A OGs were missing from both core genomes. Functional annotation found most of the 325 OGs shared by LBCA and Syn3A to be well-characterized and primarily ribosomal or metabolic, while the 105 non-LBCA OGs in Syn3A were mostly under-characterized proteins (77/105) with the remaining 28 OGs being of diverse functions (Fig. 5.4b). These results suggest that most of the genes in the minimal organism Syn3A likely descended from the LBCA core genome, but given the requirement of additional non-core genes for viability, also suggest that the LBCA core genome alone may be insufficient for survival.

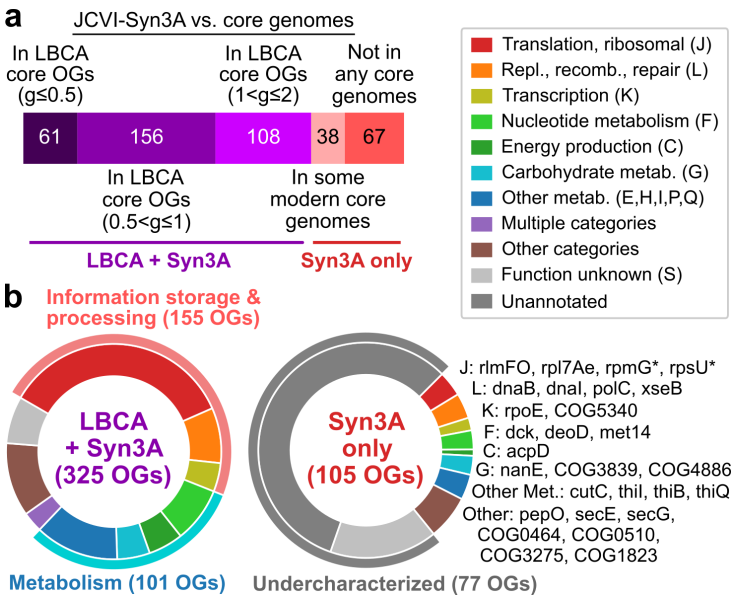


Figure 5.4. Similarities between the LBCA core genome and the genome of minimal organism JCVI-Syn3A. (a) Distribution of Syn3A orthogroups (OGs) in the LBCA and modern core genomes. Modern core genomes refer to those previously constructed for 183 bacterial species. (b) Distribution of functional categories among OGs shared by the LBCA core genome ($g = 2$) and Syn3A, and those in Syn3A only. Starred OGs refer to alternate variants of OGs present in the LBCA core genome.

5.4.7 Pathway analysis suggests a highly metabolically self-sufficient LBCA core genome

To examine the metabolic capabilities of the LBCA core genome at the pathway level, COG orthogroups were mapped to KEGG orthogroups [28] through eggNOG-emapper annotations and then to KEGG modules (Dataset D.5). 100 and 175 KEGG modules had at least one reaction active based on KEGG orthogroups present at $g = 1.0$ and $g = 2.0$, respectively, primarily from five categories: amino acid metabolism, cofactor/vitamin metabolism, carbohydrate metabolism, energy metabolism, and nucleotide metabolism (Fig. 5.5a). Reconstruction of modules in the last three categories revealed that much of central carbon metabolism and biosynthetic pathways to all major nucleotides were largely intact even at $g = 1.0$ (Fig. 5.5b).

Just three components in central carbon metabolism were missing from the LBCA

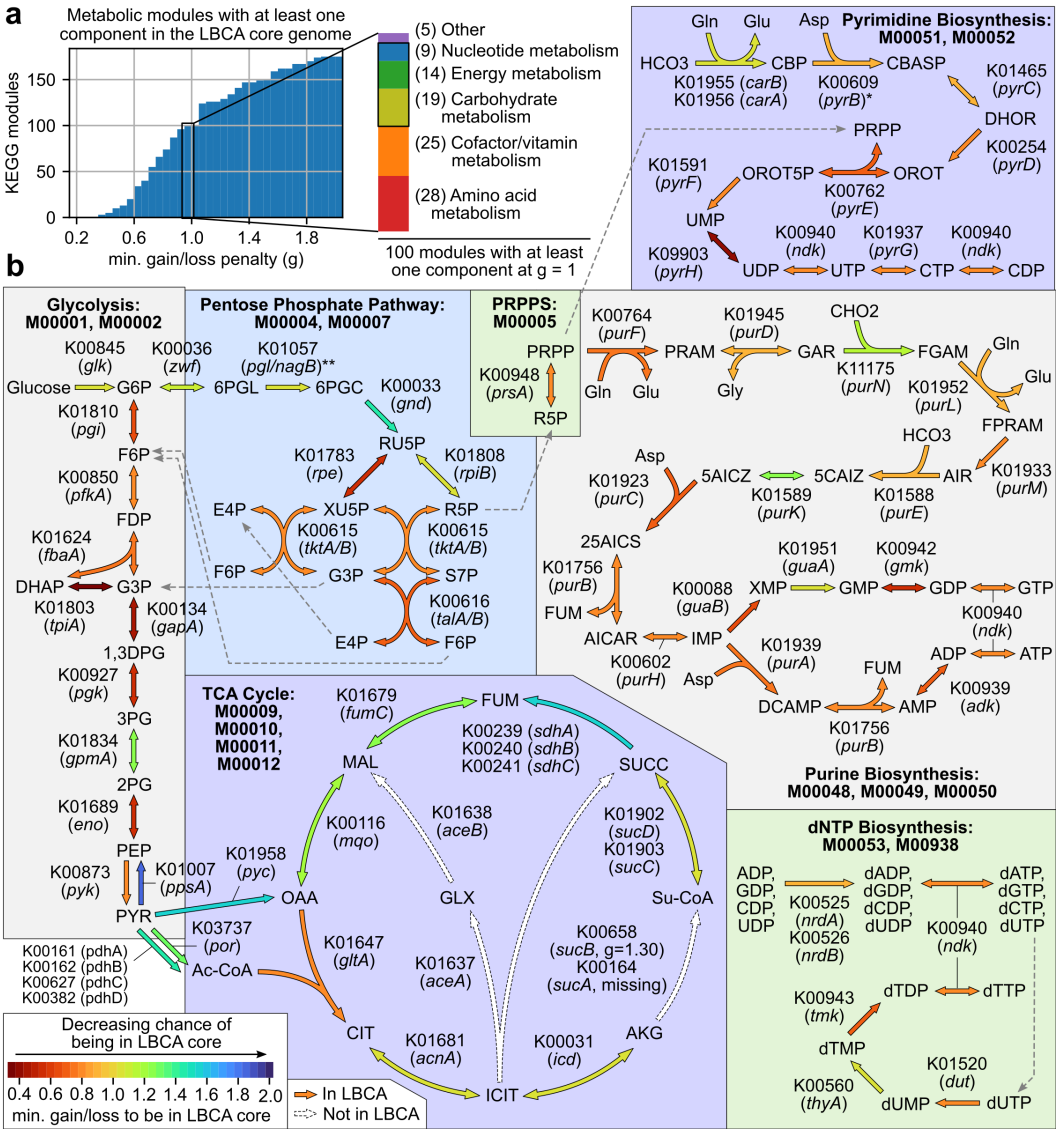
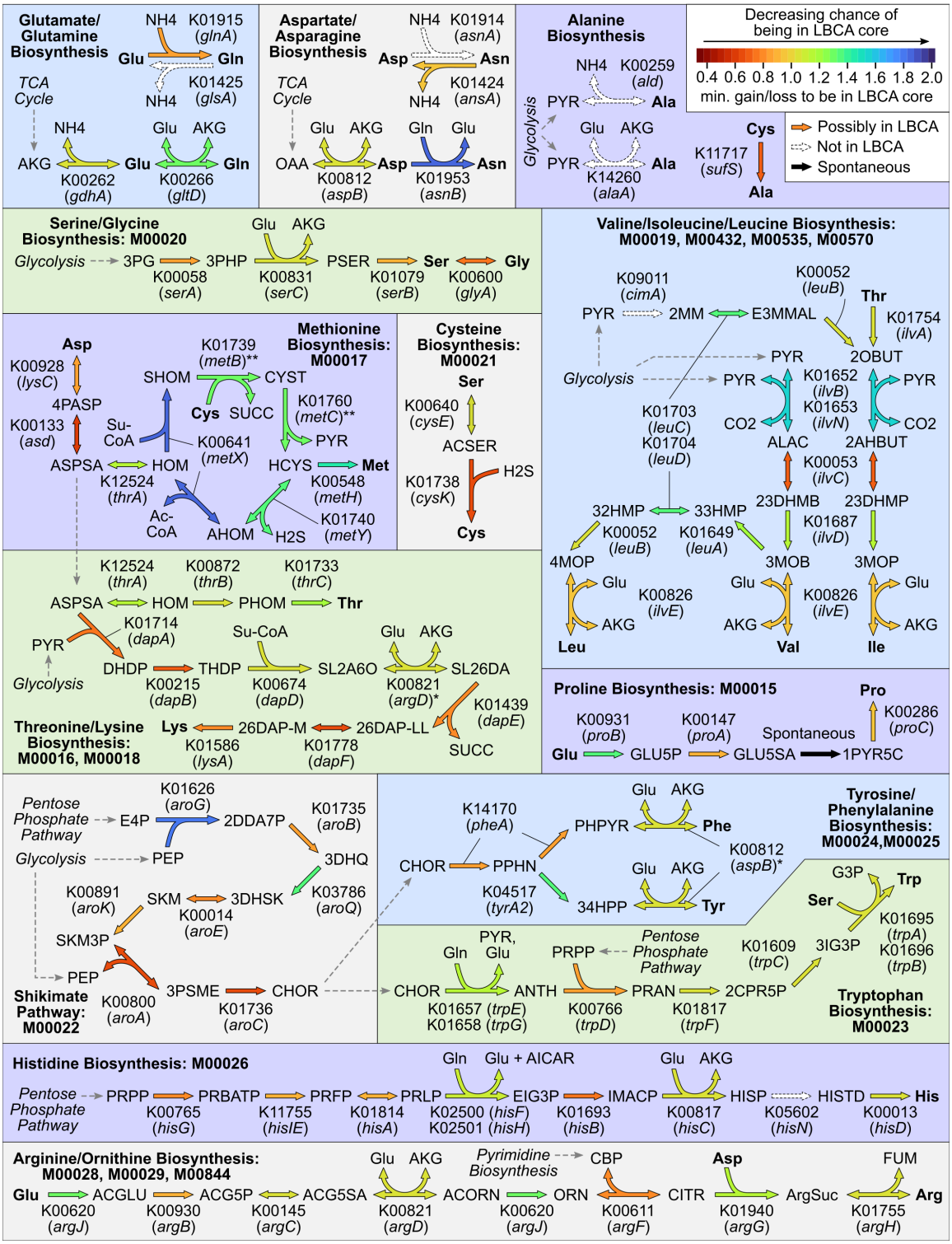


Figure 5.5. Metabolic modules involving central carbon metabolism or nucleotide metabolism represented in the LBCA core genome. (a) Distribution of KEGG modules with at least one related orthogroup (OG) present in the LBCA core genome for varying gain/loss penalty thresholds (g). Module categories at the $g = 1$ threshold or shown, with visualized modules boxed. (b) Individual reactions in the LBCA core genome related to modules involving central carbon metabolism or nucleotide metabolism. Each reaction is labeled with the corresponding KEGG OG(s) and gene name(s), and is colored by the minimum g at which the corresponding OG is present in the LBCA core (or maximum of all such g if multiple OGs are required). Missing reactions are colored white. Dashed gray arrows connect identical metabolites. For the starred reaction corresponding to K00609 (*pyrB*), the regulatory subunit K00610 (*pyrI*) is missing in LBCA but not required for catalytic activity. For the double starred reaction corresponding to K01057 (*pgl*), the closest COG orthogroup COG0363 represents sequences from both *pgl* and *nagB*. Abbreviations are available in Dataset D.5.

core genome: 1) *sucA*, the E1 component of the α-ketoglutarate dehydrogenase complex in the TCA cycle, 2) *aceA*, isocitrate lyase in the glyoxylate cycle, and 3) *aceB*, malate synthase in the glyoxylate cycle. Several reactions were only active at high g: two mechanisms of generating acetyl-CoA from pyruvate (pyruvate dehydrogenase complex and pyruvate-ferredoxin oxidoreductase) and the second half of the TCA cycle (succinate to oxaloacetate) were only active at $g = 1.2$ or higher. In nucleotide metabolism, *de novo* biosynthetic pathways to all nucleotides (A,C,G,U) and deoxyribonucleotides (A,C,G,T,U) were active in the LBCA core genome with most reactions present at $g = 1.0$. The only missing OG was *pyrI*, the regulatory subunit of aspartate carbamoyltransferase (the catalytic subunit *pyrB* is independently active [29]).

A similar reconstruction of amino acid metabolism finds *de novo* biosynthetic pathways to all 20 amino acids to be mostly intact, with most reactions present at $g = 1.2$ and no more than one reaction missing per pathway (Fig. 5.6). Six components were missing from the LBCA core genome: 1) *glsA*, glutaminase, 2) *asnA*, aspartate ammonia ligase, 3) *ald*, alanine dehydrogenase, 4) *alaA*, alanine transaminase, 5) *cimA*, citramalate synthase, and 6) *hisN*, histidinol phosphatase. Biosynthetic pathways towards lysine, phenylalanine, and tyrosine also relied on promiscuous aminotransferase activity from other pathways as annotated on KEGG (*argD* and *aspB* in place of *dapC* and *tyrB*, respectively). All but one of these absences either did not impact overall biosynthesis pathways or could be compensated by other pathways. Orthogroups *glsA* and *asnA* are not required in the biosynthesis of glutamate/glutamine and aspartate/asparagine from TCA cycle intermediates. Alanine could be generated from cysteine via *sufS* in the absence of *ald* and *alaA*, and the isoleucine precursor 2-oxobutanoate could be generated from threonine via *ilvA* in the absence of *cimA*. The absence of *hisN* did not have a clear compensatory pathway, though it is possible that the missing reaction could be catalyzed by another phosphatase or that LBCA may have had a bifunctional *hisB* capable of carrying out the reaction as observed in Enterobacteria [30].

Figure 5.6. Metabolic modules involving amino acid biosynthesis represented in the LBCA core genome. Each reaction is labeled with the corresponding KEGG orthogroup (OG) and gene name(s), and is colored by the minimum penalty ratio (g) at which the corresponding OG is present in the LBCA core (or maximum of all such g if multiple OGs are required). Missing reactions are colored white and spontaneous reactions are colored black. Amino acid inputs and outputs are bold. The biosynthesis pathways of alanine, aspartate, asparagine, glutamate, and glutamine do not have specific KEGG modules and have been added separately. Starred reactions indicate reactions that can be catalyzed by enzymes in unrelated pathways as the standard enzyme is missing from the LBCA core genome: In Lysine biosynthesis, K00821 (*argD*) can catalyze SL2A60 $\rightarrow$ SL26DA in place of K14267 (*dapC*), and in Tyrosine/Phenylalanine biosynthesis, K00812 (*aspB*) can catalyze the final steps in both pathways in place of K00832 (*tyrB*). For double starred reactions corresponding to *metB* and *metC*, the COG orthogroup COG0626 present at $g = 1.3$ represents sequences from both genes. Abbreviations are available in Dataset D.5.


Glutamate/ Glutamine Biosynthesis
NH4 K01915 (glnA)
Glu
Gln
K01425 (glsA)
NH4
TCA Cycle
NH4
Glu AKG
AKG
Glu
Gln
K00262 (gdhA)
K00266 (gltD)
Aspartate/ Asparagine Biosynthesis
NH4 K01914 (asnA)
Asp
Asn
K01424 (ansA)
NH4
TCA Cycle
Glu AKG
Gln Glu
OAA
K00812 (aspB)
Asp
K01953 (asnB)
Asn
Alanine Biosynthesis
NH4 K00259 (ald)
PYR
Ala
Glycolysis
Glu AKG
PYR
K14260 (alaA)
Ala
Cys
K11717 (sufS)
Ala
Decreasing chance of being in LBCA core
0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0
min. gain/loss to be in LBCA core
Possibly in LBCA
Not in LBCA
Spontaneous
Serine/Glycine Biosynthesis: M00020
Glycolysis
3PG
K00058 (serA)
3PHP
Glu AKG
K00831 (serC)
PSER
K01079 (serB)
Ser
K00600 (glyA)
Gly
Methionine Biosynthesis: M00017
Asp
K00928 (lysC)
4PASP
K00133 (asd)
ASPSA
K12524 (thrA)
HOM
K01739 (metB)**
SHOM
CYST
Cys
SUCC
K01760 (metC)**
PYR
Su-CoA
K00641 (metX)
HCYS
Met
K00548 (metH)
Ac-CoA
AHOM
H2S
K01740 (metY)
Cysteine Biosynthesis: M00021
Ser
K00640 (cysE)
ACSER
K01738 (cysK)
H2S
Cys
Valine/Isoleucine/Leucine Biosynthesis: M00019, M00432, M00535, M00570
K09011 (cimA)
PYR
2MM
E3MMAL
K00052 (leuB)
Thr
K01754 (ilvA)
PYR
2OBUT
Glycolysis
PYR
K01652 (ilvB)
K01653 (ilvN)
PYR
CO2
CO2
K01703 (leuC)
K01704 (leuD)
ALAC
2AHBUT
K00053 (ilvC)
23DHMB
23DHMP
32HMP
33HMP
K01687 (ilvD)
K00052 (leuB)
K01649 (leuA)
4MOP
3MOB
3MOP
Glu
K00826 (ilvE)
Glu
Glu
AKG
AKG
AKG
Leu
Val
Ile
Threonine/Lysine Biosynthesis: M00016, M00018
K12524 (thrA)
K00872 (thrB)
K01733 (thrC)
ASPSA
HOM
PHOM
Thr
K01714 (dapA)
PYR
Glycolysis
Su-CoA
Glu AKG
DHDP
THDP
SL2A6O
SL26DA
K00215 (dapB)
K00674 (dapD)
K00821 (argD)*
K01439 (dapE)
Lys
26DAP-M
26DAP-LL
SUCC
K01586 (lysA)
K01778 (dapF)
Proline Biosynthesis: M00015
Pro
K00931 (proB)
K00147 (proA)
Spontaneous
K00286 (proC)
Glu
GLU5P
GLU5SA
1PYR5C
Shikimate Pathway: M00022
Pentose Phosphate Pathway
E4P
K01626 (aroG)
2DDA7P
K01735 (aroB)
Glycolysis
PEP
3DHQ
K00891 (aroK)
SKM
3DHSK
K03786 (aroQ)
K00014 (aroE)
SKM3P
PEP
K00800 (aroA)
3PSME
K01736 (aroC)
CHOR
Tyrosine/ Phenylalanine Biosynthesis: M00024,M00025
K14170 (pheA)
CHOR
PPHN
PHPYR
Glu AKG
Phe
K00812 (aspB)*
K04517 (tyrA2)
34HPP
Glu AKG
Tyr
Tryptophan Biosynthesis: M00023
G3P
Trp
Ser
K01695 (trpA)
K01696 (trpB)
PYR, Glu
Gln Glu
PRPP
Pentose Phosphate Pathway
K01609 (trpC)
3IG3P
CHOR
ANTH
PRAN
2CPR5P
K01657 (trpE)
K01658 (trpG)
K00766 (trpD)
K01817 (trpF)
Histidine Biosynthesis: M00026
Pentose Phosphate Pathway
PRPP
K00765 (hisG)
PRBATP
K11755 (hisIE)
PRFP
K01814 (hisA)
PRLP
Gln
Glu + AICAR
K02500 (hisF)
K02501 (hisH)
EIG3P
K01693 (hisB)
IMACP
Glu AKG
K00817 (hisC)
HISP
K05602 (hisN)
HISTD
K00013 (hisD)
His
Arginine/Ornithine Biosynthesis: M00028, M00029, M00844
Glu
ACGLU
K00620 (argJ)
ACG5P
K00930 (argB)
ACG5SA
K00145 (argC)
Glu AKG
K00821 (argD)
Pyrimidine Biosynthesis
ACORN
K00620 (argJ)
ORN
CBP
K00611 (argF)
CITR
Asp
K01940 (argG)
ArgSuc
FUM
K01755 (argH)
Arg

5.5 Discussion

The vast amount of publicly available genomic data has enabled the reconstruction of evolutionary histories spanning entire domains of life, but has also opened new questions regarding the appropriate structure of such phylogenies and the nature of the ancestors at their roots. With a pangenomic perspective compatible with competing theories on universal ancestry, we reconstructed and characterized the core genome of the LBCA from the pangenomes of 183 species with three tools: 1) pangenome construction by sequence clustering, 2) core gene identification through a novel model for estimating gene frequencies and genome-specific gene recovery rates, and 3) ancestral genome reconstruction with asymmetric Wagner parsimony. We find that the 183 modern core genomes vary significantly at the level of individual genes but similarly allocate genes to different functional categories, and that the LBCA core genome was versatile with many systems related to translation machinery and biosynthetic pathways fully intact.

We first addressed the issue of sequencing artifacts and other errors leading to genes being lost during pangenome construction and confounding the identification of core genes. Rather than adopting an arbitrary cutoff of allowing core genes to be missing from a certain percentage of genomes as was common in previous studies, we estimated the underlying true gene frequencies by modeling gene presence/absence calls in observed pangenomes as Bernoulli random variables and applying maximum likelihood estimation. Applied to the 183 pangenomes, several results support the robustness and reliability of this approach. Estimated genome-specific recovery rates were correlated with existing measures of genome assembly quality and especially with CheckM's completeness metric. Estimated gene frequencies were robust to subsampling, consistent with the existence of true core genomes, had a consistent distribution shape, and yielded core genomes of similar size to those from previous studies. These results suggest that gene frequency estimation can identify core genes in a manner more faithful to their original definition.

More generally, this approach of decomposing presence/absence matrices into feature frequencies and sample recovery rates may inform future analyses of other types of binary biological data on the true distribution of observed features and sample quality.

A comparison of the 183 core genomes defined using estimated frequencies revealed that bacterial species allocate similar fractions of their core genes to specific functional categories, similar to what was observed previously for 12 pathogens [12]. In contrast, core genomes differed greatly at the gene-level with no single OG observed in all 183 core genomes, consistent with previous observations that biochemical functions rather than individual genes tend to be conserved [31]. The few genes that were nearly universally conserved across the core genomes were mostly ribosomal proteins, resembling the set of approximately 30 genes that smaller scale studies reported present in all genomes [32, 33]. We also identified 28 under-characterized genes that were strongly conserved across core genomes. These genes are ideal candidates for experimental characterization to expand the current understanding of bacterial biology, as any individual bacterial strain is likely to carry them.

We next integrated our core genomes with the Web of Life phylogeny [2] and applied asymmetric Wagner parsimony to reconstruct the core genome of the LBCA. Even under strict assumptions ($g = 1.0$), the LBCA core genome was functionally versatile with many biological systems fully or mostly intact, especially translation machinery. Just 4/56 ribosomal proteins were not present at $g = 1.0$: L25, L33, and L36 have known paralogs and have previously been found to be missing in some genomes [21, 34] and correspondingly resulted in their prediction to be in the LBCA core at the margin ($g = 1.05$), while L7ae is typically associated with archaea [35] and was correctly predicted to be missing. Similarly, just 3/20 aminoacyl-tRNA synthetases were not present at $g = 1$. Lack of CysRS and AsnRS can be compensated through alternate pathways [36, 37] and are known to be missing in some genomes [32]. However, no evidence currently exists for MetRS's absence from the LBCA, and given its critical role in the initiation of translation it is possible that

this result is an artifact of clustering and annotation. Finally, 2/11 translation factors were not present at $g = 1$: EF-G, which is believed to ancient but has known paralogs [38] was also recovered at the margin ($g = 1.05$), while RF-3's much later recovery ($g = 1.6$) is consistent with a previous prediction that RF-3 is a post-LBCA offshoot of EF-G [39]. These results confirm many previous predictions regarding the presence of fundamental biological systems in ancestral bacteria to the stronger standard of being present in the LBCA core genome.

However, a comparison with the genome of JCVI-Syn3A suggests that the LBCA core genome alone may not be viable. While the two gene sets overlap significantly, 25% of the genes in Syn3A are still missing from the LBCA core even with relaxed assumptions ($g = 2$), most of which were under-characterized. Previous multi-strain studies on gene essentiality for *Pseudomonas aeruginosa* [40] and *Streptococcus pyogenes* [41] have found core genomes to only capture a fraction of any individual strain's essential gene set, and *in vivo* genome reduction efforts have yielded organisms with significantly different sets of essential genes even when starting from strains of the same species [42, 43]. Consequently, it is likely that core genomes, including that of the LBCA, must be supplemented with accessory genes to yield a functioning organism, and is consistent with claims that core genes and essential genes are not synonymous concepts [42].

Metabolic reconstruction of the LBCA core genome provides further evidence in support of its versatility. Much of central carbon metabolism and *de novo* biosynthetic pathways to all major nucleotides and amino acids are present at $g = 1$ and fully intact at $g = 2$. These results are consistent with a previous metabolic analysis of the LBCA revealing its capability to synthesize all 4 DNA bases, 4 RNA bases, and 20 amino acids [3]. It has also been predicted that all but four amino acids could be synthesized by the ancestor of all three domains of life [44]; three of these (lysine, phenylalanine, tyrosine) could only be synthesized by our predicted LBCA core genome when assuming promiscuous activity from other aminotransferases. While most genes we predicted missing from the

LBCA core genome did not preclude the biosynthesis of essential metabolites, the most impactful was the lack of *sucA*, the E1 component of the α-ketoglutarate dehydrogenase complex. The presence of *sucB*, the E2 component, without *sucA* has been observed in *Mycobacterium tuberculosis*, which only carries an E2-like gene and exhibits no α-ketoglutarate dehydrogenase activity [45]. However, it is worth noting that the absence of these genes from the LBCA core genome does not mean that the LBCA was entirely unable to catalyze the corresponding reactions. Rather, the missing genes may have instead been part of the LBCA accessory genome, allowing some LBCA strains to carry out the reactions while others adapted alternate genetic and metabolic solutions for survival. Such intraspecies diversity in the LBCA likely contributed to the emergence of modern interspecies diversity.

The core metabolism of the LBCA also provides some clues into its contemporary environment. The absence of a direct pathway from pyruvate to alanine is compensated by *sufS* enabling the conversion of cysteine to alanine, consistent with previous theories highlighting the prevalence of sulfur chemistry in early life [46]. Comparing alternate amino acid biosynthesis pathways, the LBCA core genome contained the thermodynamically favorable pathway over the cofactor efficient pathway in all five acyl-CoA dependent pathways [47], which suggests that cofactor efficiency may not have been a major driver of selection even in the anaerobic environment of ancient bacteria. Similarly, the absence of the glyoxylate cycle, which provides a net anaplerotic reaction required for growth with certain organic acids such as acetate as sole carbon sources [48], suggests that such environments may have been too rare to establish glyoxylate cycle enzymes as part of the LBCA core genome.

Overall, the integration of pangenomics and phylogenetics enabled by the current scale of public genomic data offers a new, flexible perspective for reconstructing and characterizing ancestral species. Our analysis of 54,085 genomes and 183 species suggests that the bacterial ancestor was genetically and metabolically versatile, confirming many

previous predictions regarding genes in the LBCA to the higher standard of being in its core genome. While this work relies on the Web of Life phylogenetic tree, future analyses may consider constructing phylogenies directly from core genomes to reduce the confounding effects of horizontal gene transfer. As more genome sequences are generated for a greater diversity of species, incorporating species beyond bacteria to this analysis will enable the reconstruction of core genomes for more distant ancestral organisms. Finally, expanding the analysis from core genomes to full pangenomes would shed light on whether a species' weakly maintained accessory genes are ephemeral at an evolutionary time scale or are a persistent and co-evolving feature of the species. Ultimately, as observed biological variation appears multimodal and suggests the existence of species [49], efforts to chart the evolutionary history of life will benefit from employing both pangenomic and phylogenetic methods that offer complementary perspectives on intra- and inter-species genetic diversity.

5.6 Acknowledgements

Both authors contributed to project Conceptualization. J.C.H. developed the Methodology and Software, and conducted the Data Curation, Investigation, Formal Analysis, Validation, and Visualization of results. B.O.P. contributed to Funding Acquisition and Project Administration, and provided Supervision and Resources. J.C.H. prepared the Original Draft and both authors were involved in Review and Editing. We also thank Dr. Qiyun Zhu and Dr. Siavash Mirarab for their support in accessing and interpreting data associated with the Web of Life project.

This research was supported by grants from the Novo Nordisk Foundation (NNF-20CC0035580, awarded to B.O.P.), the National Institute of Allergy and Infectious Diseases (U01-AI124316, awarded to B.O.P.), and the National Institutes of Health (T32GM8806, awarded to J.C.H.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 5 has been submitted for publication in *Proceedings of the National Academy of Sciences*: **Jason C Hyun** and Bernhard O Palsson. "Reconstruction of the last bacterial common ancestor from 183 pangenomes reveals a versatile ancient core genome." The disseration author is the primary author.

5.7 References

[1] Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, 50(D1):D785–D794, January 2022.

[2] Qiyun Zhu, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G Sanders, Pedro Belda-Ferre, Gabriel A Al-Ghalith, Evguenia Kopylova, Daniel McDonald, Tomasz Kosciolek, John B Yin, Shi Huang, Nimaichand Salam, Jian-Yu Jiao, Zijun Wu, Zhenjiang Z Xu, Kalen Cantrell, Yimeng Yang, Erfan Sayyari, Maryam Rabiee, James T Morton, Sheila Podell, Dan Knights, Wen-Jun Li, Curtis Huttenhower, Nicola Segata, Larry Smarr, Siavash Mirarab, and Rob Knight. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat. Commun.*, 10(1):5477, December 2019.

[3] Joana C Xavier, Rebecca E Gerhards, Jessica L E Wimmer, Julia Brueckner, Fernando D K Tria, and William F Martin. The metabolic network of the last bacterial common ancestor. *Commun. Biol.*, 4(1):413, March 2021.

[4] Fouad El Baidouri, Chris Venditti, Sei Suzuki, Andrew Meade, and Stuart Humphries. Phenotypic reconstruction of the last universal common ancestor reveals a complex cell. August 2020.

[5] Madeline C Weiss, Filipa L Sousa, Natalia Mrnjavac, Sinje Neukirchen, Mayo Roettger, Shijulal Nelson-Sathi, and William F Martin. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.*, 1(9), September 2016.

[6] Eugene V Koonin. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.*, 1(2):127–136, November 2003.

[7] W Ford Doolittle and Tyler D P Brunet. What is the tree of life? *PLoS Genet.*, 12(4):e1005912, April 2016.

[8] Joel Velasco. Universal common ancestry, LUCA, and the tree of life: three distinct hypotheses about the evolution of life. *Biol. Philos.*, 33(5-6), December 2018.

[9] Nicolas Glansdorff, Ying Xu, and Bernard Labedan. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol. Direct*, 3(1):29, July 2008.

[10] C Woese. The universal ancestor. *Proc. Natl. Acad. Sci. U. S. A.*, 95(12):6854–6859, June 1998.

[11] Michael A Brockhurst, Ellie Harrison, James P J Hall, Thomas Richards, Alan McNally, and Craig MacLean. The ecology and evolution of pangenomes. *Curr. Biol.*, 29(20):R1094–R1103, October 2019.

[12] Jason C Hyun, Jonathan M Monk, and Bernhard O Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022.

[13] Miklós Csűrös. Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In *Comparative Genomics*, Lecture notes in computer science, pages 72–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[14] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial

taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004, November 2018.

[15] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, March 2010.

[16] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[17] Andrew J Page, Carla A Cummins, Martin Hunt, Vanessa K Wong, Sandra Reuter, Matthew T G Holden, Maria Fookes, Daniel Falush, Jacqueline A Keane, and Julian Parkhill. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22):3691–3693, November 2015.

[18] Bruno Contreras-Moreira and Pablo Vinuesa. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.*, 79(24):7696–7701, December 2013.

[19] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, 38(12):5825–5829, December 2021.

[20] Zoe L Watson, Fred R Ward, Raphaël Méheust, Omer Ad, Alanna Schepartz, Jillian F Banfield, and Jamie Hd Cate. Structure of the bacterial ribosome at 2 å resolution. *Elife*, 9, September 2020.

[21] Michael Y Galperin, Yuri I Wolf, Sofya K Garushyants, Roberto Vera Alvarez, and Eugene V Koonin. Nonessential ribosomal proteins in bacteria and archaea identified using clusters of orthologous genes. *J. Bacteriol.*, 203(11), May 2021.

[22] Miklós Csurös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, August 2010.

[23] Boris G Mirkin, Trevor I Fenner, Michael Y Galperin, and Eugene V Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.*, 3:2, January 2003.

[24] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, 49(D1):D274–D281, January 2021.

[25] Y I Wolf, L Aravind, N V Grishin, and E V Koonin. Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.*, 9(8):689–710, August 1999.

[26] Tõnu Margus, Maido Remm, and Tanel Tenson. A computational study of elongation factor G (EFG) duplicated genes: diverged nature underlying the innovation on the same structural template. *PLoS One*, 6(8):e22789, August 2011.

[27] Marian Breuer, Tyler M Earnest, Chuck Merryman, Kim S Wise, Lijie Sun, Michaela R Lynott, Clyde A Hutchison, Hamilton O Smith, John D Lapek, David J Gonzalez, Valérie de Crécy-Lagard, Drago Haas, Andrew D Hanson, Piyush Labhsetwar, John I Glass, and Zaida Luthey-Schulten. Essential metabolism for a minimal cell. *Elife*, 8, January 2019.

[28] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000.

[29] W D Roof, K F Foltermann, and J R Wild. The organization and regulation of the pyrBI operon in e. coli includes a rho-independent attenuator sequence. *Mol. Gen. Genet.*, 187(3):391–400, 1982.

[30] Matteo Brilli and Renato Fani. Molecular evolution of hisb genes. *J. Mol. Evol.*, 58(2):225–237, February 2004.

[31] Antoine Danchin, Gang Fang, and Stanislas Noria. The extant core bacterial proteome is an archive of the origin of life. *Proteomics*, 7(6):875–889, March 2007.

[32] Robert L Charlebois and W Ford Doolittle. Computing prokaryotic gene ubiquity: rescuing the core from extinction. *Genome Res.*, 14(12):2469–2477, December 2004.

[33] S Hansmann and W Martin. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int. J. Syst. Evol. Microbiol.*, 50 Pt 4(4):1655–1663, July 2000.

[34] Natalya Yutin, Pere Puigbò, Eugene V Koonin, and Yuri I Wolf. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, 7(5):e36972, May 2012.

[35] Nathan J Baird, Jinwei Zhang, Tomoko Hamma, and Adrian R Ferré-D'Amaré. YbxF and YlxQ are bacterial homologs of L7Ae and bind k-turns but not k-loops. *RNA*, 18(4):759–770, April 2012.

[36] C S Hamann, K R Sowers, R S Lipman, and Y M Hou. An archaeal aminoacyl-tRNA synthetase missing from genomic analysis. *J. Bacteriol.*, 181(18):5880–5884, September 1999.

[37] Miguel Angel Rubio Gomez and Michael Ibba. Aminoacyl-tRNA synthetases. *RNA*, 26(8):910–936, August 2020.

[38] Gemma C Atkinson and Sandra L Baldauf. Evolution of elongation factor G and the origins of mitochondrial and chloroplast forms. *Mol. Biol. Evol.*, 28(3):1281–1292, March 2011.

[39] A Maxwell Burroughs and L Aravind. The origin and evolution of release factors: Implications for translation termination, ribosome rescue, and quality control pathways. *Int. J. Mol. Sci.*, 20(8), April 2019.

[40] Bradley E Poulsen, Rui Yang, Anne E Clatworthy, Tiantian White, Sarah J Osmulski, Li Li, Cristina Penaranda, Eric S Lander, Noam Shoresh, and Deborah T Hung. Defining the core essential genome of pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.*, 116(20):10072–10080, May 2019.

[41] Yoann Le Breton, Ashton T Belew, Kayla M Valdes, Emrul Islam, Patrick Curry, Hervé Tettelin, Mark E Shirtliff, Najib M El-Sayed, and Kevin S McIver. Essential genes in the core genome of the human pathogen streptococcus pyogenes. *Sci. Rep.*, 5(1):9838, May 2015.

[42] Enrique Martínez-Carranza, Hugo Barajas, Luis-David Alcaraz, Luis Servín-González, Gabriel-Yaxal Ponce-Soto, and Gloria Soberón-Chávez. Variability of bacterial essential genes among closely related bacteria: The case of escherichia coli. *Front. Microbiol.*, 9, May 2018.

[43] Carlos G Acevedo-Rocha, Gang Fang, Markus Schmidt, David W Ussery, and Antoine Danchin. From essential to persistent genes: a functional approach to constructing synthetic life. *Trends Genet.*, 29(5):273–279, May 2013.

[44] Georgina Hernández-Montes, J Javier Díaz-Mejía, Ernesto Pérez-Rueda, and Lorenzo Segovia. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biol.*, 9(6):R95, June 2008.

[45] Jing Tian, Ruslana Bryk, Shuangping Shi, Hediye Erdjument-Bromage, Paul Tempst, and Carl Nathan. Mycobacterium tuberculosis appears to lack alpha-ketoglutarate dehydrogenase and encodes pyruvate dehydrogenase in widely separated genes. *Mol. Microbiol.*, 57(3):859–868, August 2005.

[46] Kenneth R Olson and Karl D Straub. The role of hydrogen sulfide in evolution and the evolution of hydrogen sulfide in metabolism and signaling. *Physiology (Bethesda)*, 31(1):60–72, January 2016.

[47] Bin Du, Daniel C Zielinski, Jonathan M Monk, and Bernhard O Palsson. Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice. *Proc. Natl. Acad. Sci. U. S. A.*, 115(44):11339–11344, October 2018.

[48] Stephen K Dolan and Martin Welch. The glyoxylate shunt, 60 years on. *Annu. Rev. Microbiol.*, 72(1):309–330, September 2018.

[49] Nicolas Galtier. Delineating species in the speciation continuum: A proposal. *Evol. Appl.*, 12(4):657–663, April 2019.

Chapter 6

Conclusions

The falling cost of genome sequencing and subsequent explosion of genomic data has shed light on the full genetic repertoires of microbial species, which far surpass the complexity of any individual organism. Developments in the construction and analysis of pangenomes have found structure within this tremendous genetic diversity, allowing more nuanced characterizations of how closely different species maintain specific genes and accelerating the discovery of novel genotype-phenotype relationships underlying complex biological phenomena. The work of this dissertation aims to systematize the analysis of pangenomes to identify constants in the genetic diversity present across different bacteria, demonstrate robust and generalizable workflows for elucidating antimicrobial resistance (AMR) at pangenome scale, and integrate pangenomics with phylogenetics to characterize the core genome of the last bacterial common ancestor and examine its implications regarding the minimum requirements for life.

Chapter 2 presents the pangenome workflow that forms the basis of this dissertation and documents the breakdown of microbial pangenomes into units of genetic diversity of increasing resolution. From pangenome shape, to genes, to sequence variants, and finally to individual mutations, this analysis finds each resolution to reveal a unique aspect of microbial diversity corresponding to genetic variation at different timescales. A comparison of twelve pathogens found that 1) a species' pangenome openness appears

to follow its phylogenetic placement, 2) the distribution of gene frequencies adopts a consistent mathematical form in line with the existence of distinct core, accessory, and unique genomes, 3) core genomes allocate similar proportions of genes to specific functional categories, 4) sequence variation in the core genome impacts a broad range of genes functions, and 5) core genome mutations tend to clusters in specific protein domains for certain genes, especially those in aminoacyl-tRNA synthetases. These results suggest that many aspects of pangenome variation are stable across species, and the continued discovery of constants in microbial genetic diversity may guide future pangenome analyses as continued genome sequencing will enable pangenome construction for more species at even greater scale.

Chapter 3 combines this pangenome framework with experimental AMR measurements to identify novel AMR genes. In light of the challenges that the clonal nature of bacteria presents in the use of traditional methods for genome-wide association studies (GWAS), this work instead employs machine learning (ML) to learn gene-AMR associations from AMR metadata and the genetic features enumerated by the pangenome workflow. Upon dissecting learned feature weights, support vector machine (SVM) ensembles were shown to outperform traditional GWAS statistical tests in recovering known AMR genes from the *S. aureus* pangenome, with feature subsampling found to be especially impactful in improving recovery. Extending this technique to two additional pangenomes for *P. aeruginosa* and *E. coli* identified a total 45 known AMR genetic features as well as 25 novel AMR gene candidates, and a case study of fluoroquinolone resistance confirmed both the reliability of the machine learning approach and the comprehensiveness of pangenome analysis when compared to existing literature.

Building on these promising results, Chapter 4 significantly expands the application of pangenomics and ML in the problem of AMR elucidation, covering twelve species and increasing both the number of genomes and number of antimicrobials analyzed by an order of magnitude. The overall workflow was refined through the introduction of noncoding

genetic features as well as the systematic annotation of known AMR genes in each pangenome, through which the evaluation of known AMR gene recovery could be directly integrated with the ML training process. These expansions revealed that 1) many AMR genes can be found in multiple species but most are confined within closely related species, 2) rare instances of AMR genes crossing between distant species, such as the introduction of TEM beta-lactamases to *S. aureus*, can meaningfully increase resistance, and 3) the improved SVM ensemble approach consistently outperforms traditional statistical tests at recovering known AMR genes, recovering 265 such genes across 127 species-drug cases, more than twice as many as Fisher's exact test. Analysis of the best models identified 142 novel AMR gene candidates, of which two were experimentally confirmed as involved in cases of conditional resistance: deletion of *cycA* confers quinolone resistance in minimal media with D-serine, and the V111D mutation in *frdD* confers beta-lactam resistance in the presence of *ampC* by altering its overlapping promoter. Overall, these two chapters provide broad confirmation of the power of pangenomics and ML to drive biological discovery for clinical applications, and demonstrates the power of scale in identifying subtle contributors to antimicrobial resistance.

Finally, Chapter 5 applies the pangenomic perspective to the problem of characterizing the last bacterial common ancestor (LBCA). Much as previous studies reconstructed individual ancestral organisms through phylogenetic analysis of diverse collections of genomes, this work aimed to reconstruct the species-wide core genome of the LBCA from a diverse collection of pangenomes. Core genomes for 183 modern species were identified using a novel statistical model to account for errors in pangenome construction and integrated with an existing bacterial phylogeny to reconstruct the LBCA core genome under asymmetric Wagner parsimony, revealing that 1) modern core genomes allocate similar fractions of genes to specific gene categories but are diverse at the level of individual genes, 2) the LBCA core genome is distinct from that of any modern species but is versatile with many fundamental biological systems intact, especially those related to translational

machinery, central carbon metabolism, and *de novo* biosynthetic pathways, and 3) yet despite this versatility, comparison against the synthetic minimal organism JCVI-Syn3A suggests that the LBCA core genome alone is likely insufficient for life and requires additional accessory genes for survival. These results suggest that many of biological systems previously believed to be ancient were not just present but closely maintained by ancestral bacteria, and highlight the distinction between core and essential genes with implications in the design of simplified workhorse organisms for bioengineering applications.

Following the rapid growth in genome sequencing data, pangenome analysis has emerged as a powerful tool for tackling questions both fundamental and practical regarding the diversity of microbial life. This dissertation presents a robust workflow for translating amorphous collections of genomes into structured pangenomes, and leverages this structure to identify constants in intraspecies genetic diversity across the bacterial kingdom, identify global patterns in AMR and novel AMR-conferring genes, and characterize the core genome of the LBCA to explore the minimum genetic requirements for life. These results demonstrate that with the appropriate analytical tools, genetic analyses of larger and larger datasets can deliver unique insights that are invisible at smaller scales. Given the countless unanswered questions that remain regarding microbial diversity and its clinical and bioengineering implications, continued development of pangenomics will be necessary to ensure that the pace of discovery tracks the pace of data collection.

Appendix A

Comparative pangenomics: Revealing conserved structures of genetic and functional diversity - Supplementary Information

A.1 Methods

A.1.1 Genome selection, pangenome construction, MLST classification, and feature identification

An initial set of genomes was taken from the PATRIC database RELEASE_NOTES (ftp.patricbrc.org/RELEASE_NOTES/, 2020-02-06), starting with ESKAPEE pathogens and WHO global priority pathogens and filtered down to 12 species with at least 100 genomes by taxon ID (Table A.1). For each species, genomes were filtered to those meeting the following quality criteria: 1) genome status is "WGS" or "Complete", 2) number of contigs is within 2.5 times the median number of contigs across all assemblies for that species, 3) number of annotated CDSs is within 3 standard deviations of the mean, and 4) total genome length is within 3 standard deviations of the mean. PATRIC Genome IDs for the selected genomes are available in Dataset A.1. Each genome was classified *in silico* by multilocus sequence type (MLST) using the mlst tool v2.18.0 (https://github.com/tseemann/mlst) [1] (Dataset A.2). A phylogenetic tree was con-

structed based on reference genomes of each species available on PATRIC, using PATRIC's Phylogenetic Tree service with the Codon Trees method and a maximum of 100 genes (Fig. A.1a) [2]. In the cases of *C. coli* and *A. baumannii*, no reference genome was available on PATRIC and representative genomes were used.

For a given species, all CDSs across all genomes (as annotated by PATRIC) were reduced to a non-redundant list and clustered by protein sequence using CD-HIT v4.6 (word size "-n" 5, minimum identity "-T" 80%, minimum alignment length "-aL" 80%, all other settings default) [3]. Each cluster was denoted a "gene" and each cluster member denoted a coding variant. For each gene, $5'$ intergenic variants were identified by locating occurrences of all coding variants of the gene across all genome assemblies and extracting the DNA sequence from the start codon to 300nt upstream. $3'$ intergenic variants were similarly identified downstream of stop codons. Intergenic variants truncated by contig boundaries were ignored.

A.1.2 Pangenome openness estimation and extrapolation with Heaps' law

To estimate pangenome openness for a given species, 100 random genome orderings were generated using two approaches: 1) *genome-based*: all available genomes were randomly shuffled, and 2) *MLST-based*: all identified MLST types were randomly shuffled and one genome was randomly sampled per MLST in the resulting order (genomes that could not be typed were grouped as a single separate subtype). For each genome ordering, the total number of unique genes encountered (pangenome size) as genomes are introduced sequentially was computed and fit to Heaps' Law using nonlinear least squares regression via SciPy's curve_fit() [4]. The mean and standard deviation of fitted Heaps' Law parameters across the 100 orderings for each method were computed.

To evaluate each method's ability to extrapolate pangenome size, Heaps' Law was fit to the first half of genomes in each genome ordering, and the mean absolute error

(MAE) was computed for the fit against both the first half (fit region) and remaining second half (extrapolation region) of genomes. The median MAE across the 100 orderings was computed for both methods for each species, as well as the relative median MAE (median MAE from the MLST-based approach divided by the median MAE from the genome-based approach).

A.1.3 Frequency-based division of pangenomes into core, accessory and unique genes

For a given species with N genomes, two distributions were computed: $P(x)$, the number of genes with frequency x, and $F(x)$, the cumulative genes with frequency less than or equal to x. To account for the observation that $P(x)$ and $P(N+1-x)$ are approximately power laws for small values, the overall frequency distribution was modeled using the following function with parameters (c_1, c_2, a_1, a_2):

$$P(x) = c_1 x^{-\alpha_1} + c_2 (N+1-x)^{-\alpha_2} \qquad x = 1, 2, ..., N$$

Since observed $P(x)$ values varied across multiple orders of magnitude, parameters of this function were fitted using the cumulative distribution, based on the integral of the $P(x)$ model and involving an extra constant parameter k:

$$F(x) = k + \frac{c_1}{1-\alpha_1} x^{1-\alpha_1} - \frac{c_2}{1-\alpha_2} (N+1-x)^{1-\alpha_2}$$

This five parameter function was fit to observed cumulative frequency distributions, using nonlinear least squares regression via SciPy's curve_fit() [4], linearly scaling the domain and range to be within $0-1$ and with initial guess $(c_1, c_2, \alpha_1, \alpha_2, k) = (1, 1, 2, 2, 1)$. The inflection point of $F(x)$, or x^*, was computed by minimizing $P(x)$ with the corresponding computed parameters in SciPy using minimize_scalar() [4]. Frequency thresholds for core, accessory, and unique genes were defined relative to this inflection point, where unique

genes were defined as those present in less than $0.1x^*$ strains, core genes as those present in more than $0.9N + 0.1x^*$ strains, and accessory genes as everything in between. Fitted parameters and derived frequency thresholds are available in Dataset A.3.

A.1.4 Orthogroup identification and enrichment testing between gene function and frequency

For each gene in each pangenome, the most commonly observed coding variant was annotated using eggNOG-emapper version 0.12.7 [5], as the representative for that gene. This annotation yielded for each gene its best orthogroup or "bestOG", COG functional category, and associated GO terms. Genes that eggNOG-emapper failed to annotate were assigned the COG category "S: Function unknown". For each species, Fisher's exact tests were applied to determine enrichment between each gene frequency group (core, accessory, unique) and COG functional category. For example, to test enrichment for COG J in the core genome, Fisher's exact test was applied between core vs. non-core genes and COG J vs. non-COG J genes. A total of 12 species $*$ 3 frequency groups $*$ 20 COGs $= 720$ tests were conducted, and significance was determined based on FWER < 0.05 under Bonferroni correction, or $p < 7 * 10^{-5}$. Log_2 odds ratios (LOR) were also computed between each frequency group and COG.

Analogous enrichment tests and LOR calculations were conducted for 414 GO terms that were observed at least 10 times in each species. A total of 12 species $*$ 3 frequency groups $*$ 414 GO terms $= 14,904$ total tests were conducted, and significance was determined based on FWER < 0.05 under Bonferroni correction, or $p < 3 * 10^{-6}$. The top 10 GO terms by mean LOR across all 12 species were reported. All LORs and p-values for both COG and GO terms are available in Dataset A.4.

To identify genes conserved across all species' core genomes, genes from different species' pangenomes were grouped by eggNOG-emapper's bestOG assignment. Gene essentiality was assigned based on comparing eggNOG-emapper predicted gene names to

essentiality predictions made in Goodall et.al. [6] and are available in Dataset A.5.

A.1.5 Analysis of intraspecies sequence-level diversity in core genomes

For each species and core gene, the frequency of each observed coding variant was counted and the Shannon entropy of this variant count distribution plus a dummy variant with frequency 1 (in order to distinguish genes with similar variant relative frequencies but different raw counts) was computed, referred to as the "coding allelic entropy" of the gene for that species. Analogous $5'$ intergenic and $3'$ intergenic allelic entropies were also computed per gene based on distributions of their respective variant types. Core genome-wide Spearman correlations were computed between these three allelic entropies for each pair of variant types, for each species.

To control for the effect of gene length on the number of unique coding variants and thus on coding allelic entropy, quantile regression was used to determine the 5 and 95% coding allelic entropy percentiles as a quadratic function of gene length [7], using Python package statsmodels [8]; the quadratic functional form was chosen as it was the simplest form that closely tracked the rolling window 5 and 95% percentiles (Fig. A.8). Functional enrichment among the most conserved and diverse core genes (determined by the 5 and 95% quantile regression percentiles) was computed similarly as for the frequency group enrichment tests, computing LORs and applying Fisher's exact tests for each COG functional category. A total of 2 groups (top/bottom 5%) $*$ 20 COGs $*$ 12 species $= 480$ tests were conducted, and significance was determined based on FWER < 0.5 under Bonferroni correction, or $p < 1 * 10^{-4}$. Similar enrichment tests were conducted for the top/bottom 5% of core genes ranked by either intergenic allelic entropy measure, using regular 5%/95% quantiles not based on quantile regression since intergenic features were fixed-length. All LORs and p-values are available in Dataset A.4.

A.1.6 Analysis of sequence-level diversity in MLST genes

DNA sequences for genes involved in PubMLST typing schemes were downloaded through the mlst tool v2.18.0 (https://github.com/tseemann/mlst). Each sequence was translated to an amino acid sequence in the frame with the fewest number of intermediate stop codons. Within each species, pangenome coding variants were mapped to translated PubMLST variants if they contained the exact sequence of the translated PubMLST variant, to yield variant-variant mappings. Pangenome genes were then mapped to PubMLST genes based on which pangenome gene had the maximum number of variant-variant mappings to a given PubMLST gene. The coding allelic entropies of the pangenome genes mapped to PubMLST genes were reported, as percentiles relative to the coding allelic entropies of all core genes for the species.

A.1.7 Analysis of sequence variation positional distribution with respect to domains

The 168 genes previously identified to be in all 12 core genomes were filtered for those with rich domain annotations. Starting with *E. coli* amino acid sequences, for each gene: 1) a multiple sequence alignment (MSA) was computed for all observed coding variants using MAFFT [9], 2) the consensus sequence was annotated for domains with InterProScan [10], 3) domains with the same InterPro accession ID were merged, and 4) domains longer than 80% of the full protein length were filtered out. This analysis yielded 76 genes with at least three domain annotations, and the amino acid sequences related to these genes for all 12 species were similarly analyzed for a total of 912 species-gene pairs annotated. Gene name annotations were assigned based on earlier eggNOG-emapper annotations of the most common sequence variant.

To quantify domain sequence variation, the entropy at each position of each species-gene MSA was computed, weighted by the relative abundance of each sequence variant. For each domain, the mean entropy across all MSA positions spanned by the domain

was computed, and the entropy's percentile was computed against the mean entropies of all subsequences of the same length within the MSA, yielding entropy percentiles for each species-gene-domain combination. To determine domains consistently variable or conserved across multiple species, the mean entropy percentile for each gene-domain pair was computed across the 12 species. P-values of mean entropy percentiles were determined against an empirically constructed distribution for the mean of 12 independent, identically distributed uniform distributions using 1,000,000 random samples (Bates distribution, based on a null hypothesis of percentiles being uniformly distributed). Significance was determined based on Benjamini-Hochberg correction (FDR < 0.05, 443 tests). All domain entropy percentiles, p-values, and significance calls are available in Dataset A.6.

Domains related to aminoacyl-tRNA synthetases (AARSs) were identified based on gene name annotations. Functional categories (editing, anticodon binding, tRNA binding, non-editing catalytic) were assigned based on InterPro text annotations of each domain. Between domains overlapping by more than 95%, only the domain with a functional annotation (rather than structural) and/or more descriptive InterPro annotation was shown as representative. A summary of domain functional assignments, evidence, and overlap filtering is available in Dataset A.6.

A.2 Supplementary Figures

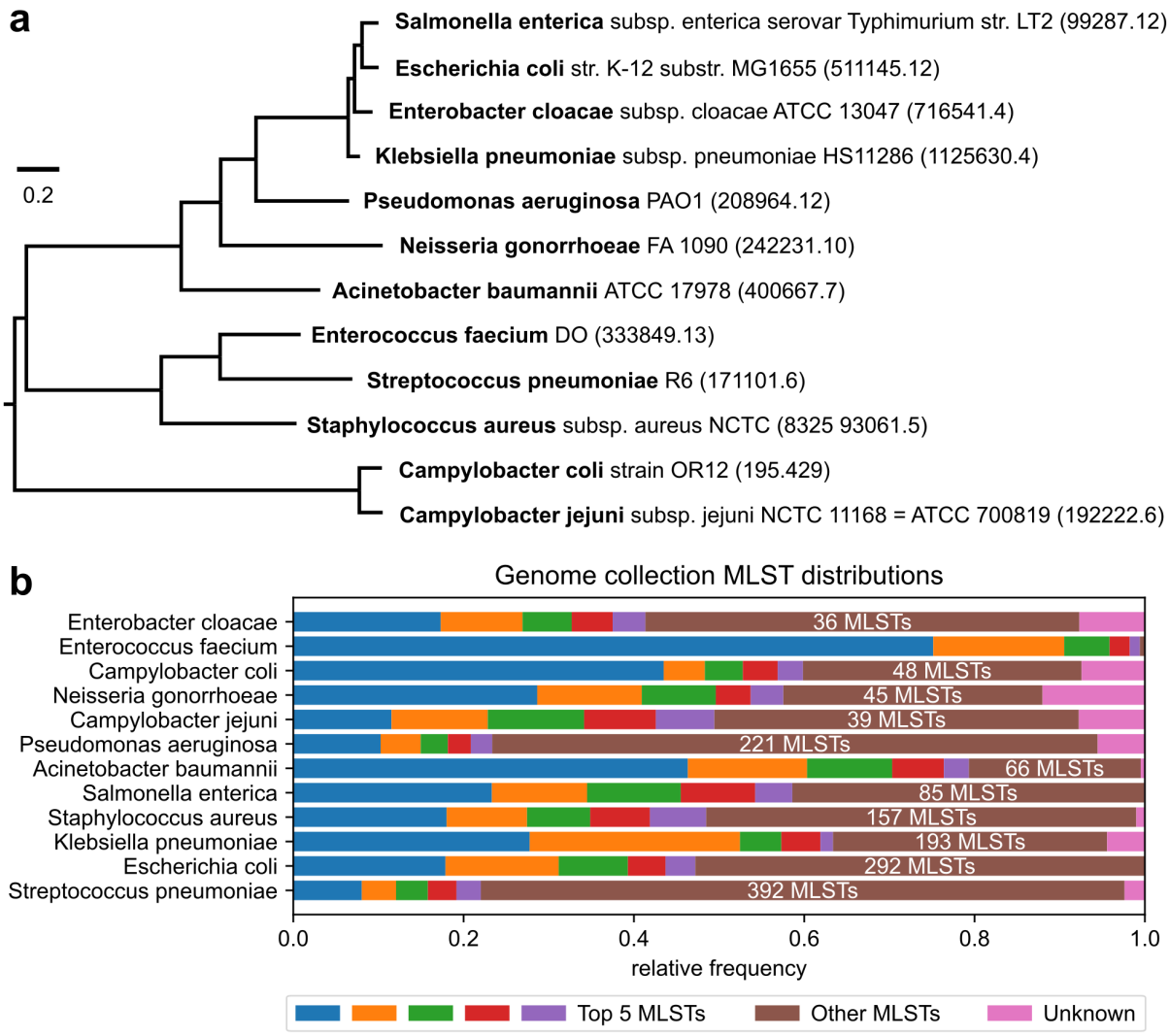


Figure A.1. Phylogenetic tree of 12 microbial species and MLST distributions. (a) Phylogenetic tree constructed for representative genomes of each species using PATRIC's Codon Tree service. Genomes are labeled by their name and PATRIC Genome ID. (b) Distribution of MLST subtypes for each species' genome collection. The relative abundance of the top 5 MLST subtypes, all other subtypes, and untyped genomes are shown per species.

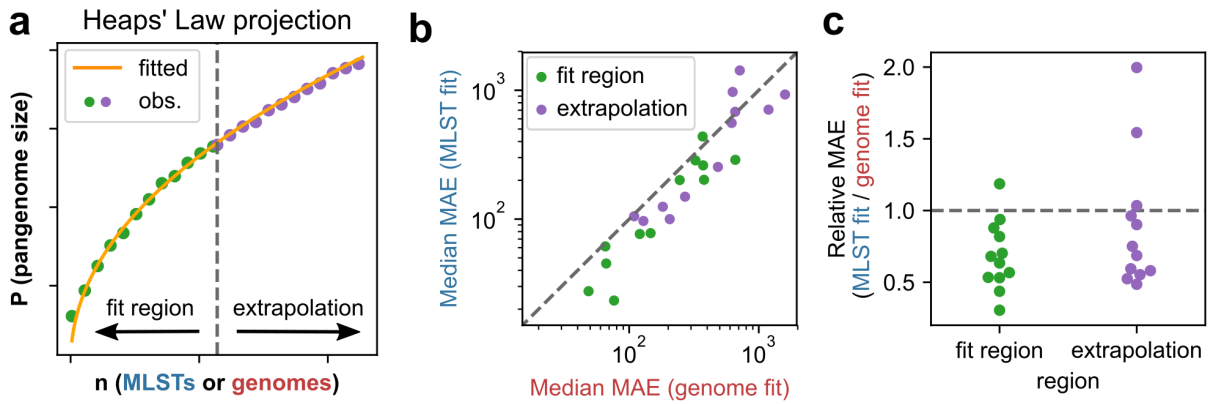


Figure A.2. Evaluation of accuracy of Heaps' Law at predicting pangenome size, with or without controlling from MLST. (a) Example fit of Heaps' Law to first half of genomes (unbalanced) or MLSTs (MLST balanced) and extrapolation to second half to evaluate pangenome size projection. (b-c) Median mean absolute error (MAE) across Heaps' Law fits for 100 random genome orderings, with or without MLST balancing, for each species in the fitting and extrapolation regions. Dotted lines indicate equal performance between the two methods.

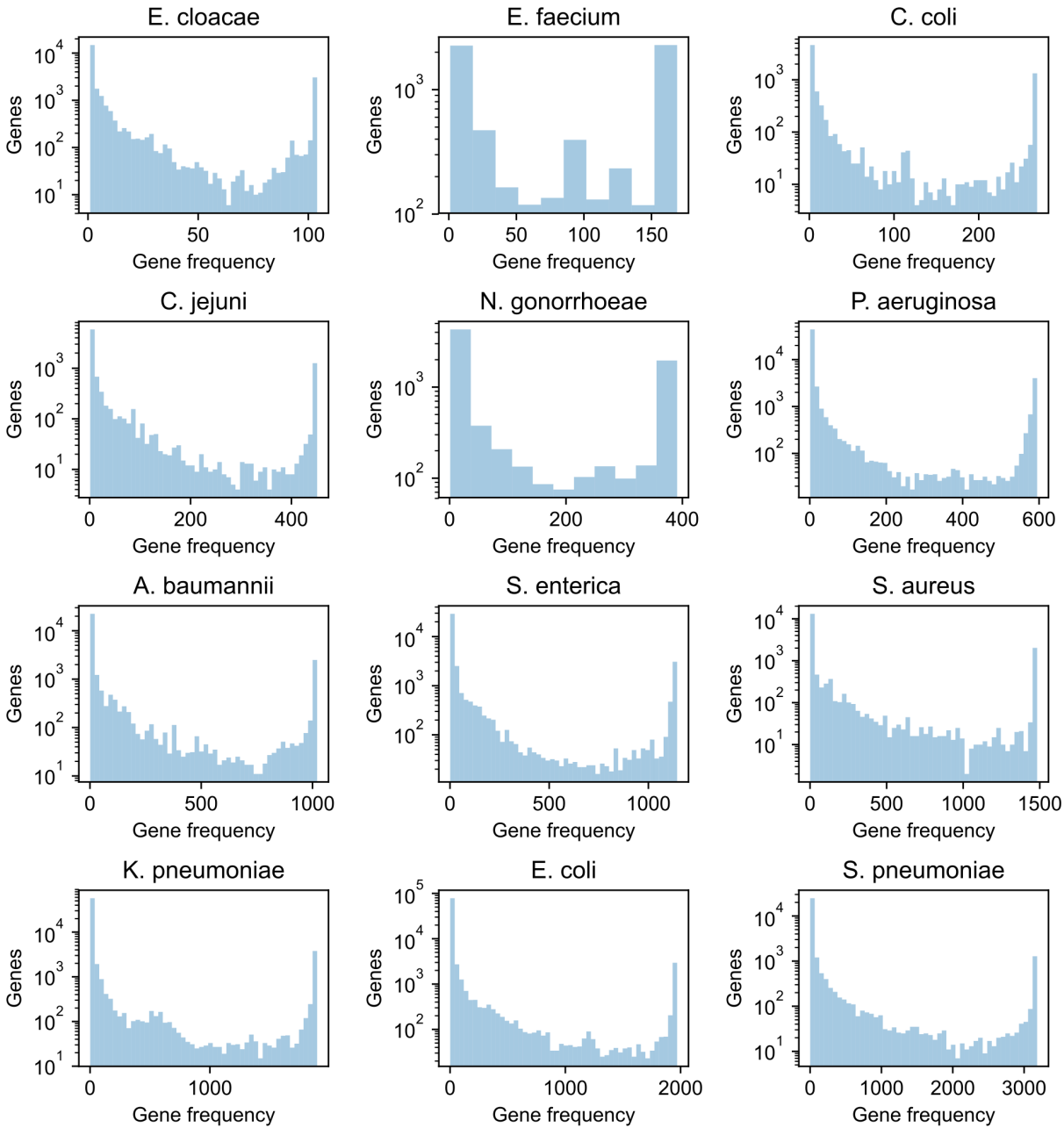


Figure A.3. Gene frequency distributions for 12 microbial species.

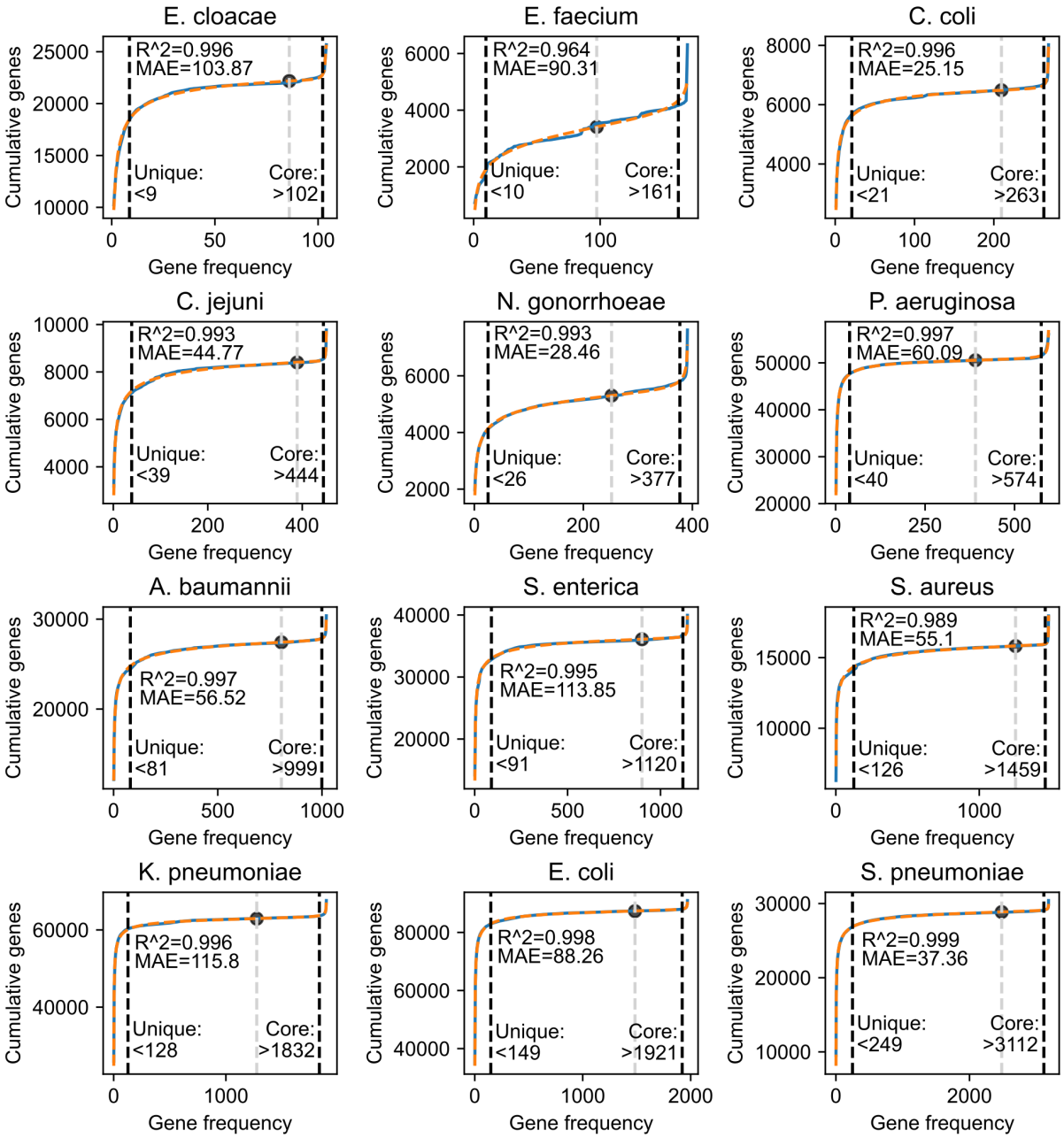


Figure A.4. Fitted cumulative gene frequency distributions and corresponding core and unique gene frequency thresholds for 12 species. Observed distributions (solid blue), fitted functions (dashed orange), and the R^2 and mean absolute errors (MAE) of the fits are shown. Fitted inflection points (black dot, dashed gray) and frequency thresholds corresponding to core and unique genes (dashed black) are also shown.

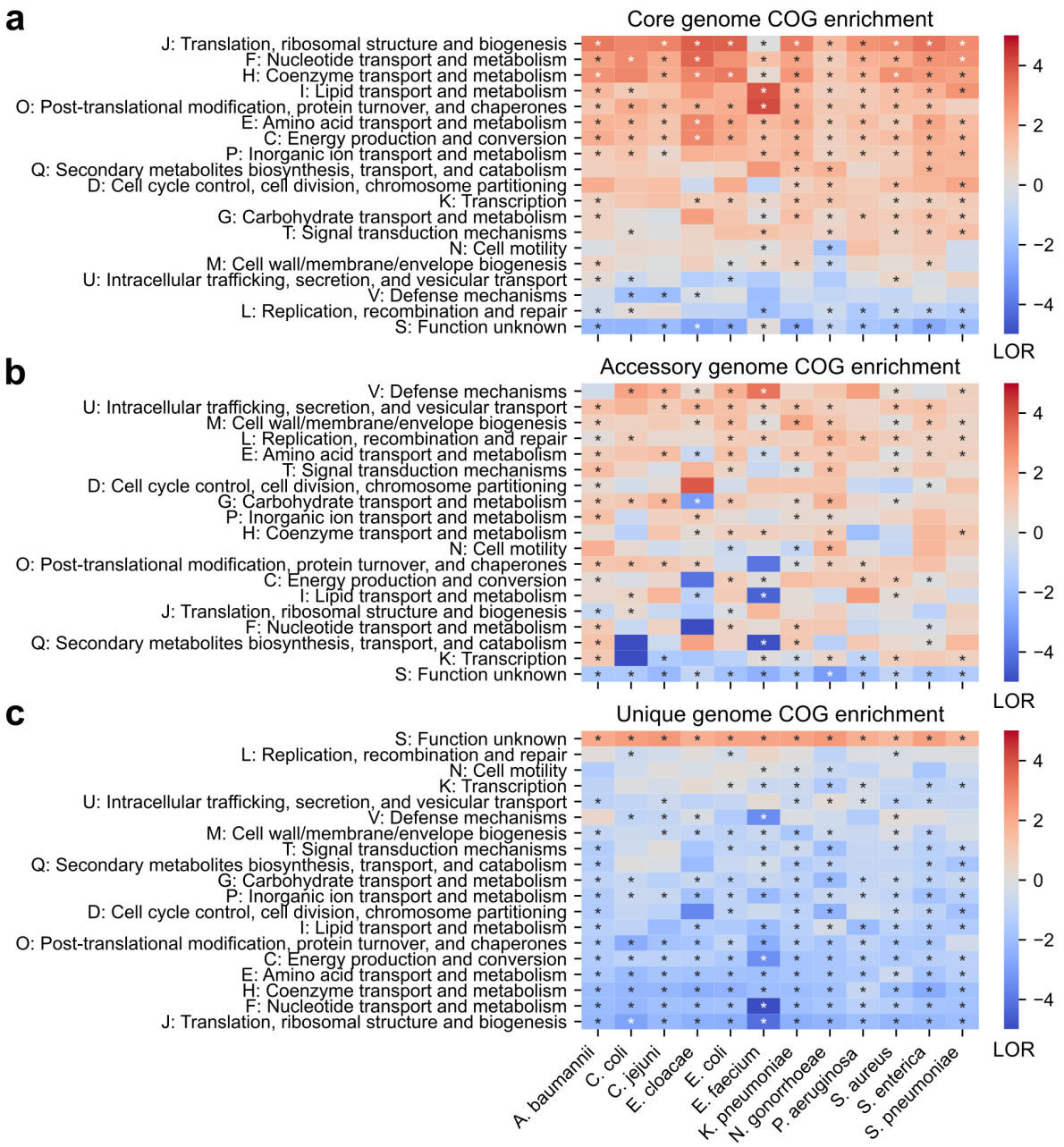


Figure A.5. COG functional group enrichment in the core, accessory, and unique genomes of 12 species. Heatmaps are colored by the $\log_2$ odds ratio (LOR) between each COG and the (a) core, (b) accessory, (c) unique genome of each species. COGs are sorted by mean LOR across all species. LOR color scales are symmetric and identical for all plots; four values are outside of the color range: F x *E. cloacae* (LOR = −7.5), Q x *C. coli* (LOR = −6.0), and K x *C. coli* (LOR = −6.9) for accessory genomes; F x *E. faecium* (LOR = −5.3) for unique genomes. Starred cells correspond to statistically significant enrichments under Fisher's Exact test with FWER < 0.05 under Bonferroni correction ($p < 7 * 10^{-5}$, 720 tests).

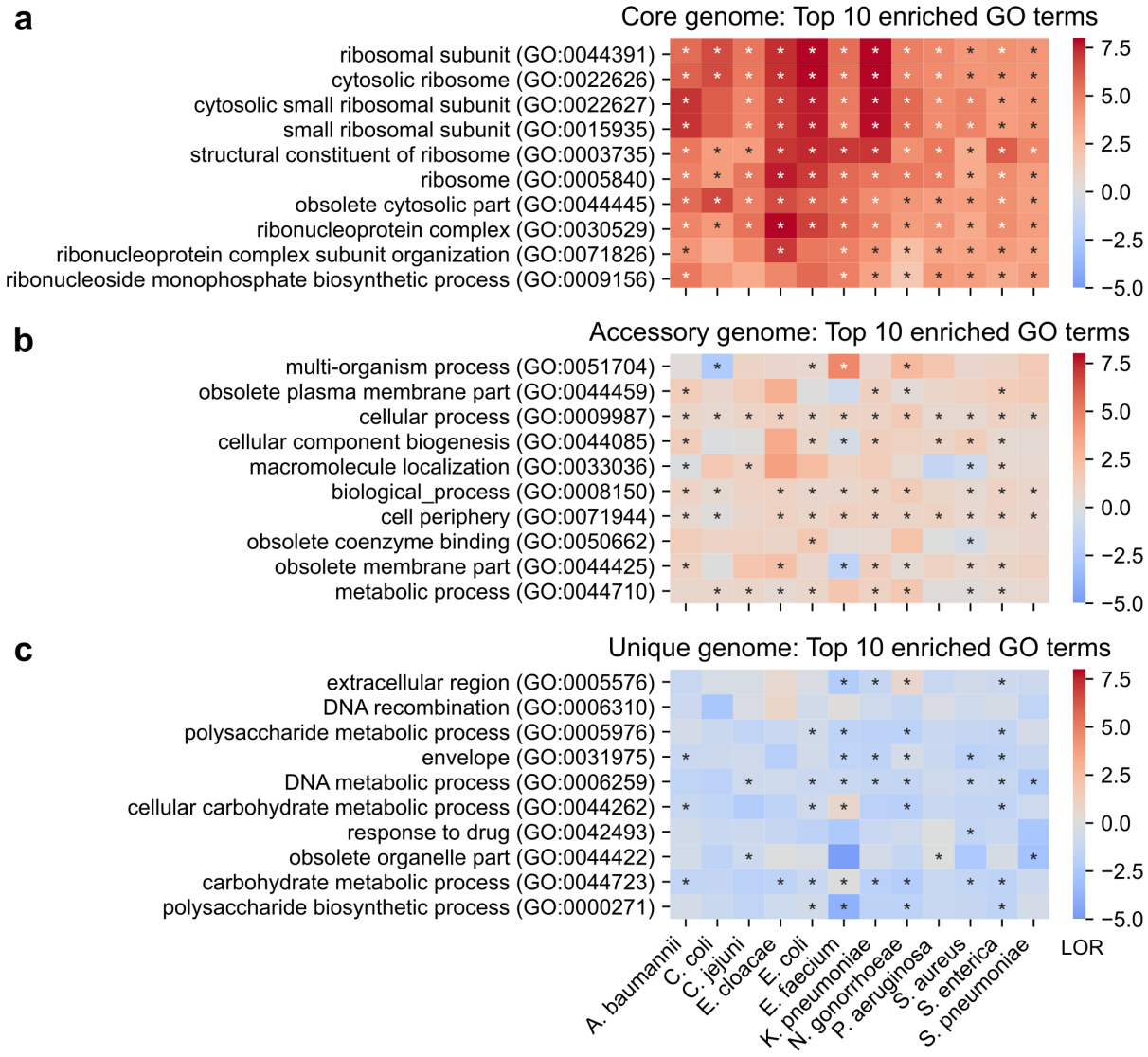


Figure A.6. Top 10 GO terms by enrichment in the core, accessory, and unique genomes of 12 species. Heatmaps are colored by the $\log_2$ odds ratio (LOR) between each GO term and the (a) core, (b) accessory, or (c) unique genome of each species. GO terms are sorted by mean LOR across all species. LOR color scales are identical for all plots. Starred cells correspond to statistically significant enrichments under Fisher's Exact test with FWER < 0.05 under Bonferroni correction ($p < 3 * 10^{-6}$, 14,904 tests).

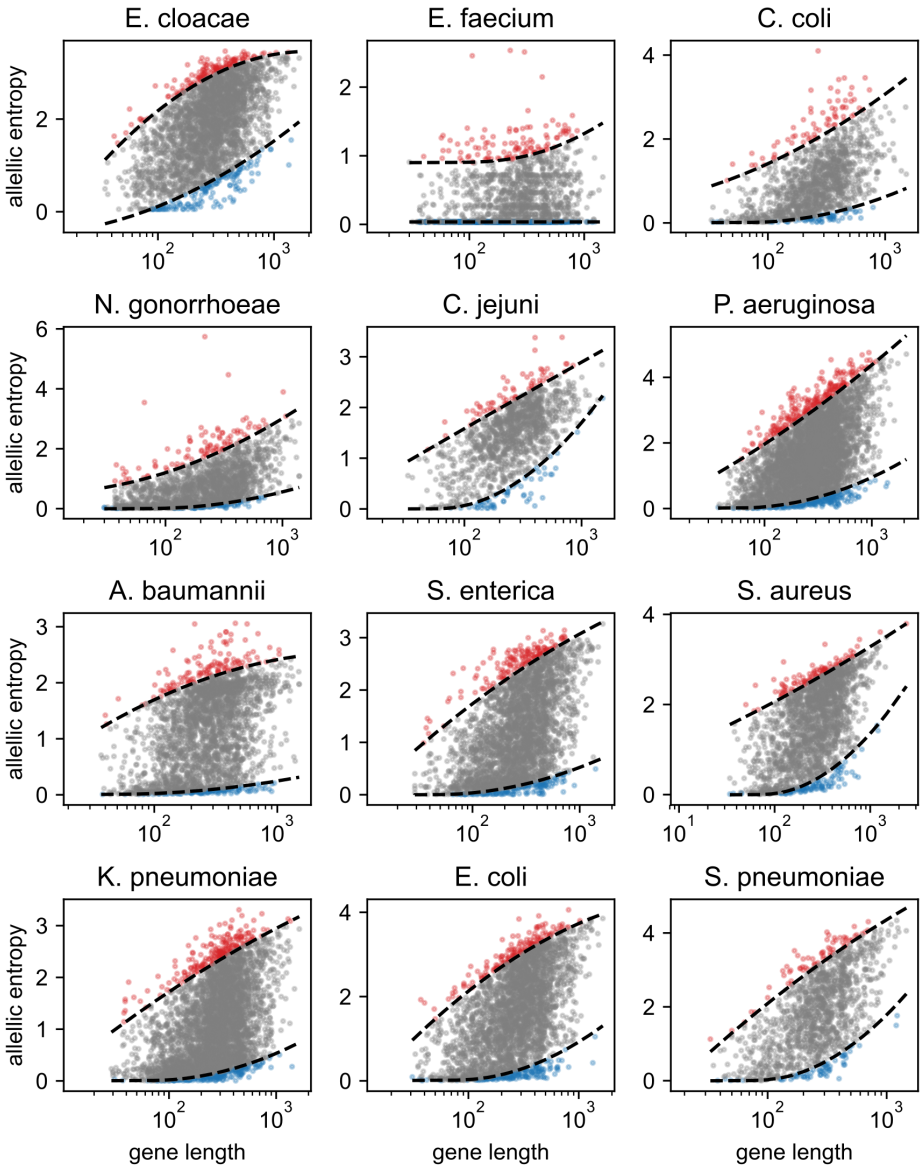


Figure A.7. Quantile regression between coding allelic entropy and gene length among core genes for 12 species. Dotted lines show quantile regressions for the 5% and 95% coding allelic entropy percentiles as a quadratic function of gene length. Red and blue dots are the most diverse and most conserved core genes, respectively, as defined by these regressions.

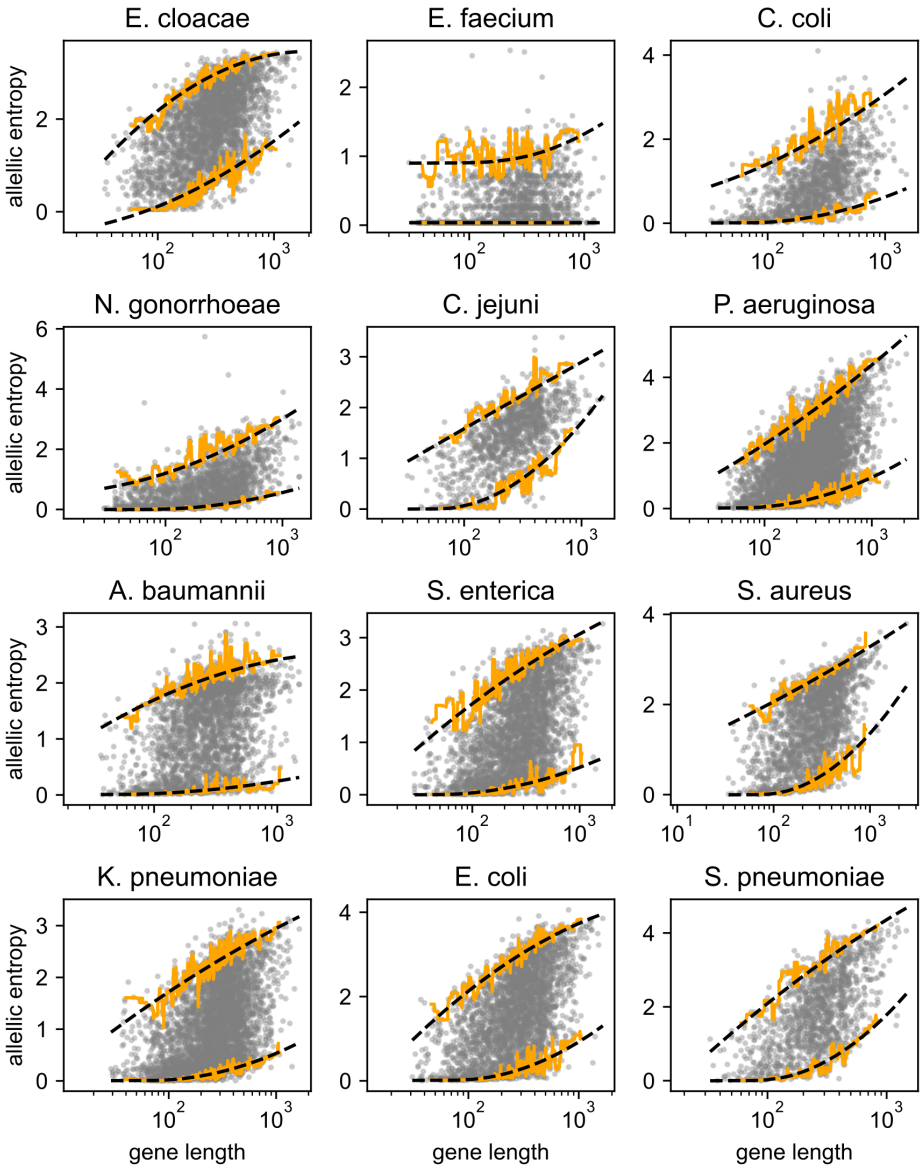


Figure A.8. Rolling window percentiles versus quantile regression between coding allelic entropy and gene length among core genes, by species. Dotted lines show quantile regressions for the 5% and 95% coding allelic entropy percentiles as a quadratic function of gene length. Orange lines show rolling 5% and 95% percentiles using windows of 50 genes.

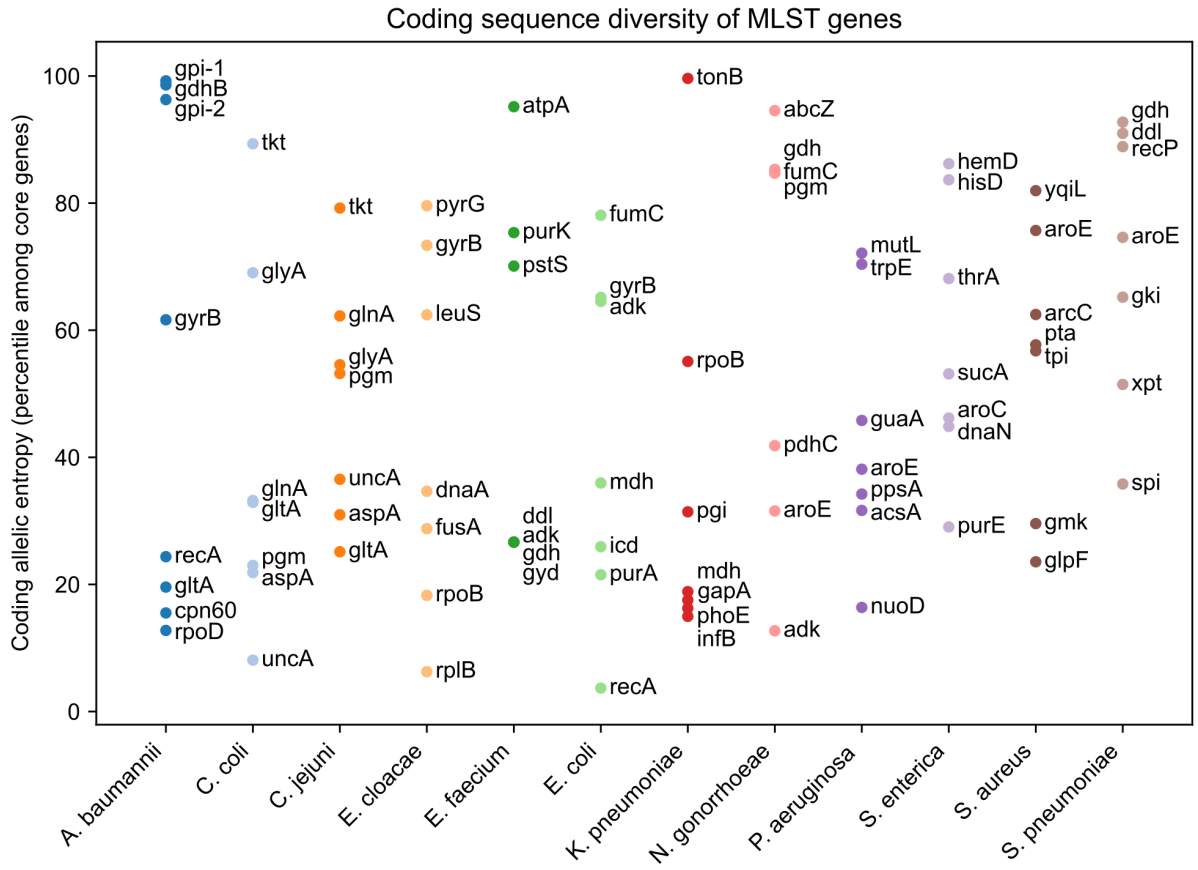


Figure A.9. Coding allelic entropies of genes used in MLST typing schemes, as percentiles among all core genes of the corresponding species. For *A. baumannii*, the MLST gene *gpi* was mapped to two pangenome gene clusters denoted gpi-1 and gpi-2, both of which include *gpi* variants defined in the *A. baumannii* MLST typing scheme.

Figure A.10. Domains with significant mutation depletion across multiple species. Species-specific mutation depletion for gene-domain pairs with significant multispecies mutation depletion (Bootstrap test, FDR < 0.05, Benjamini-Hochberg correction). Domains related to aminoacyl-tRNA synthetases are labeled purple. White cells correspond to domains that could not be annotated within the species' consensus sequence of the parent protein.

A.3 Supplementary Tables

Table A.1. Genome counts, abbreviations, and taxon IDs for species examined in the development of comparative pangenomic methods.

Species	Genomes	Abbreviation	Taxon ID
Acinetobacter baumannii	1021	AcB	470
Campylobacter coli	269	CaC	195
Campylobacter jejuni	451	CaJ	197
Enterobacter cloacae	104	EnC	550
Enterococcus faecium	169	EnF	1352
Escherichia coli	1970	EsC	562
Klebsiella pneumoniae	1895	KlP	573
Neisseria gonorrhoeae	391	NeG	485
Pseudomonas aeruginosa	595	PsA	287
Salmonella enterica	1145	SaE	28901
Staphylococcus aureus	1483	StA	1280
Streptococcus pneumoniae	3183	StP	1313

Table A.2. Heaps' Law parameter estimates for 12 species, fitted by either randomly shuffling all genomes or one genome per MLST. The two methods are denoted "By Genome" and "By MLST," respectively. Means and standard deviations from 100 iterations are shown for each species, parameter, and method. Species are sorted by Heaps' Law λ, estimated using the MLST method.

	Heaps' Law, λ		Heaps' Law, κ	
Species	By Genome	By MLST	By Genome	By MLST
N. gonorrhoeae	0.205 ± 0.009	0.198 ± 0.008	2207 ± 110	2476 ± 70
E. faecium	0.119 ± 0.012	0.218 ± 0.018	3450 ± 208	3080 ± 99
S. aureus	0.340 ± 0.010	0.295 ± 0.013	1492 ± 104	2092 ± 135
C. coli	0.312 ± 0.016	0.301 ± 0.014	1401 ± 122	1652 ± 94
C. jejuni	0.301 ± 0.014	0.319 ± 0.024	1543 ± 127	1689 ± 148
S. pneumoniae	0.325 ± 0.009	0.362 ± 0.012	2230 ± 158	2012 ± 147
E. cloacae	0.384 ± 0.023	0.428 ± 0.025	4330 ± 451	4142 ± 382
S. enterica	0.342 ± 0.019	0.430 ± 0.032	3598 ± 461	3529 ± 499
A. baumannii	0.361 ± 0.031	0.452 ± 0.038	2507 ± 542	2863 ± 494
P. aeruginosa	0.426 ± 0.016	0.454 ± 0.029	3715 ± 375	3291 ± 505
K. pneumoniae	0.406 ± 0.013	0.455 ± 0.015	3193 ± 310	3645 ± 274
E. coli	0.412 ± 0.021	0.467 ± 0.019	4053 ± 657	3732 ± 389

Table A.3. Evaluating accuracy of Heaps' Law fits, based on either randomly shuffling all genomes or one genome per MLST. The two methods are denoted "By Genome" and "By MLST," respectively. Heaps' Law was fit to the first half of genomes in pangenome size curves ("fitting region") from either method and accuracy was evaluated against the second half ("extrapolation region"). The mean absolute error (MAE) for each region was computed, and the median MAE across 100 iterations is shown, as well as relative error between the MLST vs. genome methods. Species are sorted by relative median MAE during extrapolation.

	Fitting Region			Extrapolation Region		
Species	By MLST	By Genome	Ratio	By MLST	By Genome	Ratio
N. gonorrhoeae	28	49	0.571	100	206	0.485
S. aureus	78	146	0.534	254	484	0.525
S. pneumoniae	77	121	0.636	149	270	0.552
E. coli	288	659	0.437	927	1595	0.581
A. baumannii	202	378	0.534	707	1189	0.595
C. jejuni	61	65	0.938	125	182	0.687
C. coli	45	66	0.682	97	129	0.752
E. cloacae	201	245	0.820	557	618	0.901
E. faecium	23	76	0.303	105	109	0.963
K. pneumoniae	261	372	0.702	680	658	1.033
S. enterica	285	325	0.877	970	629	1.542
P. aeruginosa	438	369	1.187	1425	714	1.996

Table A.4. Gene frequency cutoffs and gene counts for the core, accessory, and unique genomes of 12 species.

		Fitting Region		Extrapolation Region			
Species	Genomes	Core	Unique	Core	Accessory	Unique	Total
E. cloacae	104	102 (98.3%)	9 (8.3%)	2906 (11.3%)	4533 (17.7%)	18239 (71.0%)	25678
E. faecium	169	162 (95.8%)	10 (5.8%)	2155 (34.2%)	2403 (38.1%)	1752 (27.8%)	6310
C. coli	269	263 (97.8%)	21 (7.8%)	1331 (16.6%)	1046 (13.0%)	5645 (70.4%)	8022
N. gonorrhoeae	391	377 (96.4%)	25 (6.4%)	1819 (23.9%)	1675 (22.0%)	4123 (54.1%)	7617
C. jejuni	451	445 (98.6%)	39 (8.6%)	1237 (12.7%)	1432 (14.7%)	7099 (72.7%)	9768
P. aeruginosa	595	575 (96.6%)	39 (6.6%)	4585 (8.2%)	3828 (6.8%)	47622 (85.0%)	56035
A. baumannii	1021	999 (97.9%)	81 (7.9%)	2508 (8.3%)	3412 (11.2%)	24455 (80.5%)	30375
S. enterica	1145	1121 (97.9%)	90 (7.9%)	3081 (7.7%)	4087 (10.2%)	32792 (82.1%)	39960
S. aureus	1483	1460 (98.4%)	125 (8.4%)	2012 (11.2%)	1796 (10.0%)	14148 (78.8%)	17956
K. pneumoniae	1895	1833 (96.7%)	127 (6.7%)	3981 (5.9%)	3368 (5.0%)	60252 (89.1%)	67601
E. coli	1970	1921 (97.5%)	148 (7.5%)	3020 (3.3%)	5046 (5.5%)	82897 (91.1%)	90963
S. pneumoniae	3183	3113 (97.8%)	248 (7.8%)	1296 (4.3%)	2254 (7.4%)	26856 (88.3%)	30406

Table A.5. Correlations between three types of intraspecies sequence diversity for core genes across 12 species. Variant types are coding (protein sequences), 5′ intergenic (5′ IG, 300nt upstream and adjacent to the start codon), and 3′ intergenic (3′ IG, 300nt downstream and adjacent to the stop codon).

	Pearson Correlations			Spearman Correlations		
Species	Coding vs. 5′ IG	Coding vs. 3′ IG	5′ IG vs. 3′ IG	Coding vs. 5′ IG	Coding vs. 3′ IG	5′ IG vs. 3′ IG
E. cloacae	0.344	0.283	0.321	0.315	0.235	0.226
E. faecium	0.597	0.553	0.628	0.505	0.451	0.491
C. coli	0.429	0.440	0.417	0.338	0.340	0.294
N. gonorrhoeae	0.308	0.260	0.213	0.271	0.217	0.209
C. jejuni	0.287	0.303	0.395	0.269	0.300	0.376
P. aeruginosa	0.281	0.147	0.216	0.272	0.139	0.202
A. baumannii	0.380	0.272	0.311	0.384	0.253	0.296
S. enterica	0.280	0.211	0.247	0.276	0.208	0.219
S. aureus	0.258	0.205	0.233	0.247	0.182	0.227
K. pneumoniae	0.358	0.244	0.246	0.353	0.241	0.235
E. coli	0.320	0.263	0.321	0.297	0.239	0.257
S. pneumoniae	0.283	0.208	0.165	0.275	0.221	0.197
Mean	**0.344**	**0.283**	**0.309**	**0.317**	**0.252**	**0.269**
Median	**0.314**	**0.262**	**0.279**	**0.286**	**0.237**	**0.231**

A.4 Supplementary Datasets

Dataset A.1. PATRIC genome IDs for all genomes used for comparative pangenomics analysis.

Dataset A.2. MLST annotations for genomes used for comparative pangenomics analysis.

Dataset A.3. Summary of double power function fits to cumulative gene frequency distributions for 12 species. Includes derived thresholds for classifying genes as core, accessory, or unique. Also includes for each species the minimum frequency to classify a gene as core, maximum frequency to classify a gene as unique, sizes of the core/accessory/unique genomes, R^2 and MAE of the fit, and the five fitted parameters.

Dataset A.4. Log odd ratios and Fisher's exact test p-values for enrichment between gene functional groups (COGs, GO terms) and various gene categories (core, accessory, unique, highest sequence diversity, lowest sequence diversity) within each species.

Dataset A.5. Predicted gene names, COG functional categories, and TraDIS *E. coli* essentiality predictions from Goodall et.al. 2018 [6] for the 168 genes observed in the core genome of all 12 species.

Dataset A.6. Domain mutation enrichment analysis across 12 core genomes. For each gene-domain pair, includes the estimated mutation enrichment as domain entropy percentile (species-specific and species-wide averages), Bootstrap test p-values, domain InterPro accession IDs, and domain descriptions. Also includes assignment of functional categories to AARS-related domains.

A.5 References

[1] Keith A Jolley and Martin C J Maiden. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11(1):595, December 2010.

[2] James J Davis, Alice R Wattam, Ramy K Aziz, Thomas Brettin, Ralph Butler, Rory M Butler, Philippe Chlenski, Neal Conrad, Allan Dickerman, Emily M Dietrich, Joseph L Gabbard, Svetlana Gerdes, Andrew Guard, Ronald W Kenyon, Dustin Machi, Chunhong Mao, Dan Murphy-Olson, Marcus Nguyen, Eric K Nordberg, Gary J Olsen, Robert D Olson, Jamie C Overbeek, Ross Overbeek, Bruce Parrello, Gordon D Pusch, Maulik Shukla, Chris Thomas, Margo VanOeffelen, Veronika Vonstein, Andrew S Warren, Fangfang Xia, Dawen Xie, Hyunseung Yoo, and Rick Stevens. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. *Nucleic Acids Res.*, 48(D1):D606–D612, January 2020.

[3] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[4] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, March 2020.

[5] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, August 2017.

[6] Emily C A Goodall, Ashley Robinson, Iain G Johnston, Sara Jabbari, Keith A Turner, Adam F Cunningham, Peter A Lund, Jeffrey A Cole, and Ian R Henderson. The essential genome of *Escherichia coli* K-12. *MBio*, 9(1), March 2018.

[7] Roger Koenker and Kevin F Hallock. Quantile regression. *J. Econ. Perspect.*, 15(4):143–156, November 2001.

[8] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the Python in Science Conference*. SciPy, 2010.

[9] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, 30(4):772–780, April 2013.

[10] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014.

Appendix B

A machine learning approach for identifying antimicrobial resistance determinants in pangenomes - Supplementary Information

B.1 Methods

B.1.1 Genome selection and pangenome assembly

For constructing the *S. aureus*, *P. aeruginosa*, and *E. coli* pangenomes, genomes on PATRIC [1] were filtered to those that met the following criteria: 1) at least one experimentally measured AMR phenotype (MIC, disk diffusion, agar dilution, Vitek2) is associated with the genome on PATRIC, 2) sequence data is not plasmid-only, and 3) there are at most 100 contigs for *S. aureus* assemblies or at most 250 contigs for *P. aeruginosa* (for *E. coli*, contig filtering was not applied, and only 4 out of 1,588 genome assemblies had more than 250 contigs). Genome IDs for selected genomes are available in Dataset B.1. PATRIC genome annotations were used to construct pangenomes using CD-HIT v4.8.1 [2]. The sequence identity threshold was set at 0.8 and the word length was set to the default of 5.

For each pangenome, the number of genomes each gene cluster was observed in was computed. The number of core genes was calculated from an increasingly relaxed

threshold for core gene, i.e. the maximum number of genomes allowed to be missing a core gene; in all three cases the core genome size stabilizes by a threshold of 10, which is the threshold used to identify core genes in all subsequent analyses (Fig. B.1), and symmetrically to identify unique genes (i.e. genes present in no more than 10 genomes). Within each pangenome, the unique amino acid sequence variants or "alleles" of each gene were enumerated (Table B.1).

B.1.2 Mathematical representation of pangenomes and AMR phenotypes

For each pangenome, each genome was encoded as a binary vector, based on the presence or absence of every gene cluster and every allele of every gene cluster observed for that species; this yielded a sparse binary matrix encoding the genetic content at both the gene and allele level (Fig. 3.2a). The number of features was reduced by only analyzing core genes at the allele level, and analyzing non-core genes at the gene level. For each drug, experimental AMR phenotypes were converted to binary vectors by directly converting raw PATRIC AMR annotations "Susceptible" to 0 and both "Resistant" and "Intermediate" to 1. The distribution of binarized phenotypes, typing methods, and typing standards associated with these annotations are in Table B.2.

B.1.3 Identification of known AMR genes in the *S. aureus* pangenome

Known AMR genes against antibiotics examined for *S. aureus* were compiled from literature and the CARD database, retrieved on November 26, 2018 [3]. CARD entries were filtered down to those referencing any of the drugs examined (ciprofloxacin, clindamycin, erythromycin, gentamicin, trimethoprim, sulfamethoxazole, tetracycline) or their drug classes (fluoroquinolone, lincosamide, macrolide, aminoglycoside, trimethoprim, sulfonamide, tetracycline). Representative protein sequences for these genes were taken from either UniProt or CARD (Dataset B.2) and were aligned to the alleles in the *S.*

aureus pangenome using blastp. Hits with an e-value below 10^{-50} and identity $> 90\%$ were treated as true AMR determinants.

Curated AMR genes were classified into four broad mechanistic categories (Fig. B.2a): 1) Mutant Site, genes that are direct targets to a given drug that can acquire AMR-conferring mutations, 2) Efflux, genes involved in efflux pumps or regulation of efflux pumps, 3) Modifies Site, genes that protect the direct targets of a given drug, such as by ribosomal modification, and 4) Modifies Drug, genes that cleave, modify, or otherwise inactivate the drug molecule. The frequency and LOR for alleles of curated AMR genes were plotted (Fig. B.2b-c). As most such alleles were very rare and observed AMR phenotypes for many drugs were highly biased towards resistant cases, a modified form of LOR with weighted pseudocounts was computed to more accurately capture the extent of enrichment and address frequent zeroes in contingency tables:

$$\text{LOR} = \log_2 \left(\frac{\left(AR + \frac{R}{R+S}\right)\left(NS + \frac{S}{R+S}\right)}{\left(AS + \frac{S}{R+S}\right)\left(NR + \frac{R}{R+S}\right)} \right), \quad \begin{matrix} R = AR + NR \\ S = AS + NS \end{matrix}$$

where AR is the number of resistant genomes with the allele, AS is the number of susceptible genomes with the allele, NS is the number of susceptible genomes without the allele, and NR is the number of resistant genomes without the allele. This adjustment has the following properties: 1) an allele that is not observed ($AR = AS = 0$) has a non-informative LOR of 0, 2) a universal allele observed in all genomes ($NS = NR = 0$) has a non-informative LOR of 0, and 3) the total adjustment to the contingency table is 2, which is common for other pseudocounts strategies for addressing contingency tables with zeroes, such as adding 0.5 to all cells.

B.1.4 Comparison of statistical tests and SVM ensemble models for predicting AMR determinants in *S. aureus*

For the *S. aureus* pangenome, Fisher's exact test and Cochran-Mantel-Haenszel's test (CMH) were applied between each drug and genetic feature. For CMH, genome subgroups were determined through hierarchical clustering on the genetic feature matrix, implemented in SciPy using pairwise Jaccard distances and average linkage; these clusters were found to be consistent with metadata regarding genome subtype (Fig. 3.1). The two smallest clusters were also treated as a single subgroup for CMH testing. Features were filtered based on significance after either a Bonferroni correction (FWER < 0.05) or Benjamini-Hochberg correction (FDR < 0.05) (Table B.3), then ranked by p-value with fractional ranking for ties.

For each drug, four different types of SVM ensembles of 500 SVMs each were trained to predict AMR phenotype from the *S. aureus* genetic feature matrix, using different resampling strategies (Fig. 3.2a). Within an ensemble, each of the 500 constituent models were trained using one of the following sampling strategies:

1. SVM: Random subsets of 80% of genomes.
2. SVM-RSE: Random subspaces with 80% of genomes and 50% of features.
3. SVM-RSE-U: From each hierarchical clustering subgroup, randomly sample n genomes, where $n = 80\%$ of the size of the smallest cluster. Randomly select 50% of features.
4. SVM-RSE-O: From each hierarchical clustering subgroup, randomly sample n genomes, where $n = 80\%$ of the size of the largest cluster. Randomly select 50% of features.

SVMs were implemented in scikit-learn, using square hinge loss weighted by class frequency to address class imbalance issues. L1 regularization was included to enforce

sparsity for feature selection. For each species-drug case, genomes without AMR phenotype data were ignored. Features were ranked based on the average feature weight across all SVMs in a given ensemble; in cases where features were subsampled, a feature's average weight was calculated from only SVMs that had access to that feature. For each drug, this yielded a list of top hits associated with resistance (largest positive weights/top ranking features) and a list of top hits associated with susceptibility (largest negative weights/bottom ranking features). Both statistical tests and the four SVM ensemble types were compared based on the number and rank of a priori curated AMR determinants detected (Fig. 3.2b).

B.1.5 Application of SVM-RSE to predict AMR determinants in *S. aureus*, *P. aeruginosa*, and *E. coli*

The SVM-RSE approach described earlier was applied to a total of 16 species-drug cases across the three species to identify genetic features associated with AMR from experimentally observed AMR phenotypes (Fig. 3.3a). For each case, after training an SVM-RSE on the species' genetic feature matrix and drug's AMR phenotype vector, the top 50 hits associated with resistance were assessed for known AMR determinants and verified through a literature search (Table 3.2). In examining the *P. aeruginosa*-amikacin case, known aminoglycoside-modifying enzymes were identified in the pangenome using the same process for curating *S. aureus* AMR genes (Table B.4). LORs were computed using the method as for the curated *S. aureus* AMR genes.

To assess the "null" level of predictive performance, another SVM-RSE was trained for each species-drug case in which AMR phenotypes were randomly shuffled. For both the original and permuted ensembles, the performance of each of their 500 constituent SVMs was evaluated by computing the Matthews correlation coefficients (MCCs) on out-of-bag samples, or genomes not used for training (Fig. B.3). To assess the overall predictive performance of the ensemble, the SVM-RSE approach was treated as a voting

classifier, in which the SVM-RSE prediction is the majority prediction of its 500 constituent SVMs. 5-fold cross validation experiments with the SVM-RSE were conducted for each species-drug case, and the average and standard error of the accuracy, MCC, precision, recall, and area under receiver operating curve (AUROC) for the testing set across all folds were computed (Fig. 3.3b). ROC curves for each fold were also computed (Fig. B.4).

B.1.6 Assessing stability of SVM-RSE selected features for different core gene thresholds

The core genome of each pangenome was defined using three thresholds: the set of genes missing in 1) no more than 10 genomes (default), 2) no more than 2% of all genomes, and 3) no more than 10% of all genomes. These core gene thresholds were used to encode each pangenome in terms of its core gene alleles and non-core genes as described earlier, yielding three distinct matrices per pangenome (i.e. the genome by gene and allele matrix in Fig. 3.2a). The SVM-RSE analysis was repeated for each pangenome matrix to predict AMR-associated features for all species-drug cases. The top 50 resistance-associated features and top 50 susceptibility-associated features yielded by each threshold for each species-drug case were identified and combined into a single top feature set for each threshold; pairs of these top feature sets across different thresholds were compared by identifying what fraction of features were shared (selected under both thresholds), not shared (available under both thresholds but selected in only one), or differentially encoded (available under only one threshold and impossible to be shared) (Fig. B.5).

B.1.7 Assessing enrichment for highly variable genes among selected features

The total number of unique alleles observed for each core gene ("allele count") was computed for each species' pangenome. For each species-drug case, the mean and median allele count of core genes for which at least one allele was selected by SVM-RSE to be

associated with resistance ("selected core genes") was computed. This was compared to the mean and median allele count of all core genes for each species (Fig. B.6a-b). For each species, the full allele count distribution of selected core genes was compared to that of all core genes for the species-drug case with the largest difference in mean allele count between selected and all core genes (Fig. B.6c-e).

B.1.8 Assessing enrichment for plasmid genes among selected features

To identify which genetic features were located on plasmids, every contig in every genome assembly was compared to known plasmids in PLSDB (version 2019_10_07) [4] using MASH [5] set to a distance threshold of 0.01, i.e. contigs with distance < 0.01 to a known plasmid were marked as plasmid contigs. All alleles found on plasmid contigs and all genes for which a majority of unique alleles were found on plasmid contigs were treated as plasmid features; all other features were treated as chromosomal. For each species-drug case, the number of plasmid and chromosomal features in the top 50 features selected by SVM-RSE was computed along with the odds ratio for plasmid features with respect to all features for that species. As plasmid features are predominantly non-core genes, this calculation was also repeated for just non-core features to more accurately reflect enrichment for plasmid features (Table B.5).

B.1.9 Analysis of *gyrA* and *parC* mutations with respect to fluoroquinolone resistance

The top 10 hits associated with either resistance (highest feature weights) or susceptibility (lowest feature weights) for the *S. aureus*-ciprofloxacin, *P. aeruginosa*-levofloxacin, and *E.coli*-ciprofloxacin cases were filtered down to just alleles of *gyrA* and *parC*. Mutations for these alleles were called relative to the corresponding protein sequence in the following reference genomes: N315 (NC_002745.2) for *S. aureus*, PAO1 (NC_002516.2) for *P. aeruginosa* and K12 MG1655 (U00096.3) for *E. coli*. Individual mutations for these

alleles were compared to those known to confer resistance to FQs (Table 3.3). Across all *gyrA* and *parC* alleles in each pangenome, the most abundant alleles were selected (top 8 for *S. aureus* and *P. aeruginosa*, top 12 in *E. coli*) and the LOR for resistance to FQ was computed for each allele individually, as well as for each *gyrA/parC* pairing to identify potential interactions (Fig. B.7). This pairwise interaction analysis was repeated for all pairs between the top 10 hits associated with resistance by SVM-RSE for the three FQ cases (Fig. B.8).

B.1.10 Extracting candidate AMR determinants from SVM-RSE weights

For each of the 16 species-drug cases, the top 10 hits associated with resistance were filtered down to higher confidence candidates for novel AMR determinants using the following steps: Features already known to be associated with AMR were removed. Features annotated as transposases, phage proteins, or other mobile elements were also removed, as their function may be attributable to their position rather than just their presence or sequence. For core gene alleles, mutations were called relative to the corresponding gene in a reference genome (same as in the FQ case study), and only alleles with at least one mutation highly enriched for resistance were kept (>95% of genomes with the mutation are resistant). These mutations were further characterized by their location in predicted domains or other structural features from InterPro (Table 3.4, Fig. 3.4); only mutations present in at least 5 genomes are shown. For non-core genes, the most common allele of the gene cluster was identified as the dominant allele, and genes with high sequence variability were filtered out to remove noisy gene calls (i.e. cases where >10% of the instances of that gene have an edit distance > 10 from the dominant allele). Of the remaining non-core genes, the dominant alleles were annotated using InterProScan [6] and further filtered down to those with at least one domain annotation. LORs for both core gene alleles and non-core genes were computed using the method as for the curated *S. aureus* AMR genes.

B.2 Supplemental Discussion

B.2.1 Analysis of genetic diversity in *S. aureus*, *P. aeruginosa*, and *E. coli* pangenomes

For each pangenome, genes were classified based on how frequently they were observed: core (missing in 0-10 genomes), accessory (missing in >10 genomes, present in >10 genomes), or unique (present in 1-10 genomes), (Table B.1, Fig. B.9a). This classification found 2,221, 4,700, and 3,107 genes to be core genes for *S. aureus*, *P. aeruginosa*, and *E. coli*, respectively, which is consistent with previously observed core genome sizes when accounting for the number of genomes examined and different thresholds for identifying core genes (Table B.6).

To quantify pangenome openness, or the propensity for new genomes to carry previously unidentified genes, the relationship between the number of genomes and the new gene rate was modeled using Heaps' Law [7]. From 2,000 random permutations of genome order for each pangenome, the Heaps' Law exponent, α (in which a value of 1 represents a closed pangenome and 0 represents a completely open pangenome), was estimated as 0.83 for *S. aureus* and 0.71 for both *P. aeruginosa* and *E. coli* (Fig. B.9b); this result is consistent with previous observations that *S. aureus* harbors a relatively closed pangenome [7].

To examine differences in function between core, accessory, and unique genes, eggNOG-mapper [8] was used to assign a Clusters of Orthologous Groups (COG) functional category to every gene in each pangenome. We find that the distributions of gene functions in the three core genomes are very similar, as well as for the three accessory genomes and three unique genomes (Fig. B.10, Fig. B.11); in other words, a gene's function is associated with its frequency across multiple species. This observation was quantified by computing $\log_2$ odds ratios (LORs) between the core/accessory/unique gene sets and each COG functional category (Fig. B.9c).

We find several functions to be enriched in either the core or non-core genomes for all three species (Table B.7, Table B.8). Genes related to translation, ribosomal structure, energy production/conversion, or the transport/metabolism of core metabolites were more likely to be found in the core genome ($p < 0.00001$), while genes related to DNA replication/recombination/repair or have unknown function were more likely to be found among accessory genes ($p < 0.00001$). The core-enriched functions recapitulate the essentiality of translation and core metabolism for survival, while non-core-enriched functions appear to be more niche-specific (non-core genes involved in DNA manipulation were comprised primarily of mobile elements). Ultimately, it appears that different species allocate similar fractions of their genes towards specific functional groups, at least among genes that have been functionally characterized.

B.2.2 Supplemental Discussion - Methods

Comparison of pangenome size and openness

Genes were categorized by frequency: for a pangenome of n genomes, a gene is 1) core, if found in all n genomes, 2) near-core, if missing from at most 10 and at least 1 genomes, 3) accessory, if found in at least 11 genomes and missing from at least 10 genomes, 4) near-unique, if found in at least 2 genomes and at most 10 genomes, or 5) unique, if found in exactly 1 genome (Fig. B.9a). Subsequent analyses combine core/near-core as just "core", and unique/near-unique as just "unique". Pangenome openness was estimated using Heaps' law:

$$\mathrm{NGR}(N) = kN^{-\alpha}$$

where N is the number of genomes, NGR is the new gene rate, or number of new genes introduced per genome, and k and α are fitted parameters [7]. NGR was estimated as the median NGR from permuting the order of genomes 2,000 times, and Heaps' law parameters were fitted by linear regression between log(NGR) and log(N) (Fig. B.9b).

Functional characterization of pangenomes

All alleles from all three pangenomes were assigned a Clusters of Orthologous Groups (COG) functional category using the public eggNOG-mapper v1 server [8]. Each gene cluster was assigned the majority COG category of its alleles, weighted by the number of genomes containing each allele. Rare instances in which genes had no majority COG or had a mixture of multiple COGs assigned were ignored, and COG distributions were visualized with and without the "S: Function Unknown" category (Fig. B.10).

To assess the stability of functional distributions, genes within each pangenome were categorized as core, accessory, or unique based on a similar threshold X from the pangenome size analysis: genes missing in up to X genomes were labeled core, genes in up to X genomes were labeled unique, and all other genes were labeled accessory. The distribution of COG categories for core, accessory, and unique genes were plotted as X increased from 1 to 10 (Fig. B.11). For enrichment analysis, the $\log_2$ odds ratio (LOR) and Fisher's exact test p-value between each COG category and each gene subset (core, accessory, unique) were computed with the threshold $X = 10$ (Fig. B.9c, Table B.7, Table B.8). LORs for a given COG category and gene subset were computed as follows:

$$\text{LOR} = \log_2\left(\frac{CI * DO}{CO * DI}\right)$$

where CI is the number of genes with the COG in the gene subset, CO is the number of genes with the COG outside the gene subset, DO is the number of genes with a different COG outside the gene subset, and DI is the number of genes with a different COG in the gene subset. Undefined LORs were replaced with 0, as all such cases had very few genes with which to examine significant enrichment. The category "B: Chromatin structure and dynamics" was ignored for this analysis as only one gene was annotated with that functional category.

B.3 Supplementary Figures

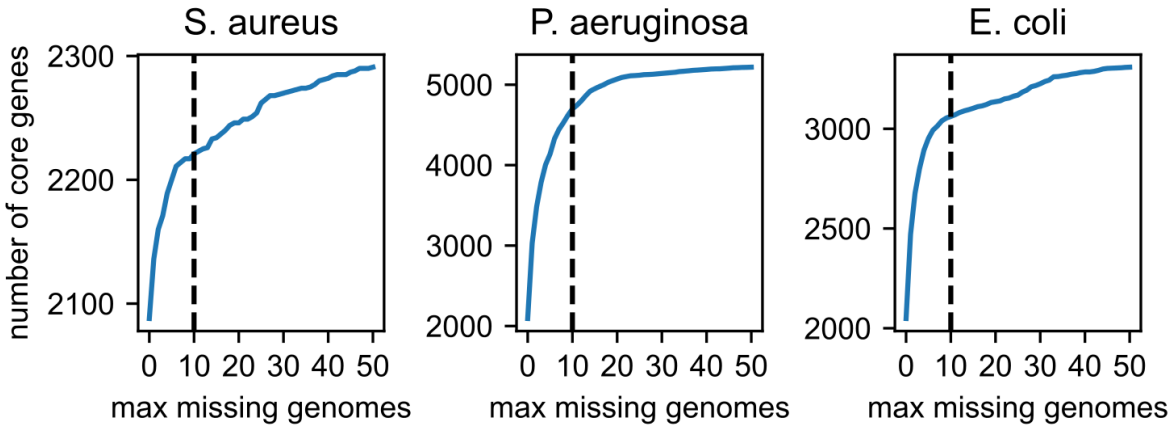


Figure B.1. Core genome size for *S. aureus*, *P. aeruginosa*, and *E. coli* at different core gene thresholds. For each pangenome, the threshold for classifying a gene as a core gene was relaxed from allowing at most 0 to at most 50 genomes to be missing the gene. The threshold of 10 genomes used for subsequent analyses is shown.

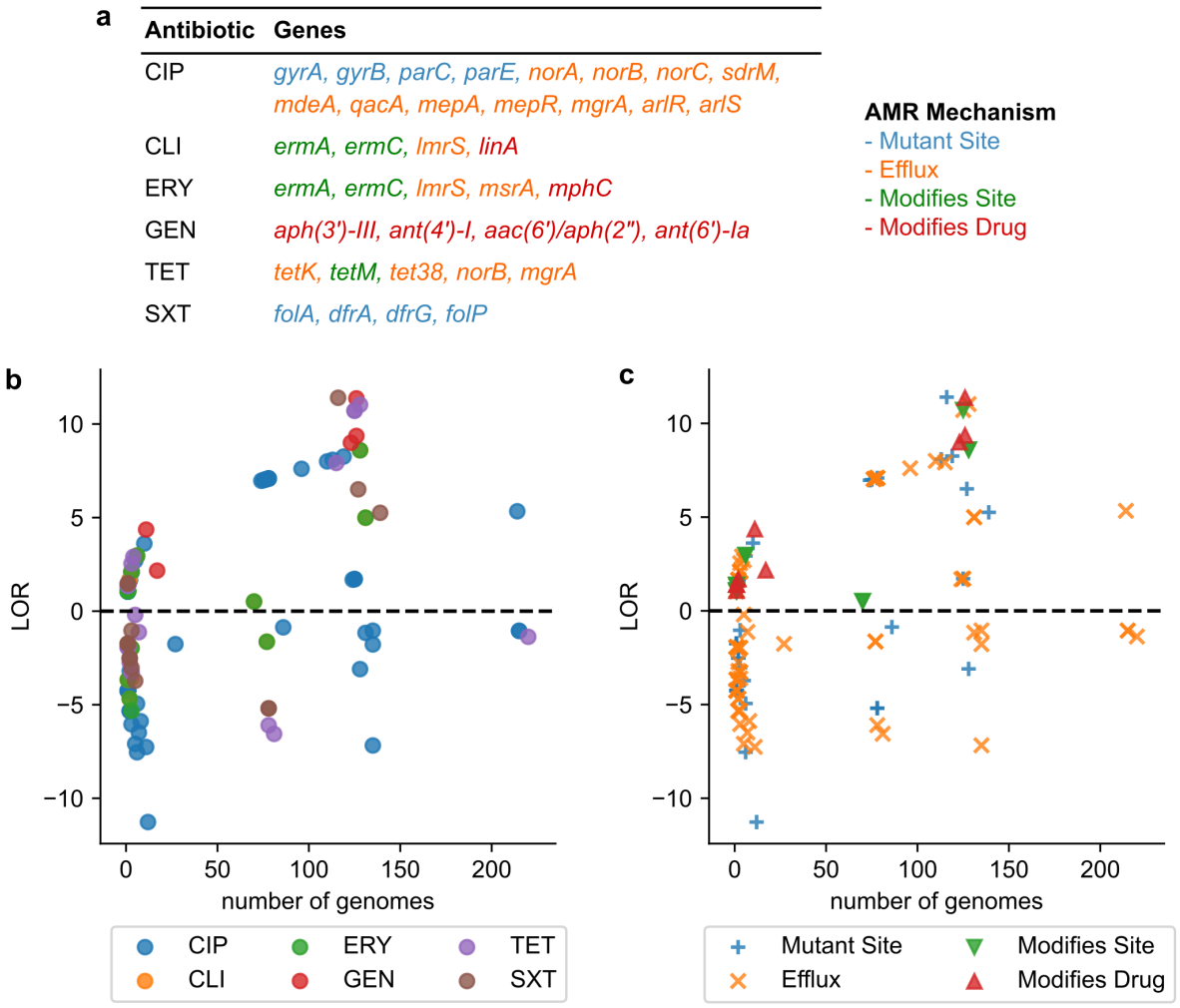


Figure B.2. Mechanism and distribution of known AMR genes identified in the *S. aureus* pangenome. (a) Each known AMR gene detected in the *S. aureus* pangenome was assigned to one of four broad mechanistic categories. For each allele of each known AMR gene, the number of genomes it is present in and the $\log_2$ odds ratio (LOR) for resistance against the appropriate drug was plotted, labeled by (b) drug or (c) mechanism.

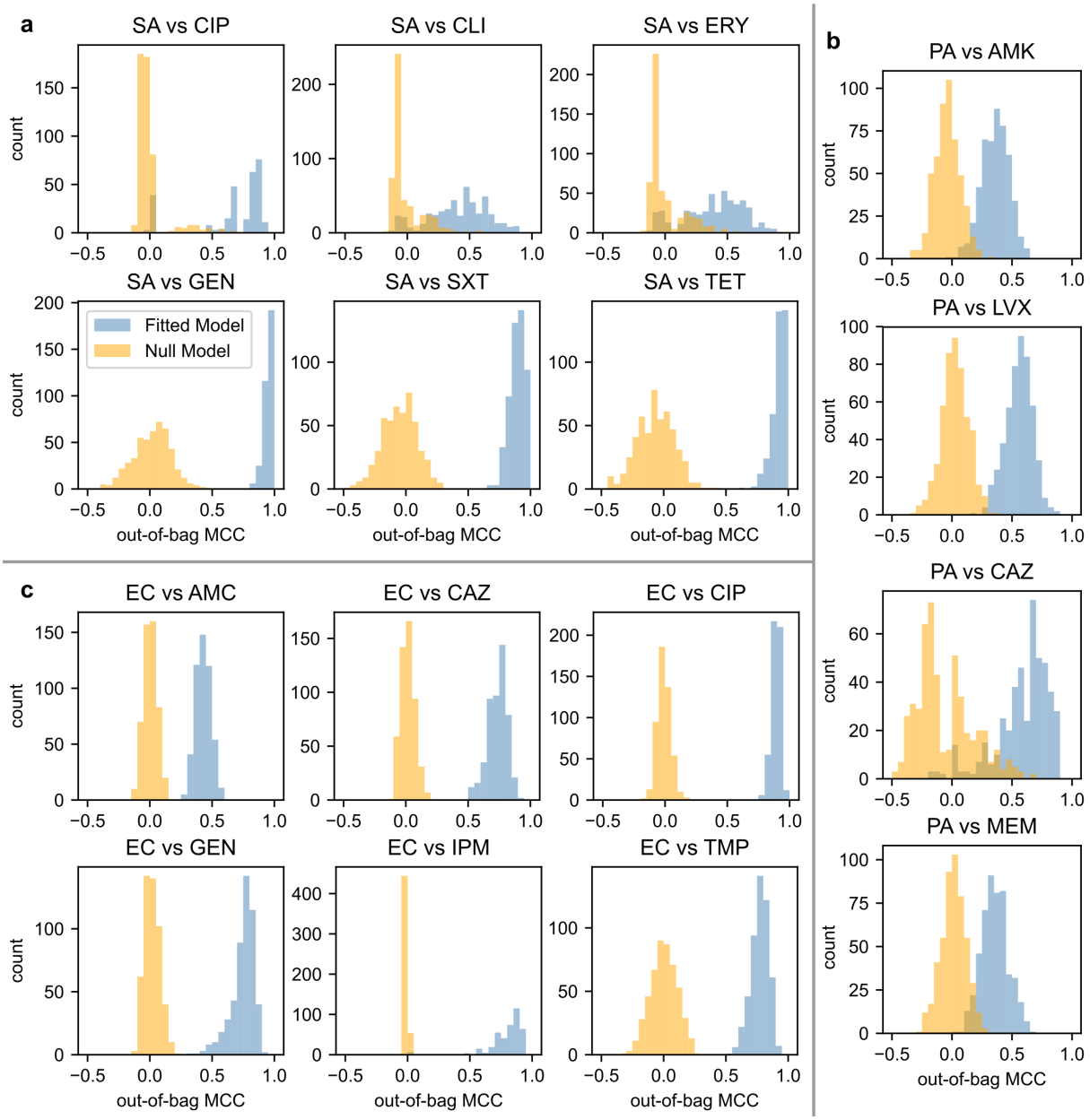


Figure B.3. Out-of-bag performance of individual SVMs in each SVM-RSE compared to null models for 16 species-drug cases. For each of the 16 species-drug cases across (a) *S. aureus*, (b) *P. aeruginosa*, and (c) *E. coli*, the performance of each of the 500 constituent SVMs used in the corresponding SVM-RSE was assessed as the Matthews correlation coefficients (MCCs) when predicting AMR phenotypes for out-of-bag genomes (those not used for training), shown in blue. The out-of-bag MCCs of constituent SVMs of SVM-RSEs trained using randomly shuffled AMR phenotype annotations are shown in orange.

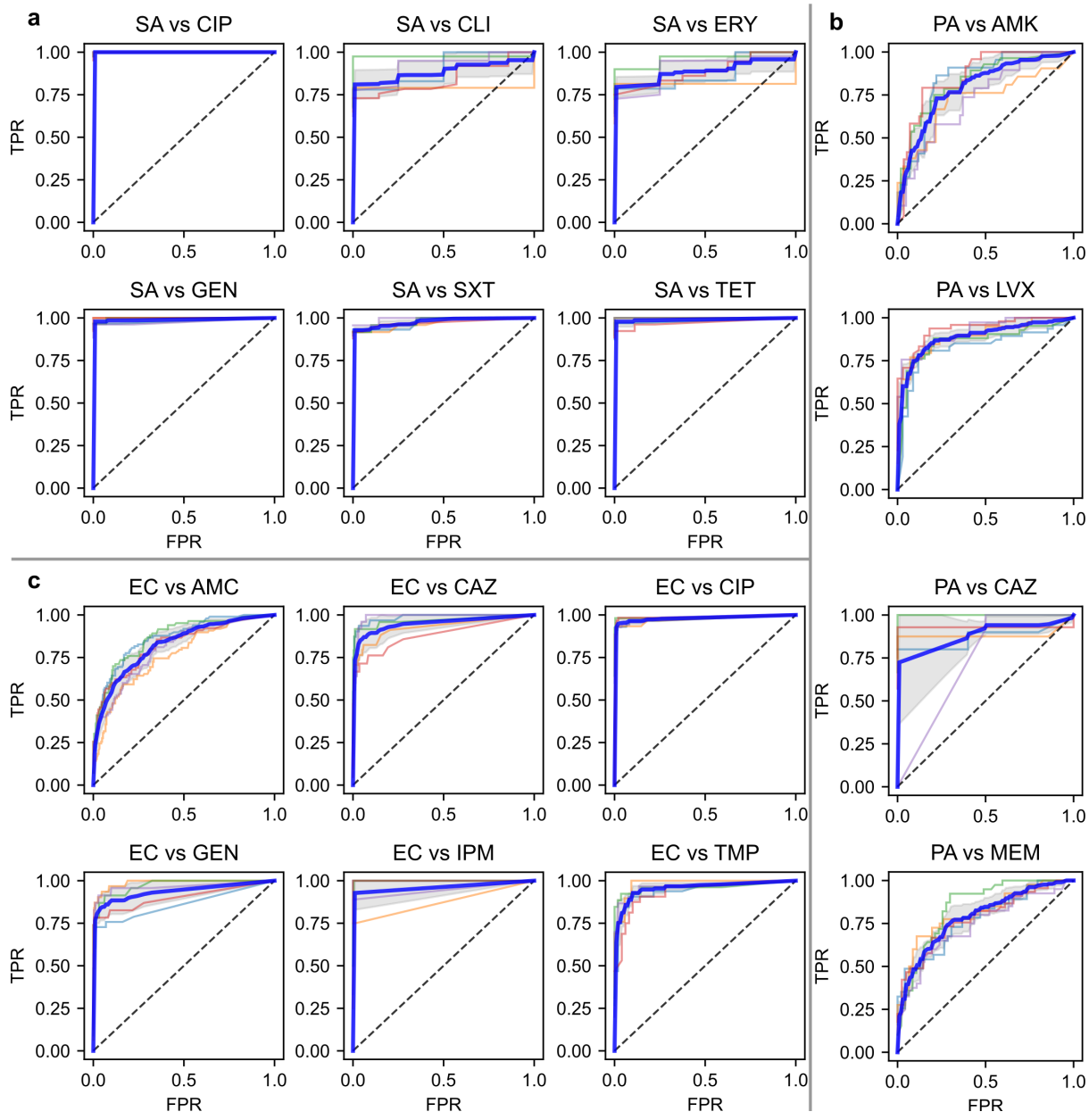


Figure B.4. Receiver operating curves of SVM-RSE models from 5-fold cross validation for 16 species-drug cases. ROC curves for each of the 16 species-drug cases across (a) *S. aureus*, (b) *P. aeruginosa*, and (c) *E. coli*. The dark blue curves are mean ROC curves from 5-fold cross validation, the lighter curves are individual ROC curves corresponding to each fold, and the grayed areas are within one standard deviation of the mean ROC curve.

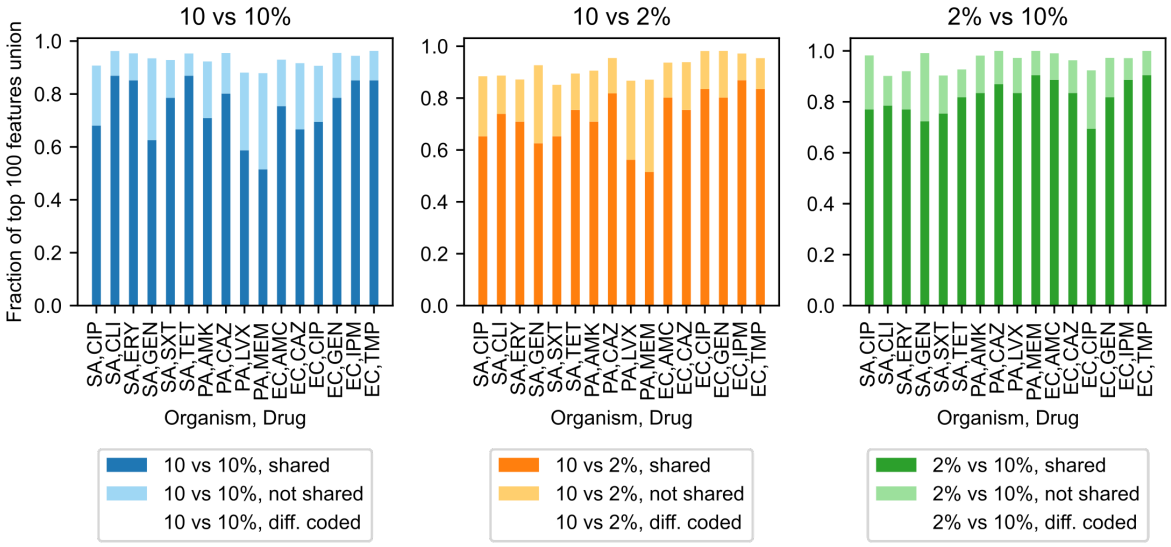


Figure B.5. Consistency of top SVM-RSE features by weight for different core gene thresholds. The top 100 features (top 50 resistance-associated and top 50 susceptibility-associated) were identified using SVM-RSE for three different core gene thresholds (10: missing from at most 10 genomes, 10%: missing from at most 10% of all genomes, 2%: missing from at most 2% of all genomes). For each pair of thresholds, the fraction of shared vs. non-shared features in the union of their top 100 feature sets were computed. Non-shared features were classified as either "not shared", where both representations contain the feature, or "diff. coded", where the feature is only available under one of the thresholds.

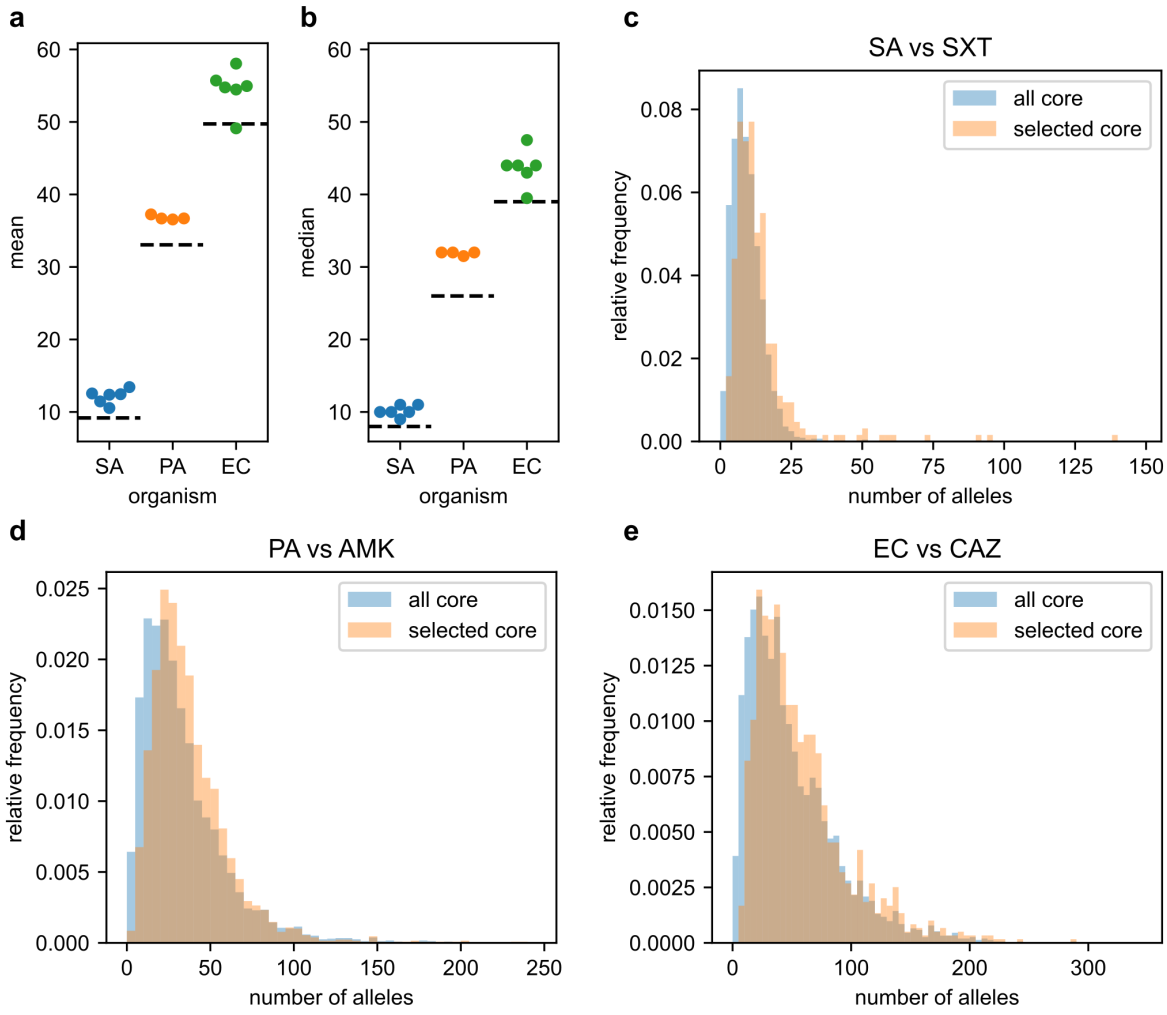


Figure B.6. Sequence variability of core gene alleles selected by SVM-RSE. For each species-drug case, the distribution of the number of alleles of all core genes was compared to that of core genes for which at least one allele was selected by SVM-RSE to be associated with resistance or susceptibility. The (a) mean and (b) median of the selected core gene allele count is shown for each case, compared to the mean and median for all core genes of the corresponding species (dotted lines). For each species, the allele count distributions are shown for the case with the largest difference in mean allele count, (c) *S. aureus* vs. sulfamethoxazole/trimethoprim, (d) *P. aeruginosa* vs. amikacin, and (e) *E. coli* vs. ceftazidime.

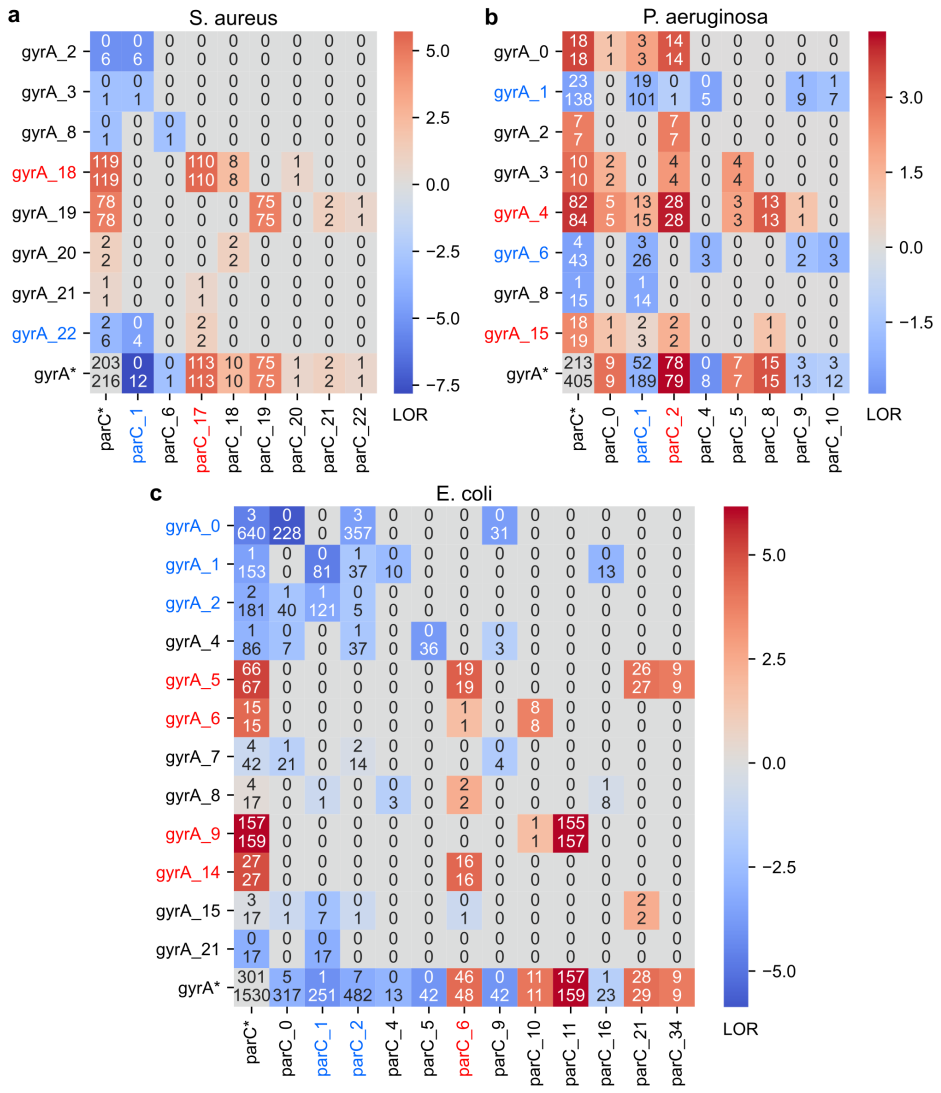


Figure B.7. Interactions between *gyrA* and *parC* alleles in resistance against fluoroquinolones. Log_2 odds ratios (LORs) for fluoroquinolone resistance were calculated for each *gyrA/parC* allele pairing and compared to individual alleles in (a) *S. aureus*, (b) *P. aeruginosa*, and (c) *E. coli*. Each cell shows the number of resistant genomes with the allele above, the total number of genomes with the allele below, and is colored by LOR; row and column totals do not add up as only the top 8 (for *S. aureus* and *P. aeruginosa*) or top 12 (for *E. coli*) most frequently observed *gyrA* and *parC* alleles are shown. Alleles among the top 10 features detected by SVM-RSE to be associated with fluoroquinolone resistance are in red, while those the SVM-RSE associated with susceptibility are in blue.

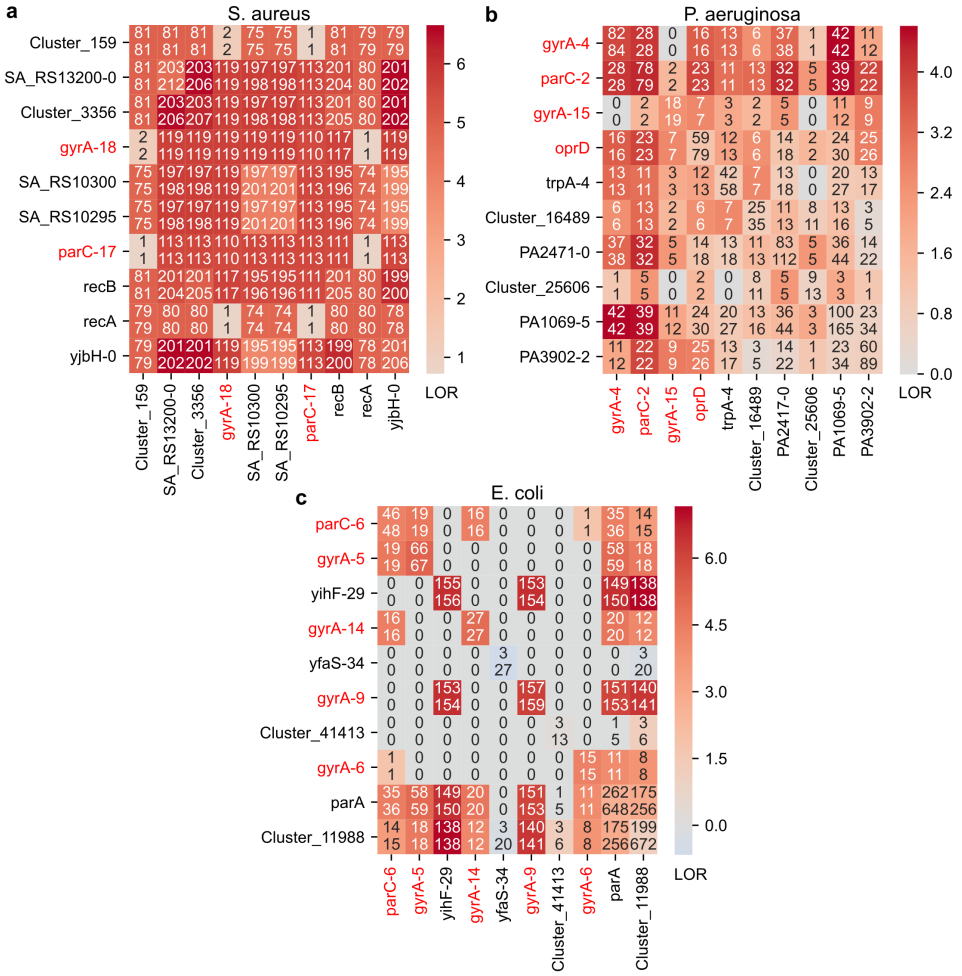


Figure B.8. Interactions between the top model-predicted hits for fluoroquinolone resistance. For each of the top 10 genetic features predicted by SVM-RSE to be associated with fluoroquinolone resistance in (a) *S. aureus*, (b) *P. aeruginosa*, and (c) *E. coli*, $\log_2$ odds ratios (LORs) for resistance were computed for each feature individually as well as for every top feature pairing. Each cell shows the number of resistant genomes with the allele above, the total number of genomes with the allele below, and is colored by LOR. Gene features are denoted by either their gene name, reference genome locus tag, or "Cluster_#" in cases the coding sequence could not be confidently mapped to a known gene. Allele features are denoted as "gene name-allele number". Features known to confer resistance are in red.

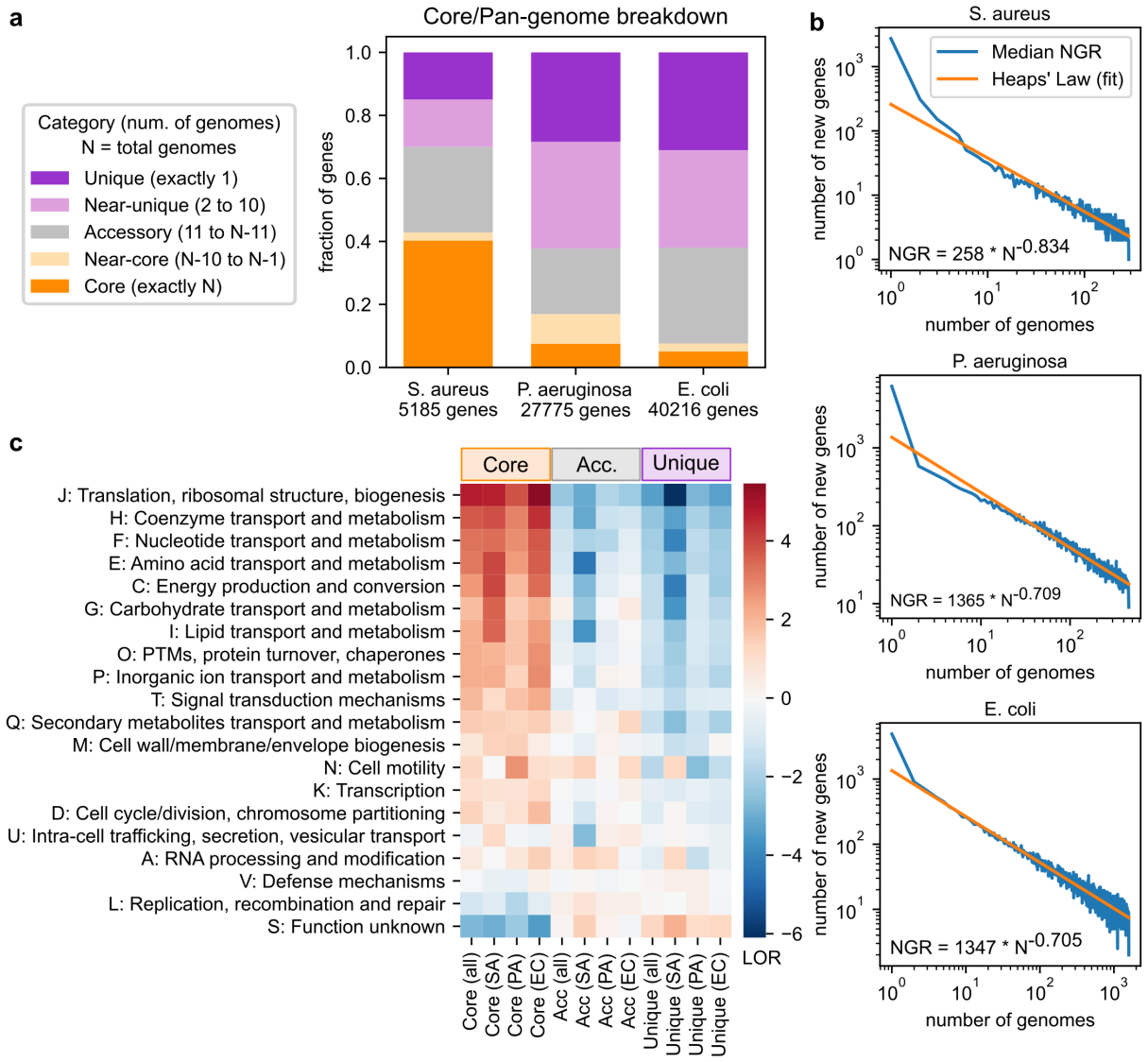


Figure B.9. Comparison of gene frequency, diversity, and functional distributions in the *S. aureus*, *P. aeruginosa*, and *E.coli* pangenomes. a) Distribution of genes categorized by frequency within each pangenome: i) core: present in all genomes, ii) near-core: missing from at most 10 genomes, iii) accessory: missing from >10 genomes and present in >10 genomes, iv) near-unique: present in 2-10 genomes, v) unique: present in exactly 1 genome. (b) Estimation of pangenome openness using Heaps' Law. The total number of genes (pangenome size) and number of genes in all genomes (core genome size) was computed as genomes were introduced sequentially from either the *S. aureus* (SA), *P. aeruginosa* (PA), or *E. coli* (EC) pangenome. Each value represents the median from 2,000 random permutations of genome order. The new gene rate (NGR) was fitted to Heaps' Law, in which a more negative exponent represents a more closed pangenome. (c) Log_2 odds ratios (LORs) between individual functional categories and the core, accessory (acc), and unique genomes for each species individually and combined.

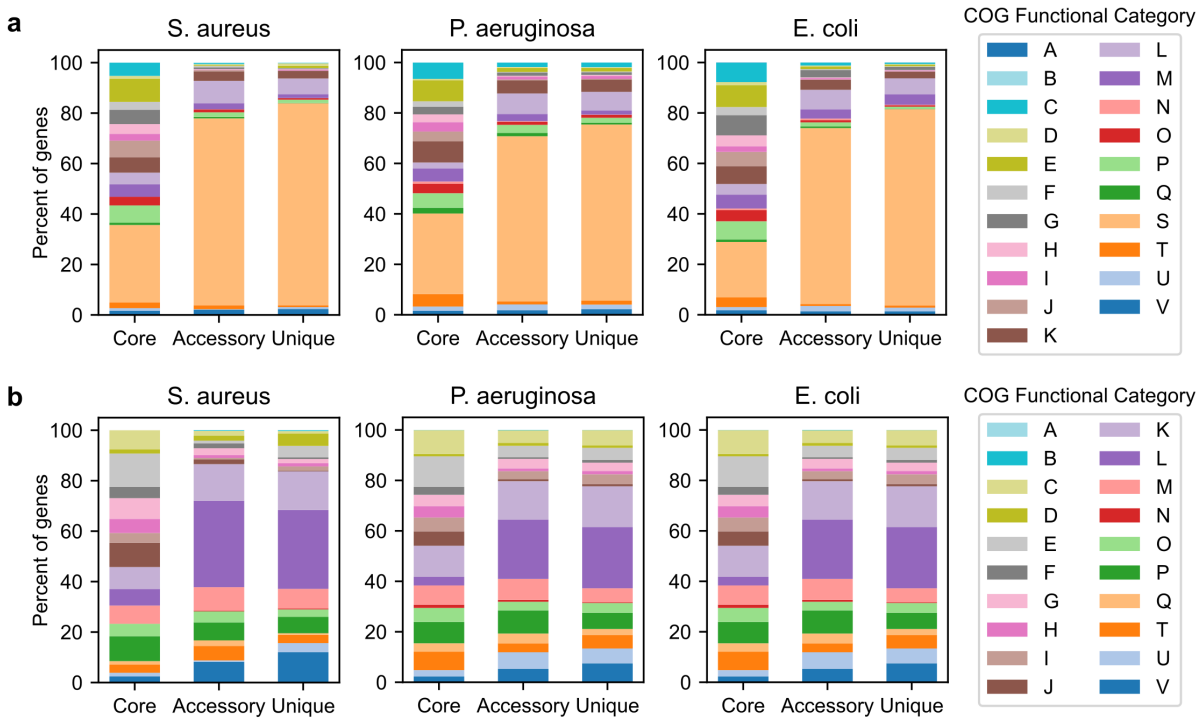


Figure B.10. Distribution of gene functions in the pangenomes of *S. aureus*, *P. aeruginosa*, and *E. coli*. The distribution of gene functional categories based on Clusters of Orthologous Groups (COGs) in the core, accessory, and unique genomes are shown, either (a) including, or (b) excluding the "S: Function unknown" category.

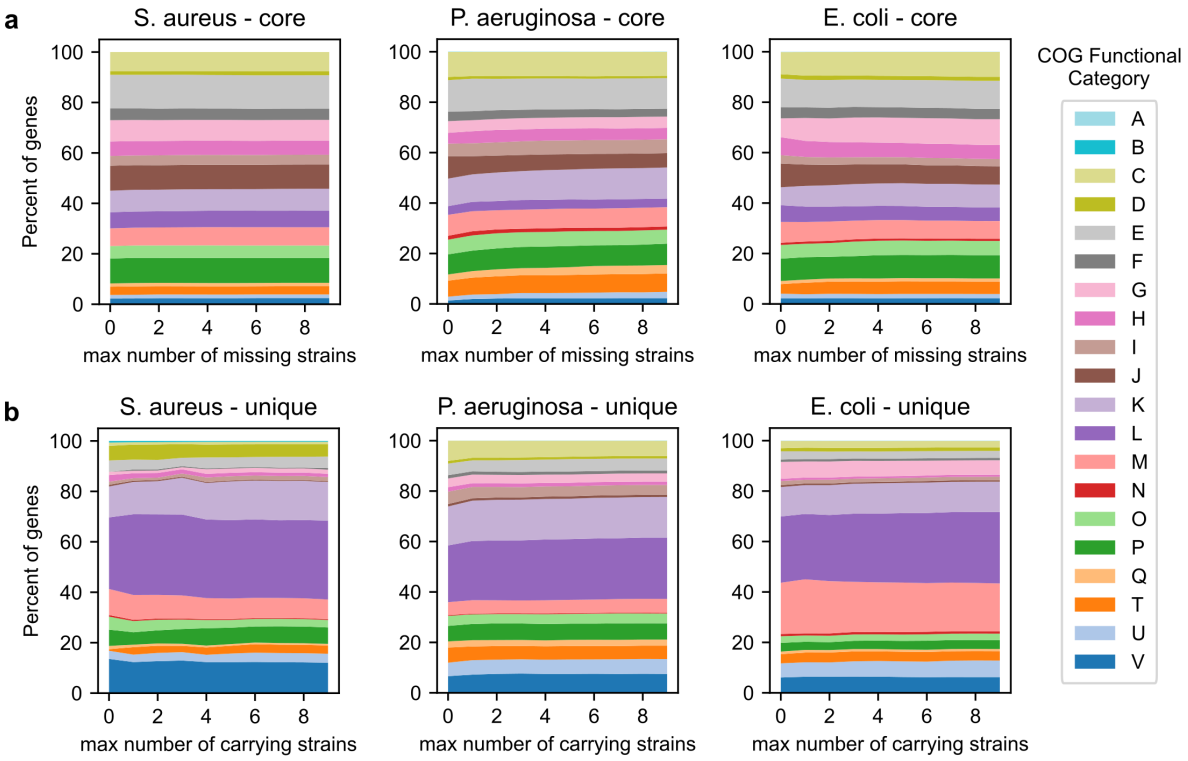


Figure B.11. Distribution of gene functions in the *S. aureus*, *P. aeruginosa*, and *E. coli* pangenomes for different thresholds for core and unique genes. For each species, the set of genes in the (a) core genome was assembled for different core gene thresholds (the maximum number of genomes allowed to be missing a core gene), and (b) analogously for unique genes comprising the unique genome (the maximum number of genomes allowed to carry a unique gene). The "S: Function unknown" functional category is not shown.

B.4 Supplementary Tables

Table B.1. Number of core, accessory, and unique genes and associated alleles in the pangenomes of *S. aureus*, *P. aerguinosa*, and *E. coli*.

	S. aureus	*P. aeruginosa*	*E. coli*
core genes	2221	4700	3062
accessory genes	1409	5795	12218
unique genes	1555	17280	24936
total genes	**5185**	**27775**	**40216**
core gene alleles	20390	155350	152264
accessory gene alleles	9083	89151	228449
unique gene alleles	2262	27304	39022
total alleles	**31735**	**271805**	**419735**

Table B.2. AMR phenotypes of PATRIC genomes and corresponding typing methods and standards for *S. aureus*, *P. aerguinosa*, and *E. coli*.

		Phenotype Counts		**Typing Method**				**Typing Standard**	
Species	Drug	Resistant	Susceptible	Agar Dilution	Disk Diffusion	MIC	Vitek2	CLSI	EUCAST
S. aureus	CIP	203	13	0	0	216	0	212	4
S. aureus	CLI	201	20	0	211	10	0	217	4
S. aureus	ERY	201	20	0	211	10	0	217	4
S. aureus	GEN	136	85	0	211	10	0	217	4
S. aureus	SXT	141	80	0	211	10	0	217	4
S. aureus	TET	125	96	0	211	10	0	217	4
P. aeruginosa	AMK	114	291	0	0	405	0	405	0
P. aeruginosa	CAZ	56	18	50	0	24	0	24	50
P. aeruginosa	LVX	213	192	0	0	405	0	405	0
P. aeruginosa	MEM	201	229	25	0	405	0	405	25
E. coli	AMC	464	1058	1094	1	17	410	17	1505
E. coli	CAZ	135	1397	1094	1	27	410	25	1507
E. coli	CIP	301	1229	1094	0	26	410	24	1506
E. coli	GEN	133	1397	1094	0	26	410	24	1506
E. coli	IPM	23	1096	1094	0	25	0	23	1096
E. coli	TMP	149	263	0	0	2	410	1	411

Table B.3. Number of significant features associated with antimicrobial resistance in *S. aureus*, as detected by Fisher's exact tests and Cochran-Mantel-Haenszel tests.

	Associated with Resistance			**Associated with Susceptibility**		
Drug	Fisher	CMH (Bonf.)	CMH (B-H)	Fisher	CMH (Bonf.)	CMH (B-H)
CIP	58	20	1662	608	31	1263
CLI	1276	0	0	80	0	0
ERY	1436	0	0	90	0	0
GEN	2138	149	1730	1892	2	269
TET	2117	6	6	1881	0	0
SXT	2110	22	37	1876	11	18

Tests were applied between each drug and each genomic feature (n = number of genomes; $n = 216$ for CIP, $n = 221$ for all other drugs). Bonferroni correction was applied to Fisher's exact tests (FWER < 0.05). Results with either a Bonferroni correction (Bonf, FWER < 0.05) or Benjamini-Hochberg correction (B-H, FDR < 0.05) are shown for the Cochran-Mantel-Haenszel (CMH) tests. Multiple hypothesis corrections were done with number of tests = m = 11,485.

Table B.4. Aminoglycoside-modifying enzymes identified by sequence homology in the *P. aeruginosa* pangenome compared to amikacin resistance phenotypes. Log_2 odds ratios (LOR) are shown, using weighted pseudocounts to address zeroes in the contingency table (see Methods).

Gene	Gene Class	Count	Res.	Sus.	LOR
AAC(3)-Id	aminoglycoside acetyltransferase	13	11	2	3.5
AAC(3)-IIc	aminoglycoside acetyltransferase	2	2	0	3.0
AAC(3)-IIIb	aminoglycoside acetyltransferase	3	0	3	-2.4
AAC(6)-33	aminoglycoside acetyltransferase	1	1	0	2.2
AAC(6')-Ib	aminoglycoside acetyltransferase	1	1	0	2.2
AAC(6')-Ip	aminoglycoside acetyltransferase	1	1	0	2.2
aadA2	aminoglycoside adenylyltransferase	36	24	12	2.6
aadA4	aminoglycoside adenylyltransferase	2	1	1	0.9
aadA6/16	aminoglycoside adenylyltransferase	5	3	2	1.7
aadA7	aminoglycoside adenylyltransferase	6	5	1	3.0
ANT(2")-Ia	aminoglycoside nucleotidyltransferase	3	1	2	0.3
APH(3')-Ib	aminoglycoside phosphotransferase	5	2	3	0.7
APH(3')-Ia	aminoglycoside phosphotransferase	1	1	0	2.2
APH(3')-IIb	aminoglycoside phosphotransferase	400	114	286	3.0
APH(3')-VI	aminoglycoside phosphotransferase	2	2	0	3.0
APH(6)-Ic	aminoglycoside phosphotransferase	4	2	2	1.1
APH(6)-Id	aminoglycoside phosphotransferase	34	18	16	1.6

Table B.5. Enrichment for plasmid over chromosomally encoded genetic features selected by SVM-RSE. For each species-drug case, the number of plasmid vs. chromosomally encoded features among the top 50 resistance-associated hits was counted, and the odds ratio for plasmid over chromosomal was computed. Rows without a specified drug show the total plasmid and chromosomal feature count for that species. This analysis was repeated for just non-core genes, where most plasmid features are present.

		All Possible Features			**Non-core Genes**		
Species	Drug	plasmid	chromosomal	odds ratio	plasmid	chromosomal	odds ratio
S. aureus	-	144	23214	-	121	2847	-
S. aureus	CIP	0	50	0	0	20	0
S. aureus	CLI	4	46	14.39	4	18	5.37
S. aureus	ERY	4	46	14.391	4	16	6.05
S. aureus	GEN	2	48	6.80	2	15	3.17
S. aureus	SXT	3	47	10.49	3	23	3.12
S. aureus	TET	7	43	27.53	7	16	10.86
P. aeruginosa	-	459	178046	-	306	22849	-
P. aeruginosa	AMK	0	50	0	0	27	0
P. aeruginosa	CAZ	2	48	16.36	0	17	0
P. aeruginosa	LVX	0	50	0	0	16	0
P. aeruginosa	MEM	0	50	0	0	27	0
E. coli	-	2261	187165	-	1109	36053	-
E. coli	AMC	7	43	13.51	7	37	6.18
E. coli	CAZ	1	49	1.69	1	42	0.77
E. coli	CIP	0	50	0	0	24	0
E. coli	GEN	3	47	5.29	3	45	2.17
E. coli	IPM	2	48	3.45	2	38	1.71
E. coli	TMP	1	49	1.69	1	39	0.83

Table B.6. Comparison of estimates for *S. aureus*, *P. aeruginosa*, and *E. coli* core genome sizes.

Reference	Species	Genomes	Core genes	Method
This study	*S. aureus*	288	2221	genes missing in at most 10 genomes
Fuchs et al. 2018 [9]	*S. aureus*	32	2115	genes missing in at most 1 genome
Bosi et al. 2016 [10]	*S. aureus*	64	1441	Estimated from genome permutations
This study	*P. aeruginosa*	456	4700	genes missing in at most 10 genomes
Subedi et al. 2018 [11]	*P. aeruginosa*	22	4910	genes present in <99% of genomes
Valot et al. 2015 [12]	*P. aeruginosa*	17	5233	genes in all genomes
Ozer et al. 2014 [13]	*P. aeruginosa*	12	5316	genes in all genomes
This study	*E. coli*	1588	3107	genes missing in at most 10 genomes
Kaas et al. 2012 [14]	*E. coli*	186	3051	genes present in <95% of genomes
Lukjancenko et al. 2010 [15]	*E. coli*	53	1472	genes in all genomes (excluding Shigella)
Rasko et al. 2008 [16]	*E. coli*	17	∼2200	Estimated from genome permutations

Previous core genome size estimates are shown, based on varying genome set sizes and definitions of a core gene. Discrepancies between this study and previous studies may be due to either a stricter definition of a core gene (in the case of *S. aureus* or *E. coli*) or a much smaller genome set in the case of *P. aeruginosa*. The method "Estimated from genome permutations" refers to fitting a function (usually exponential) to the size of the core genome vs. number of genomes based on many permutations of genome order. This approach extrapolates the number of core genes present in all genomes for an infinite number of genomes and can be considered the strictest definition of core genome.

Table B.7. Fisher's exact test p-values between each COG functional category and the combined core, accessory, or unique genomes of *S. aureus*, *P. aeruginosa*, and *E. coli*. Tests were applied between each COG and gene category (n = number of genes = 72,220). COGs are ordered by effect size.

COG Functional Category	Core	Acc	Unique
J: Translation, ribosomal structure, biogenesis	<0.00001	<0.00001	<0.00001
H: Coenzyme transport/metabolism	<0.00001	<0.00001	<0.00001
F: Nucleotide transport/metabolism	<0.00001	<0.00001	<0.00001
E: Amino acid transport/metabolism	<0.00001	<0.00001	<0.00001
C: Energy production and conversion	<0.00001	<0.00001	<0.00001
G: Carbohydrate transport/metabolism	<0.00001	0.00047	<0.00001
I: Lipid transport/metabolism	<0.00001	<0.00001	<0.00001
O: PTMs, protein turnover, chaperones	<0.00001	0.00013	<0.00001
P: Inorganic ion transport/metabolism	<0.00001	0.31713	<0.00001
T: Signal transduction mechanisms	<0.00001	<0.00001	<0.00001
Q: Secondary metabolites transport/metabolism	<0.00001	0.00028	<0.00001
M: Cell wall/membrane/envelope biogenesis	<0.00001	0.33189	<0.00001
N: Cell motility	<0.00001	<0.00001	<0.00001
K: Transcription	<0.00001	0.69589	<0.00001
D: Cell cycle/division, chromosome partitioning	<0.00001	0.62140	<0.00001
U: Trafficking, secretion, vesicular transport	0.06502	0.00001	0.0086
A: RNA processing and modification	0.66545	0.65512	0.35414
V: Defense mechanisms	0.67894	0.14832	0.10492
L: Replication, recombination and repair	<0.00001	<0.00001	0.13947
S: Function unknown	<0.00001	<0.00001	<0.00001

Table B.8. Fisher's exact test p-values between each COG functional category and the individual core, accessory, and unique genomes of *S. aureus* (SA), *P. aeruginosa* (PA), and *E. coli* (EC). Tests were applied between each COG and gene category (n = number of genes; n = 5,152 for *S. aureus*, n = 27,435 for *P. aeruginosa*, n = 39,633 for *E. coli*). COGs are ordered by effect size.

	Core Genomes			**Accessory Genomes**			**Unique Genomes**		
COG	SA	PA	EC	SA	PA	EC	SA	PA	EC
J	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
H	<0.00001	<0.00001	<0.00001	<0.00001	0.00002	0.00001	<0.00001	<0.00001	<0.00001
F	<0.00001	<0.00001	<0.00001	0.00046	0.00001	0.00691	<0.00001	<0.00001	<0.00001
E	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.0003	<0.00001	<0.00001	<0.00001
C	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.19327	<0.00001	<0.00001	<0.00001
G	<0.00001	<0.00001	<0.00001	<0.00001	0.53968	<0.00001	<0.00001	<0.00001	<0.00001
I	<0.00001	<0.00001	<0.00001	0.00001	0.00094	0.96853	0.00011	<0.00001	<0.00001
O	<0.00001	<0.00001	<0.00001	0.01354	0.00489	0.5218	<0.00001	<0.00001	<0.00001
P	<0.00001	<0.00001	<0.00001	<0.00001	0.08958	0.4422	<0.00001	<0.00001	<0.00001
T	0.00067	<0.00001	<0.00001	0.86057	<0.00001	0.00105	0.00073	<0.00001	<0.00001
Q	0.01269	<0.00001	<0.00001	0.90916	0.0751	<0.00001	0.00933	<0.00001	<0.00001
M	<0.00001	<0.00001	0.00013	0.04235	0.05049	0.00288	<0.00001	<0.00001	0.47482
N	1	<0.00001	0.00609	0.93916	0.9712	<0.00001	0.87672	<0.00001	<0.00001
K	<0.00001	<0.00001	<0.00001	0.1288	0.29539	0.00004	0.00116	<0.00001	<0.00001
D	0.18291	0.00346	<0.00001	0.07709	0.78917	0.95705	0.90148	0.01029	0.00013
U	0.02353	0.45176	0.15074	0.007	0.01099	<0.00001	0.99719	0.12774	0.00033
A	1	0.86136	0.44756	0.93916	0.48628	0.91384	0.87672	0.27756	0.83678
V	0.17506	0.01038	0.1114	0.82478	0.17517	0.93426	0.25457	0.00164	0.45539
L	0.00003	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.98869	<0.00001	0.00401
S	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001	0.00004	<0.00001	<0.00001	<0.00001

B.5 Supplementary Datasets

Dataset B.1. PATRIC genome IDs for *S. aureus*, *P. aeruginosa*, and *E. coli* genomes used in the development of SVM-RSE for AMR.

Dataset B.2. Protein sequences for known AMR-conferring genes in *S. aureus* annotated for SVM-RSE benchmarking. Contains representative protein sequences of genes known to be associated with resistance against ciprofloxacin, clindamycin, erythromycin, gentamicin, sulfamethoxazole, tetracycline, and trimethoprim. Files named "*drug*_card_amr.faa" contain sequences that were extracted from the CARD database, retrieved November 26, 2018. File "other_amr.faa" contains additional sequences for AMR-conferring genes from literature and UniProt compiled independent of CARD.

Dataset B.3. Protein sequences for the top 50 resistance-associated genetic features identified by SVM-RSE for 16 species-drug cases. Files are named *species*_*drug*_top_hits_seqs.faa, which each contain all protein sequences relevant to the top 50 hits of the corresponding species-drug case. For selected alleles, the exact protein sequence of the allele is included. For selected genes, the protein sequences of all alleles of that gene observed in the species's pangenome are included.

Dataset B.4. Annotations for the top 50 resistance-associated genetic features identified by SVM-RSE for 16 species-drug cases. Includes the following annotation for each genetic feature: 1) ranking from SVM-RSE, 2) the name of the common allele for selected genes, 3) locus tag of the best aligned reference sequence in the corresponding reference genome, if any, 4) gene name of the reference sequence, if available, 5) gene name assigned by eggNOG, if available, and 6) gene functional annotation by eggNOG.

Dataset B.5. Additional figure-associated data the SVM-RSE analysis of AMR.

B.6 References

[1] Alice R Wattam, David Abraham, Oral Dalay, Terry L Disz, Timothy Driscoll, Joseph L Gabbard, Joseph J Gillespie, Roger Gough, Deborah Hix, Ronald Kenyon, Dustin Machi, Chunhong Mao, Eric K Nordberg, Robert Olson, Ross Overbeek, Gordon D Pusch, Maulik Shukla, Julie Schulman, Rick L Stevens, Daniel E Sullivan, Veronika Vonstein, Andrew Warren, Rebecca Will, Meredith J C Wilson, Hyun Seung Yoo, Chengdong Zhang, Yan Zhang, and Bruno W Sobral. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, 42(Database issue):D581–91, January 2014.

[2] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[3] Baofeng Jia, Amogelang R Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K Tsang, Briony A Lago, Biren M Dave, Sheldon Pereira, Arjun N Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E Williams, Jonathan G Frye, Tariq Elsayegh, Daim Sardar, Erin L Westman, Andrew C Pawlowski, Timothy A Johnson, Fiona S L Brinkman, Gerard D Wright, and Andrew G McArthur. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 45(D1):D566–D573, January 2017.

[4] Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.*, 47(D1):D195–D202, January 2019.

[5] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, June 2016.

[6] Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, May 2014.

[7] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11(5):472–477, October 2008.

[8] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.*, 34(8):2115–2122, August 2017.

[9] Stephan Fuchs, Henry Mehlan, Jörg Bernhardt, André Hennig, Stephan Michalik, Kristin Surmann, Jan Pané-Farré, Anne Giese, Stefan Weiss, Linus Backert, Alexander Herbig, Kay Nieselt, Michael Hecker, Uwe Völker, and Ulrike Mäder. AureoWiki the repository of the staphylococcus aureus research and annotation community. *Int. J. Med. Microbiol.*, 308(6):558–568, August 2018.

[10] Emanuele Bosi, Jonathan M Monk, Ramy K Aziz, Marco Fondi, Victor Nizet, and Bernhard Ø Palsson. Comparative genome-scale modelling of staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. *Proc. Natl. Acad. Sci. U. S. A.*, 113(26):E3801–9, June 2016.

[11] Dinesh Subedi, Ajay Kumar Vijay, Gurjeet Singh Kohli, Scott A Rice, and Mark Willcox. Comparative genomics of clinical strains of pseudomonas aeruginosa strains isolated from different geographic sites. *Sci. Rep.*, 8(1):15668, October 2018.

[12] Benoît Valot, Christophe Guyeux, Julien Yves Rolland, Kamel Mazouzi, Xavier Bertrand, and Didier Hocquet. What it takes to be a pseudomonas aeruginosa? the

core genome of the opportunistic pathogen updated. *PLoS One*, 10(5):e0126468, May 2015.

[13] Egon A Ozer, Jonathan P Allen, and Alan R Hauser. Characterization of the core and accessory genomes of pseudomonas aeruginosa using bioinformatic tools spine and AGEnt. *BMC Genomics*, 15:737, August 2014.

[14] Rolf S Kaas, Carsten Friis, David W Ussery, and Frank M Aarestrup. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse escherichia coli genomes. *BMC Genomics*, 13:577, October 2012.

[15] Oksana Lukjancenko, Trudy M Wassenaar, and David W Ussery. Comparison of 61 sequenced escherichia coli genomes. *Microb. Ecol.*, 60(4):708–720, November 2010.

[16] David A Rasko, M J Rosovitz, Garry S A Myers, Emmanuel F Mongodin, W Florian Fricke, Pawel Gajer, Jonathan Crabtree, Mohammed Sebaihia, Nicholas R Thomson, Roy Chaudhuri, Ian R Henderson, Vanessa Sperandio, and Jacques Ravel. The pangenome structure of escherichia coli: comparative genomic analysis of e. coli commensal and pathogenic isolates. *J. Bacteriol.*, 190(20):6881–6893, October 2008.

Appendix C

Global pathogenomic analysis for antimicrobial resistance in twelve species - Supplementary Information

C.1 Methods

C.1.1 Genome selection

An initial set of genomes was taken from the PATRIC database RELEASE_NOTES (ftp.patricbrc.org/RELEASE_NOTES/, 2021-07-21) and filtered down to 12 species based on taxon ID. For each species, genomes were filtered to those meeting the four criteria previously described [1], as well as having CheckM fine consistency of at least 87% [2]. Drugs with at least 100 experimental antimicrobial (AMR) measurements were identified per species, and genomes were filtered to those with data for at least one of those drugs. Selected PATRIC Genome IDs are available in Dataset C.1.

C.1.2 Pangenome construction and genetic feature identification

For each species, genes, protein sequence variants ("alleles"), and ORF-flanking sequence variants were identified using the approach previously described [1]. Briefly, protein sequences were clustered using CD-HIT v4.6 with minimum identity 80% and minimum alignment length 80% [3]. Each cluster was treated as a gene, cluster members

as alleles, and the 300bp upstream of the start codon of each occurrence of the gene as 5′ variants (and analogously the 300bp downstream of stop codons as 3′ variants). A similar clustering analysis was conducted for non-coding features (annotated by PATRIC as "transcript", "tRNA", "rRNA", or "misc_binding") to identify non-coding feature clusters and nucleotide sequence variants, using CD-HIT-EST v4.6 with the same parameters [3]. The species-wide genetic variation covered by these six feature types was represented as a binary matrix based on presence/absence calls of each feature for each genome.

C.1.3 Processing antimicrobial resistance phenotypes and SIR phenotype inference from MICs

Species-drug cases with at least 100 experimental susceptible-intermediate-resistant (SIR) phenotypes or minimum inhibitory concentration (MIC) measurements among selected genomes were identified. For each case, the most common SIR testing standard (i.e. CLSI, EUCAST) was identified from either the metadata or manual curation of BioProject accession IDs (Dataset C.2), referred to as the species-drug case's primary standard.

MIC values were mapped to SIRs by first filtering MICs for exact mg/L values (opposed to bounded MICs) derived from one of the following laboratory typing methods: agar_dilution, agar_dilution_or_etest, bd_phoenix, bd_phoenix_and_etest, broth_microdilution, etest, mic, mic broth microdilution, liofilchem, sensititre, vitek_2. For each species-drug case, MIC-SIR mappings were generated for all MIC-SIR value pairs reported in at least three genomes under the primary standard. Ambiguous mappings (MIC value mapped to multiple SIRs) and inconsistent sets of mappings (instances where a susceptible MIC is greater than an intermediate or resistant MIC, or where a resistant MIC is less than an intermediate or susceptible MIC) were removed. A MIC-to-SIR inference scheme was developed as follows:

- Exact MICs: Mapped directly to the corresponding SIR if possible.

- Upper bounded and unmapped exact MICs: If MIC $\leq$ largest MIC mapped to susceptible, it is mapped to susceptible.
- Lower bounded and unmapped exact MICs: If MIC $\geq$ smallest MIC mapped to resistant, it is mapped to resistant.
- Other unmapped exact MICs: If the MIC value is within the range of MICs mapped to intermediate, it is mapped to intermediate.

MICs used and MIC-SIR mappings are available in Dataset C.2. Reported and inferred SIRs were combined for subsequent analyses. For genomes with multiple conflicting SIRs for a single drug (i.e. measured by different methods), the most common SIR across all methods and inferences was selected, with directly reported SIRs breaking ties and perfect ties ignored. The final set of SIRs are available in Dataset C.2. SIRs were binarized into susceptible and non-susceptible by converting "susceptible", "susceptible-dose dependent", and "non-resistant" to 0s, and "resistant", "intermediate", "non-susceptible" and "IS" to 1s for subsequent analyses.

C.1.4 Identification and classification of known AMR genes

All protein sequences for each species were annotated using RGI v5.2.0 with CARD ontology v3.1.3 [4]. To link AMR genes identified by RGI to specific drugs, a directed graph was constructed from the CARD ontology using ARO accession IDs as nodes and adding directed edge "U$\rightarrow$V" whenever:

- U corresponds to a gene and U has the relationship "is_a" to V.
- U corresponds to a drug and V has the relationship "is_a" to U.
- U has the relationship "part_of", "regulates", "confers_resistance_to_antibiotic", or "confers_resistance_to_drug_class" to V.

- V has the relationship "has_part" to U.

A gene was labeled as conferring resistance to a drug if there exists a path from the gene's node to the drug's node in this graph. 23S rRNAs, 16S rRNAs, and 50S rRNAs were manually identified from PATRIC text annotations of noncoding features and were similarly linked to specific drugs using the graph, starting from nodes ARO:3000336, ARO:3003211, and ARO:3005003, respectively.

Additional drug-specific AMR genes were identified from PATRIC text annotations. Sequences with annotations containing a drug name or identical to the annotation of an RGI-identified AMR gene were identified and manually curated for probable known AMR genes (curated annotations are available in Dataset C.3). For machine learning purposes, all features associated with a gene cluster containing a sequence linked to resistance for a drug by RGI or PATRIC text annotation were treated as known AMR features.

C.1.5 AMR gene cross-species comparison, location prediction, and TEM beta-lactamase analysis

All alleles of identified AMR genes across all species were combined, de-duplicated, and re-clustered with the same CD-HIT parameters for cross-species analysis. Gene-level functional annotations for re-clustered AMR genes were inherited from corresponding allele-level annotations from RGI and PATRIC. AMR gene functional categories were assigned based on RGI annotations when available and PATRIC annotations otherwise, with categories occurring less than 50 times grouped as "other" (Dataset C.3).

All contigs from all assemblies were labeled as plasmid or chromosomal based on *de novo* predictions from PlasFlow v1.1 on default settings [5]. Contigs were also mapped to known plasmids in PLSDB version 2021_06_2327 using MASH v2.3 with minimum shared kmers 500/1000 [6]. For a given AMR gene cluster, each instance of each allele was assigned a location based on the PlasFlow prediction (chromosome, plasmid, unassigned) for the contig containing that instance. The overall location of a gene or allele was assigned

as 1) "chromosome" if >90% of instances were chromosomal, 2) "plasmid" if >90% of instances were plasmid, 3) "chromosome-leaning" if >50% of instances were chromosomal, 4) "plasmid-leaning" if >50% of instances were plasmid, and 5) "ambiguous" otherwise (Dataset C.3).

Complete TEM-family beta-lactamases (blaTEMs) were identified by filtering all AMR alleles for mention of "TEM" in the RGI or PATRIC annotation, and for length at least 272aa (95% length of TEM-1). Mutations were called from pairwise global alignment of each allele to TEM-1 using the Biopython Align module [7] with scores match=1, mismatch=-3, open_gap_score=-5, and extend_gap_score=-2. N/C-terminal deletions were ignored. Known TEM variants were identified based on exact matches in the CARD database. For the *S. aureus* analysis, cefoxitin was identified as the only drug for which MIC data was available among blaTEM-carrying *S. aureus* genomes. Known AMR genes among *S. aureus* genomes with cefoxitin MIC data were identified from exact matches to entries related to beta-lactams in the CARD database.

C.1.6 Implementation, evaluation, and hyperparameter optimization of SVM ensembles

For each species-drug case, SVM ensembles were trained to classify genomes as susceptible or non-susceptible (intermediate or resistant, referred to as "resistant") based on the species' genetic feature presence/absence matrix. To accelerate training, feature count was reduced in three stages: 1) features present or missing in less than 3 genomes were removed, 2) perfectly correlated features were merged, and 3) remaining features were sorted by log odds ratio (LOR) for resistant genomes, and features with the 25,000 highest and 25,000 lowest LORs were retained. SVM ensembles were implemented using scikit-learn classes LinearSVC and BaggingClassifier [8], with square hinge loss weighted by class frequency to address class imbalance issues and L1 regularization to enforce sparsity in feature selection.

Model performance was evaluated in 5-fold cross validation experiments. Phenotype prediction accuracy was scored as the mean Matthews correlation coefficient (MCC) on the test set. Biological relevance was scored using the following equation:

$$\text{GWAS Score} = \sum_{r \in \text{known}} 0.5^{(r-1)/10}$$

where r corresponds to the ranks of known AMR features associated with the specific drug when sorted by feature importance, with $r = 1$ corresponding to the highest feature importance. Feature importance was computed as the absolute value of the mean of the feature's coefficients across all SVMs in the ensemble with access to the feature (i.e. selected during feature subsampling). For ties, the average rank was assigned to all tied features.

Hyperparameter (HP) ranges for SVM ensembles were first evaluated on 10 representative species-drug cases (Table C.4), starting with 256 HP combinations from four HPs: number of estimators per ensemble, fraction of samples per estimator, fraction of features per estimator, and the SVM regularization term C. For each representative case, the highest MCC and GWAS scores were computed across all HP combinations. The smallest subset of HP combinations was identified such that the best scores in the subset were within 90% of the best scores across all combinations, across all representative cases. The initial and reduced HP ranges are available in Table C.6.

SVM ensembles were trained for each HP combination in the reduced set across 127 species-drug cases with at least 100 SIRs, 10 known AMR genes, and minority phenotype frequency >5%. For each case, the optimal HP set was selected by ranking all HP sets by either MCC or GWAS score, taking the average of the two ranks as the HP set's overall rank, and selecting the HP set with the highest overall rank. Comparison with Fisher's exact test was conducted by applying Fisher's exact test between each genetic feature and AMR phenotype, ranking features by p-value, and comparing the top 10, 20, or 50

features from either method.

C.1.7 Identification of candidate antimicrobial resistance determinants

The models for each of the 127 species-drug cases were filtered down to those achieving mean test MCC>80% during 5-fold cross validation. From each remaining model, the top 10 features by feature weight absolute value were identified, and filtered for those that 1) were not already known AMR genes, 2) occurred in at least 10 genomes with SIR data for the corresponding drug, and 3) had positive feature weight and LOR for resistance. Remaining feature-drug pairs were tested for whether the feature was significantly associated with resistance, applying Fisher's exact test for SIRs and Brunner-Munzel tests for MICs. Tests were applied for the specific drug and drugs of the same class for which at least 5 SIRs (for Fisher's exact) or 5 MICs (for Brunner-Munzel) were available, and significance was determined at FWER < 0.05 with Bonferroni correction (2,008 Fisher's exact tests and 1,393 Brunner-Munzel tests were conducted, with significance thresholds of $p < 2.5 * 10^{-5}$ and $p < 3.6 * 10^{-5}$, respectively). To evaluate co-occurrence with known AMR genes, "strong effect" AMR features were identified for each drug, defined as those occurring in at least 5 genomes of which at least 90% are resistant. Each candidate feature-drug pair was assigned a score based on the sum of 1) number of drugs with significant association based on SIRs, 2) number of drugs with significant association based on MICs, and 3) number of drugs for which at least one resistant genome with the feature does not also have any strong effect AMR features. Candidate scores are available in Dataset C.7.

The top 10 features by score for each species-drug class pair were identified, yielding 142 candidates which were categorized by function as annotated by PATRIC and additionally by eggNOG-emapper v2.1.6-43 [9]. Genes that were poorly annotated, related to mobile genetic elements (transposases, insertion elements, phage elements, integrases,

plasmid maintenance), or known to be associated with a specific AMR mechanism for an unrelated drug class were their own categories, and the remaining genes were categorized as well-characterized candidates. For perfectly correlated features, coding features were selected over noncoding features. For variant-level features, mutations were determined against the most common variant of the parent gene cluster using the Biopython Align module [7].

C.1.8 Generation of *frdD* and *cycA E. coli* mutants

E. coli BW25113 knockout mutants $\Delta cycA$ and $\Delta ampC$ were taken from the Keio collection [10]. The *frdD* mutations referred to in this study were introduced into both BW25113 and $\Delta ampC$ using a Cas9-assisted Lambda Red homologous recombination method. Golden gate assembly was first used to construct a plasmid vector harboring both Cas9 and lambda red recombinase genes under the control of an L-arabinose inducible promoter, a single guide RNA sequence, and a donor fragment generated by PCR which contained the desired mutation and around 200bp flanking both sides of the Cas9 target cut site as directed by the guide RNA. After allowing the transformed cells to recover for 2 hours at 30°C, L-arabinose was added to the media and the cells were allowed to grow for 3-5 hours at which time a portion of the culture was plated. Single colonies were screened using ARMS PCR and amplicons spanning the mutation site, generated with primers annealing to the genome upstream and downstream of the sequence of the donor fragment contained in the plasmid, were confirmed with Sanger sequencing. Confirmed isolates were cured of the plasmid by growth at 37°C. Both of the *frdD* mutations that were introduced in this study fell within a guide RNA target sequence. Because Cas9 has a tolerance for some single base mismatches in the guide RNA, a second mismatch was engineered into the guide RNA so that the guide RNA had two mismatches with respect to the successfully mutated target sequence and only one tolerated mismatch with respect to the starting strain. In one case, an intermediary strain was first constructed in

which all of the codons falling within the guide RNA target sequence were switched to synonymous ones maximizing the base changes. A second round of Cas9-assisted Lambda Red homologous recombination was then used to restore those codons to their original sequences and introduce the desired mutation at the same time.

C.1.9 Cell growth conditions and measurements

Two media were used for both *cycA* and *frdD* experiments: 1) Mueller-Hinton Broth (Sigma-Aldrich, SKU: 70192-500G) supplemented with 49mM $MgCl_2$ and 69mM $CaCl_2$, and 2) M9 minimal medium (47.8mM Na_2HPO_4, 22mM KH_2PO_4, 8.6mM NaCl, 18.7mM NH_4Cl, 2mM $MgSO_4$, 0.1mM $CaCl_2$) supplemented with 2g/L glucose. For *cycA* experiments, media were also supplemented with either 10mM glycine, D-serine, D-alanine, L-alanine, DL-alanine (50:50 mixture of D- and L-alanine) or nothing, for a total of 12 possible supplemented media.

Cell densities for strain-media-antibiotic combinations were measured in biological triplicates as follows: Fresh culture samples were prepared (OD600 = $\sim$0.05) in each relevant media. Sample solutions were loaded to Costar flat-bottom 96-well plates (Corning, catalog no. 3370), with antibiotics added to varying concentrations (8, 16, 31, 62, or 125 μg/L ciprofloxacin for *cycA* experiments, and 0.25, 0.5, 1, 2, 4, or 8 mg/L ampicillin for *frdD* experiments). Plates were incubated in a microplate reader (Tecan Infinite200 PRO) with shaking at 37°C, and ODs were read every 15 minutes. Maximum cell density for each condition and replicate was calculated as the maximum OD600 over 12 hours after inoculation minus the minimum OD600 observed for the corresponding media without inoculum. Significant differences in cell density between pairs of strains or conditions were determined with Welch t-tests (FDR < 0.05, Benjamini-Hochberg correction).

C.2 Supplementary Figures

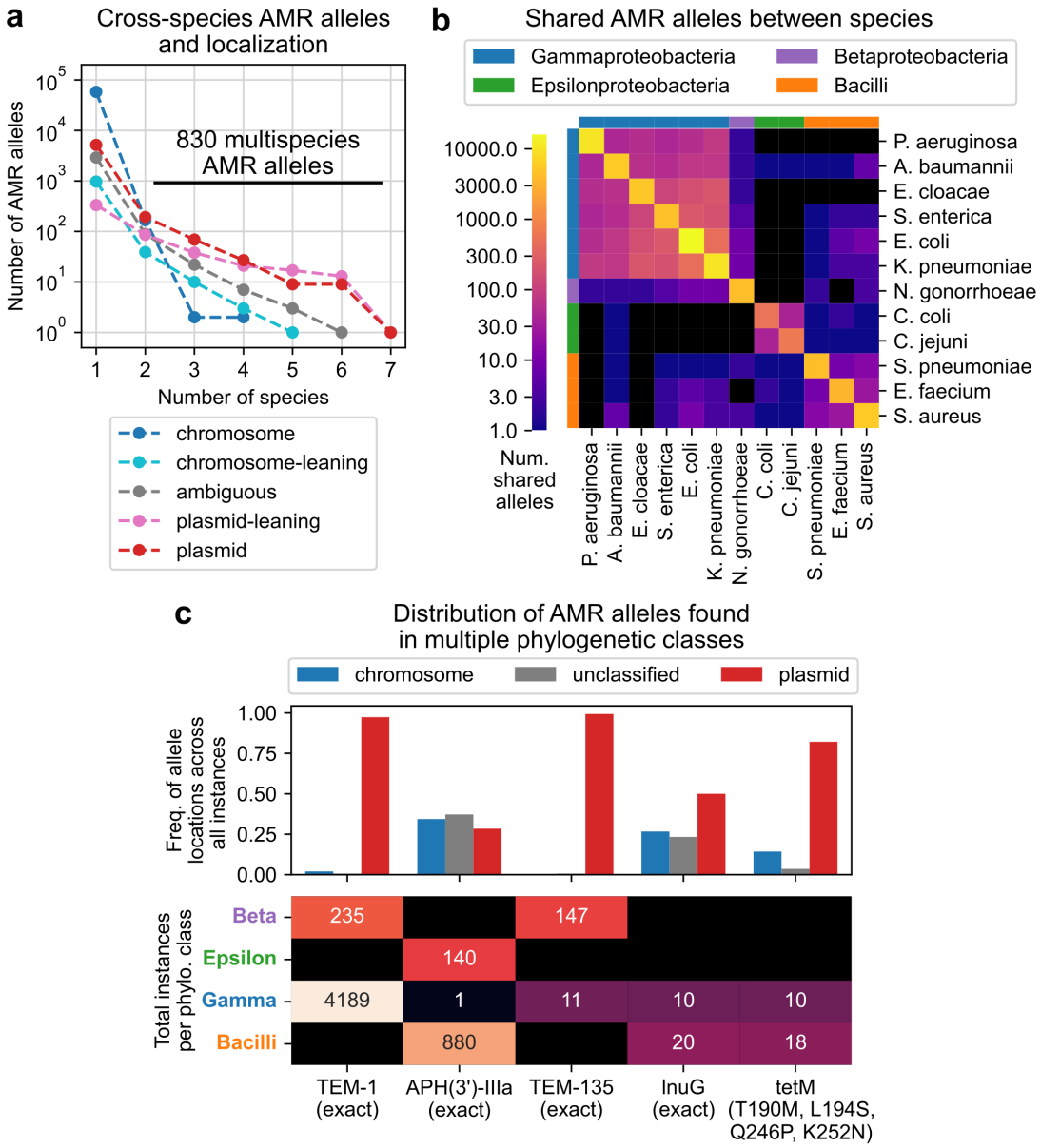


Figure C.1. Cross-species analysis of 68,324 antimicrobial resistance gene alleles and gene locations. (a) Relationship between the number of species an AMR allele is observed in and tendency to be plasmid-encoded. (b) Number of AMR alleles shared between each pair of species, compared to species phylogenetic class. (c) Distribution of predicted gene locations and total occurrences per phylogenetic class for AMR alleles appearing at least 10 genomes in multiple classes.

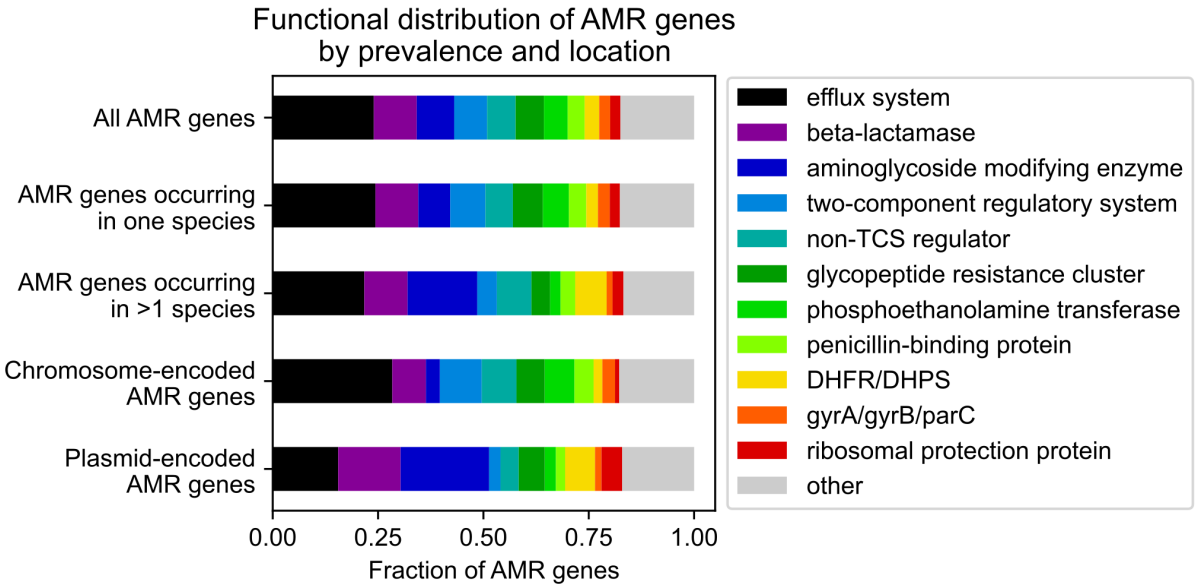


Figure C.2. Distribution of 6,332 antimicrobial resistance (AMR) genes across 12 species by functional category, cross-species prevalence, and gene location. Abbreviated functions are two-component regulatory system (TCS), dihydrofolate reductase (DHFR), and dihydropteroate synthase (DHPS).

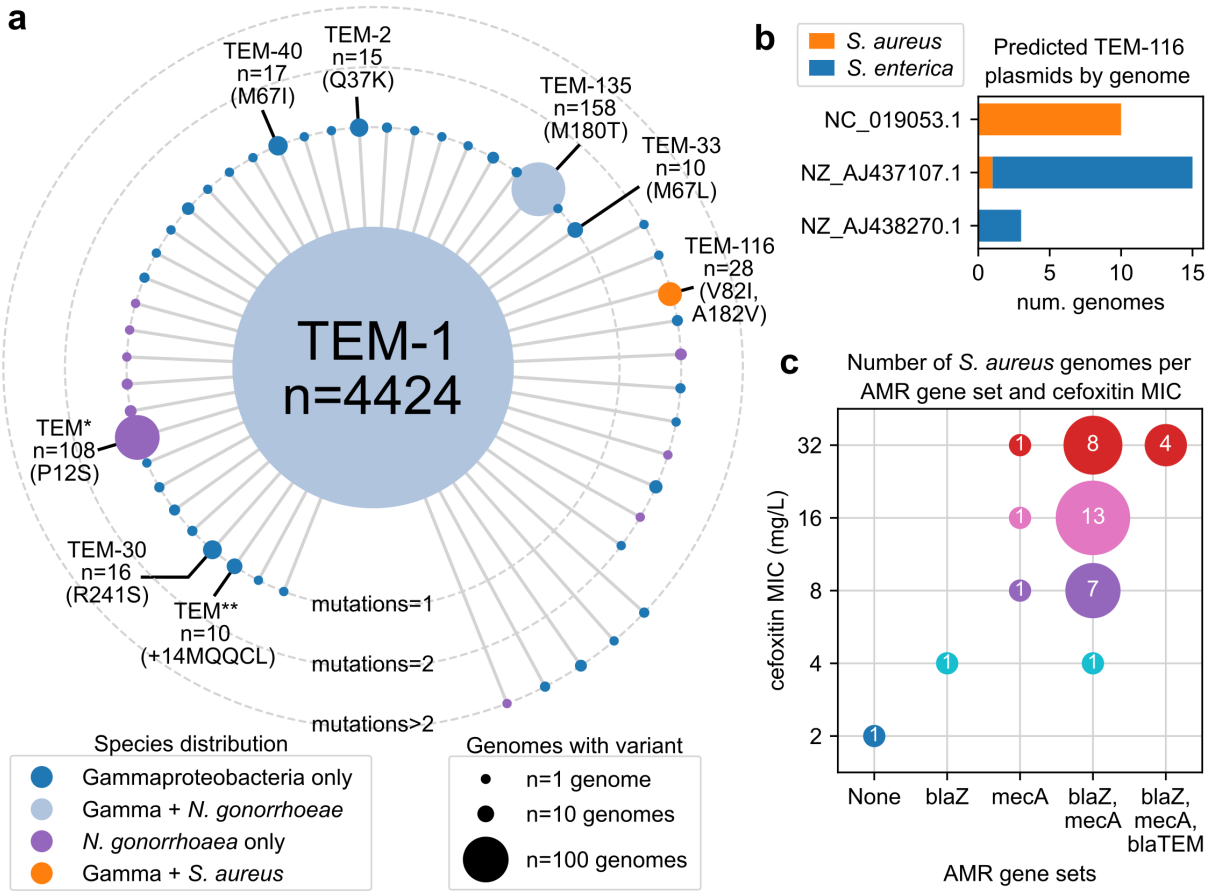


Figure C.3. Distribution of TEM-family beta-lactamases (blaTEMs) observed in 4,861 genomes across 8 species. (a) Distribution of blaTEM alleles with respect to genome count, harboring species, and mutations relative to the TEM-1 allele. Alleles occurring in at least 10 genomes are labeled. TEM* and TEM** refer to unnamed blaTEM alleles. (b) Distribution of plasmids containing the TEM-116 allele observed in both gram-positive and gram-negative genomes. (c) Relationship between beta-lactam associated AMR genes and cefoxitin minimum inhibitory concentration (MIC) in *S. aureus*. Numeric labels correspond to the number of genomes observed with the corresponding AMR genes and MIC.

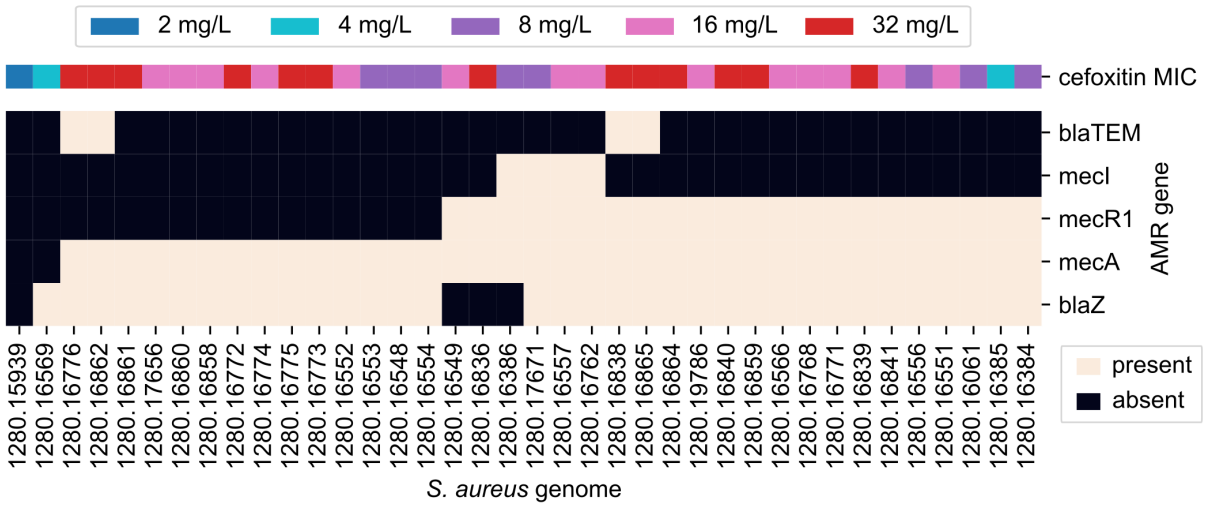


Figure C.4. Cefoxitin MIC versus beta-lactam resistance genes in *S. aureus*. Rows and columns have been ordered by hierarchical clustering with Jaccard distances and average linkage.

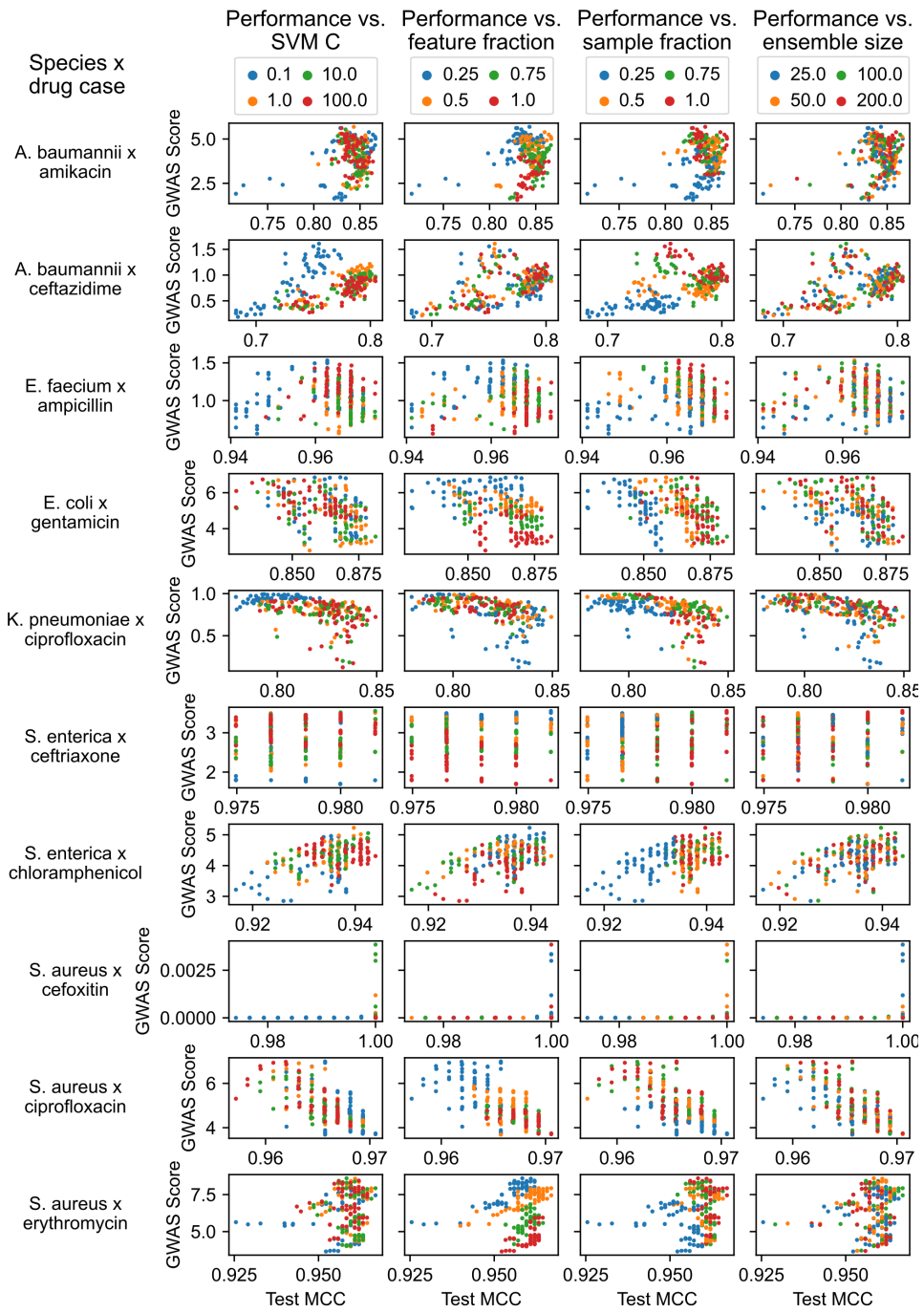


Figure C.5. Impact of SVM ensemble hyperparameters on AMR phenotype prediction performance and recovery of known AMR genes. Each row corresponds to a single AMR prediction problem between a species and a drug, and each column corresponds to a varied hyperparameter. X-axes show average MCCs on the test set from 5-fold cross validation (CV) experiments, y-axes show average GWAS scores across the CV experiment.

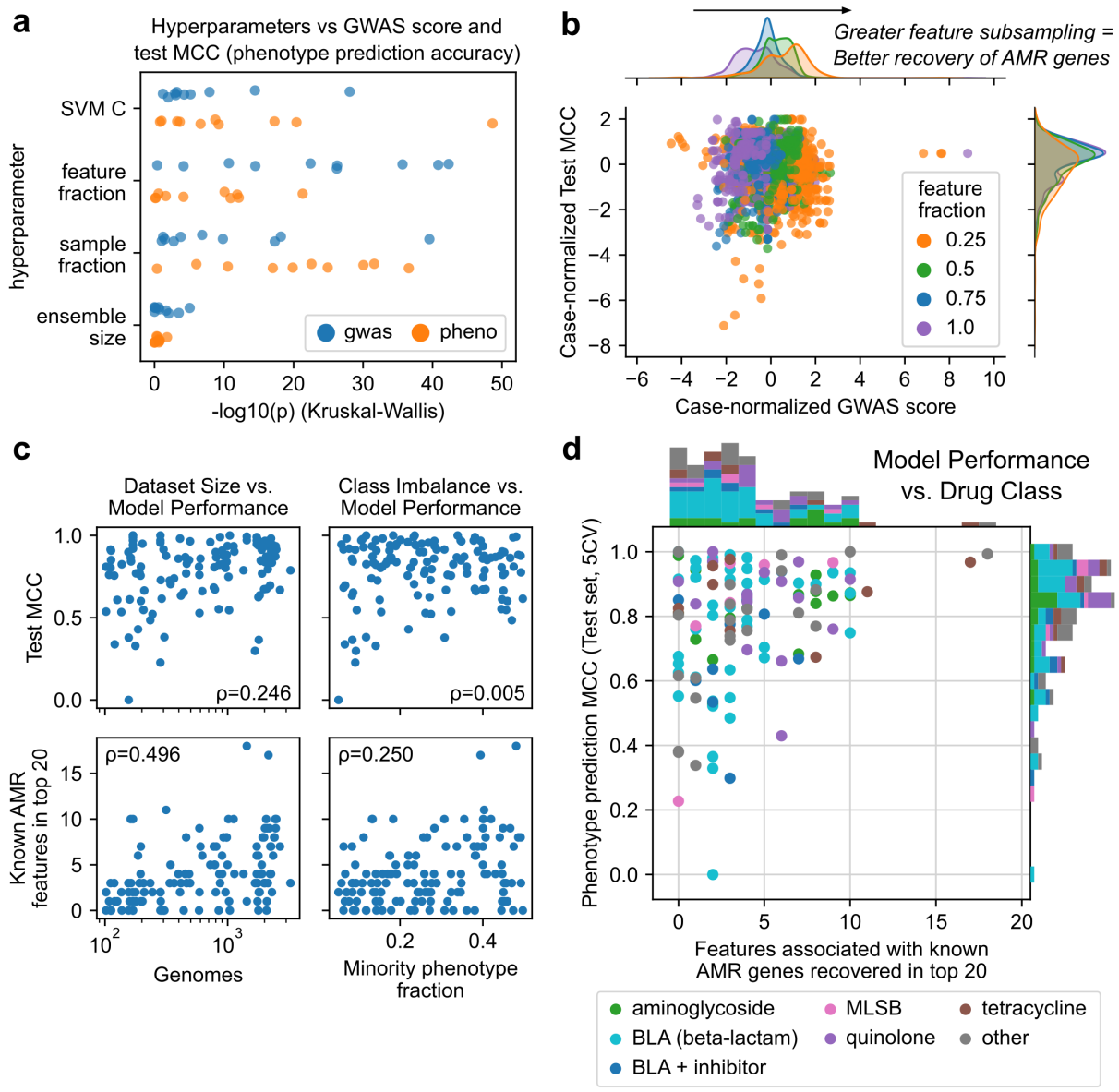


Figure C.6. Overall impact of SVM ensemble hyperparameters on performance and global performance of hyperparameter-optimized ensembles. (a) Kruskal-Wallis ANOVA tests between hyperparameters and either AMR phenotype prediction performance (Matthews correlation coefficient, MCC) or biological relevance (GWAS score) across 10 representative species-drug cases. (b) Impact of feature subsampling on model performance for representative cases. MCCs and GWAS scores have been normalized to mean 0 and standard deviation 1, within their specific species-drug cases. (c) Performance of hyperparameter-optimized SVM ensembles on 127 cases versus dataset size and extent of class imbalance. Spearman correlation coefficients are shown. (d) Performance of the 127 SVM ensembles versus drug class.

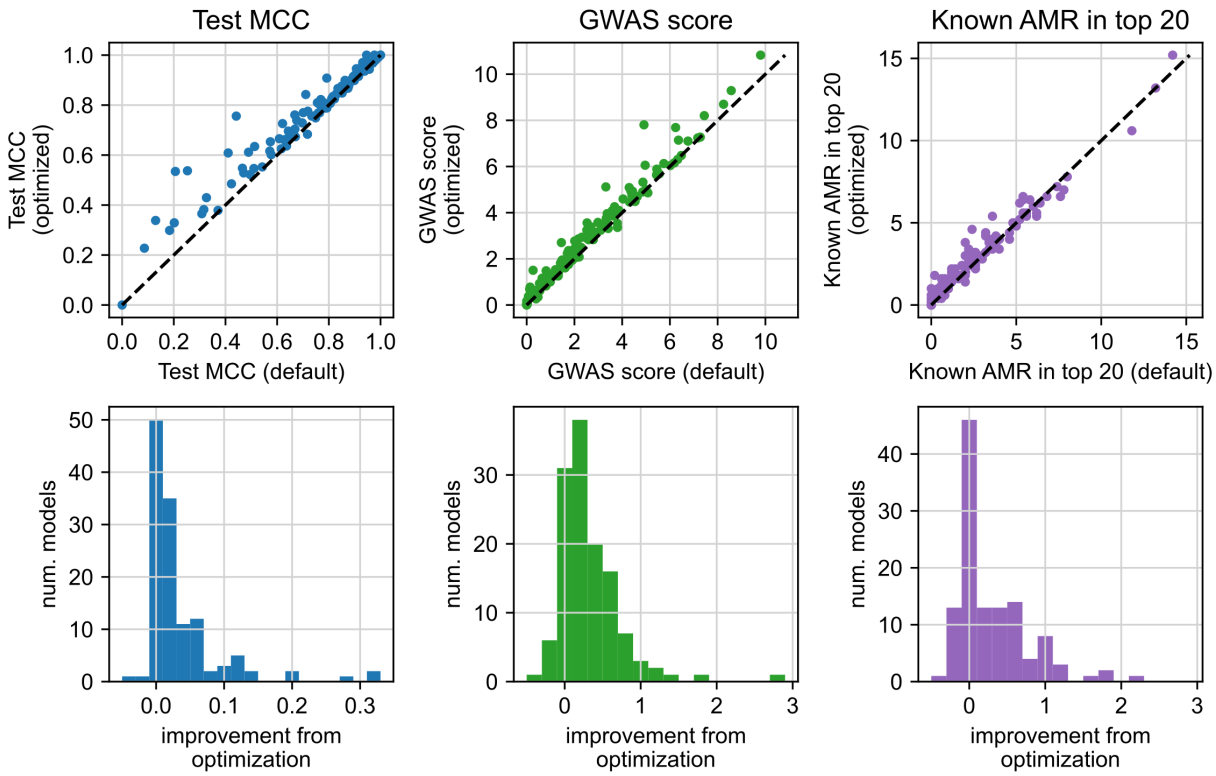


Figure C.7. Performance improvements from hyperparameter optimization of SVM ensembles. Performance by three metrics is shown for SVM ensembles trained using "default" fixed hyperparameters (C = 1, feature fraction = 50%, sample fraction = 75%, ensemble size = 50) compared to ensembles with hyperparameters optimized to balance both phenotype prediction accuracy and known AMR gene recovery. Performance is based on means from 5-fold cross validation.

C.3 Supplementary Tables

Table C.1. Genetic feature counts across 12 species selected for AMR analysis.

		ORF-associated				**Non-ORF-associated**	
species	genomes	genes	alleles	5' flanking variants	3' flanking variants	noncoding clusters	noncoding variants
A. baumannii	1327	43888	239729	262420	275216	177	1725
C. coli	283	8299	44804	32011	31527	77	231
C. jejuni	452	9801	48710	35975	36276	82	381
E. cloacae	484	51969	325374	394577	406576	167	1716
E. faecium	1432	53518	174176	116209	113574	119	890
E. coli	3856	148886	1024333	1005186	1003017	435	5562
K. pneumoniae	3022	94147	562721	584114	596130	356	5010
N. gonorrhoeae	5962	18225	199856	122707	117320	74	693
P. aeruginosa	1050	67217	448882	558844	581936	136	903
S. enterica	3302	56315	262555	255349	261757	322	4932
S. aureus	2248	22649	210749	177169	173244	135	1826
S. pneumoniae	3737	36070	295716	269652	281790	110	1669

Table C.2. Enrichment of AMR gene categories in multispecies over single species genes and plasmid over chromosomal genes. For each comparison, "genes" is the total number of genes in the category, "LOR" is $\log_2$ odds ratio, and "p-value" is Fisher's exact test p-value. Starred values are significant at FWER < 0.05 (Bonferroni correction, 36 tests). Categories are sorted by LOR for plasmid over chromosomal genes.

	Multispecies over Single species			**Plasmid over Chromosomal**		
AMR gene category	genes	LOR	p-value	genes	LOR	p-value
aminoglycoside modifying enzymes (AMEs)	563	1.272	<0.00001*	455	2.957	<0.00001*
ribosomal protection proteins	154	0.113	0.72879	115	2.348	<0.00001*
chloramphenicol acetyltransferases (CATs)	91	1.001	0.00656	77	1.850	<0.00001*
dihydrofolate reductases (DHFRs), dihydropteroate synthases (DHPSs)	219	1.498	<0.00001*	191	1.845	<0.00001*
rpoB variants	74	-0.718	0.24769	60	1.540	0.00007*
rRNA methyltransferases (rRNA MTases)	152	0.061	0.81658	132	1.049	0.00008*
beta-lactamases	645	0.015	0.90646	541	1.004	<0.00001*
glycopeptide resistance clusters (RCs)	425	-0.721	0.00217	357	-0.123	0.50527
gyrA/gyrB/parC variants	165	-1.021	0.01013	141	-0.979	0.00174
non-two-component system (non-TCS) regulators	429	0.358	0.06540	388	-0.993	<0.00001*
penicillin-binding proteins (PBPs)	254	-0.257	0.41414	211	-1.088	0.00002*
aminoacyl-tRNA synthetases (AARSs)	101	-1.454	0.01001	82	-1.094	0.00939
efflux	1519	-0.215	0.08753	1346	-1.107	<0.00001*
phosphoethanolamine (pEtN) transferases	358	-1.373	<0.00001*	316	-1.485	<0.00001*
other	410	-0.218	0.34748	362	-1.811	<0.00001*
two-component system (TCS) regulators	494	-0.900	0.00005*	429	-1.988	<0.00001*
cya variants	57	-0.544	0.45547	50	-3.273	0.00002*
porins	74	-1.595	0.01988	69	-3.762	<0.00001*

Table C.3. Distribution of frequently observed TEM-family beta-lactamases by species. Mutations are shown relative to the TEM-1 variant. Variants are ordered by total count. Species are abbreviated as *A. baumannii* (AcB), *E. cloacae* (EnC), *E. coli* (EsC), *K. pneumoniae* (KlP), *N. gonorrhoeae* (NeG), *P. aeruginosa* (PsA), *S. enterica* (SaE), and *S. aureus* (StA).

		Number of genomes with variant by species								
Mutations	Variant	AcB	EnC	EsC	KlP	NeG	PsA	SaE	StA	Total
-	TEM-1	414	171	1692	1179	235	1	732	-	4424
M180T	TEM-135	-	-	8	1	147	-	2	-	158
P12S	*	-	-	-	-	108	-	-	-	108
V82I, A182V	TEM-116	-	-	-	-	-	-	17	11	28
M67I	TEM-40	-	-	16	1	-	-	-	-	17
R241S	TEM-30	-	-	16	-	-	-	-	-	16
Q37K	TEM-2	-	14	-	1	-	-	-	-	15
M67L	TEM-33	-	-	7	1	-	-	2	-	10
14insMQQCL	*	-	-	-	-	-	-	10	-	10

*No exact match in the CARD database

Table C.4. Representative species-drug cases for SVM hyperparameter testing. Columns "n", "% Sus." and "AMR genes" refer to the number of genomes with AMR data, the fraction of those genomes that are susceptible, and the number of known AMR genes identified that are associated with the drug for that species, respectively.

Species	Species Class	Drug	Drug Class	n	% Sus.	AMR genes
A. baumannii	γ-proteobacteria	amikacin	aminoglycoside	924	49.1	150
A. baumannii	γ-proteobacteria	ceftazidime	beta-lactam	960	15.3	151
E. coli	γ-proteobacteria	gentamicin	aminoglycoside	2447	86.6	263
K. pneumoniae	γ-proteobacteria	ciprofloxacin	quinolone	2089	19.1	227
S. enterica	γ-proteobacteria	ceftriaxone	beta-lactam	2259	84.2	154
S. enterica	γ-proteobacteria	chloramphenicol	other	2348	82.4	116
E. faecium	Bacilli	ampicillin	beta-lactam	1432	14.2	81
S. aureus	Bacilli	cefoxitin	beta-lactam	1041	23.5	68
S. aureus	Bacilli	ciprofloxacin	quinolone	1725	62.2	42
S. aureus	Bacilli	erythromycin	other	2014	71.7	44

Table C.5. Negative $\log_{10}$ p-values for Kruskal-Wallis tests between SVM ensemble hyperparameters and model performance. Performance metrics compared are phenotype prediction performance (mean test set MCC during 5-fold cross validation) and recovery of known AMR genes among model features (GWAS score).

	Hyperparameter vs. Test set MCC				**Hyperparameter vs. GWAS score**			
AMR Case	SVM C	feature fraction	sample fraction	ensemble size	SVM C	feature fraction	sample fraction	ensemble size
A. baumannii, amikacin	1.009	12.123	6.016	0.75	1.967	26.219	17.297	0.711
A. baumannii, ceftazidime	17.258	0.187	24.916	0.58	3.212	0.402	39.543	0.011
E. coli, gentamicin	3.256	10.935	31.611	0.051	2.963	35.68	1.47	5.058
E. faecium, ampicillin	8.738	11.894	17.011	0.307	7.905	22.447	9.831	1.674
K. pneumoniae, ciprofloxacin	20.421	1.621	19.877	0.051	28.08	4.177	2.767	3.495
S. aureus, cefoxitin	48.629	0.117	0.353	0.001	14.449	10.696	1.244	0.001
S. aureus, ciprofloxacin	9.228	21.314	10.536	0.416	1.211	26.275	18.174	0.059
S. aureus, erythromycin	3.635	10.07	22.577	0.723	3.065	40.817	6.849	0.158
S. enterica, ceftriaxone	0.791	4.16	36.582	0.618	4.264	42.314	1.036	0.59
S. enterica, chloramphenicol	6.62	0.579	30.051	1.804	5.181	14.501	3.774	2.055

Table C.6. SVM ensemble hyperparameter ranges used during optimization. Larger initial ranges were used for evaluating representative species-drug cases, from which reduced ranges were derived for evaluating against all species-drug cases.

Hyperparameter	Initial range	Reduced range
Number of estimators	25, 50, 100, 200	25, 50
Fraction of samples per estimator	25%, 50%, 75%, 100%	50%, 75%, 100%
Fraction of features per estimator	25%, 50%, 75%, 100%	25%, 50%, 75%, 100%
C (SVM regularization term)	0.1, 1, 10, 100	0.1, 1, 10
Unique combinations	**256**	**72**

Table C.7. Impact of input data properties on SVM ensemble performance. Impact was assessed with Kruskal-Wallis tests for categorical properties (species, drug class) and Spearman correlation for quantitative properties (number of genomes, minority phenotype fraction). Rows are sorted by p-value.

Performance Metric	Data Metric	Statistical Test	p-value
AMR genes in top 20	num. genomes	Spearman-R	<0.00001
Test MCC in 5CV	species	Kruskal-Wallis	<0.00001
AMR genes in top 20	minority fraction	Spearman-R	0.00464
Test MCC in 5CV	num. genomes	Spearman-R	0.00537
AMR genes in top 20	species	Kruskal-Wallis	0.00721
Test MCC in 5CV	drug class	Kruskal-Wallis	0.30250
AMR genes in top 20	drug class	Kruskal-Wallis	0.48732
Test MCC in 5CV	minority fraction	Spearman-R	0.95993

Table C.8. Summary of known AMR gene-drug mappings recovered by Fisher's exact test but missed by the SVM ensemble approach. Drugs abbreviated are quinupristin-dalfopristin (Q-D) and trimethoprim-sulfamethoxazole (SXT).

Species	Drug	Missed Gene	Model Test MCC (5CV)	Gene Rank in Model	Statistically Unique Features in Top 50*	Possible Failure Mode
Acinetobacter baumannii	tobramycin	*aadA*	0.87	33	44	Feature rank below top 20 threshold
Enterobacter cloacae	cefepime	*blaKPC*	0.61	-	46	Poor model performance
Enterococcus faecium	Q-D	*eatAv*	1.00	-	1	High correlation among top features
Enterococcus faecium	teicoplanin	*vanXA*	1.00	21	11	Feature rank below top 20 threshold
Enterococcus faecium	teicoplanin	*vanA*	1.00	21	11	Feature rank below top 20 threshold
Enterococcus faecium	teicoplanin	*vanHA*	1.00	21	11	Feature rank below top 20 threshold
Enterococcus faecium	vancomycin	*vanZA*	0.99	-	16	High correlation among top features
Enterococcus faecium	vancomycin	*vanYA*	0.99	-	16	High correlation among top features
Escherichia coli	norfloxacin	*parC*	0.43	-	36	Poor model performance
Klebsiella pneumoniae	SXT	*sul1*	0.84	-	13	High correlation among top features
Neisseria gonorrhoeae	erythromycin	*mtrR*	0.77	-	48	Poor model performance
Staphylococcus aureus	ciprofloxacin	*arlR*	0.97	-	46	-

*Refers to the number of features remaining among the SVM model's top 50 features by weight after collapsing perfectly correlated features together. Lower values correspond to more highly correlated features.

Table C.9. 13 gene flanking noncoding variants predicted to be associated with resistance against specific drug classes for individual species.

Species	Drug Class	Accession (Gene)	Predicted Gene Product	Mutations*	Resistant/ Susceptible	LORs
5' flanking variants						
Acinetobacter baumannii	beta-lactam	AGQ10471.1 (*pqqA*)	Coenzyme PQQ synthesis protein A	-12_2delTG ATTTAAT CAAGTG**	CTX=73/0 CRO=73/0 AMP=73/0	CTX=6.4 CRO=6.4 AMP=6.4
Acinetobacter baumannii	tetracycline	WP_000096554.1 (*clpV*)	T6SS AAA+ chaperone	Most common variant	TET=241/12 MIN=71/18	TET=3.3 MIN=2.3
Campylobacter coli	quinolone	WP_002805020.1	GNAT acetyl-transferase	Most common variant	CIP=94/71 NAL=95/70	CIP=6.0 NAL=5.5
Campylobacter coli	quinolone	WP_002783313.1	Transmembrane transport protein, MFS	Most common variant	NAL=97/115 CIP=95/117	NAL=5.5 CIP=5.4
Escherichia coli	beta-lactam	AAG56074.1 (*ymgF*)	Small inner membrane protein	9*** (see below)	CTX=28/0 CAZ=30/1 CXM=14/1	CTX=6.7 CAZ=5.7 CXM=4.6
Escherichia coli	quinolone	WP_000017703.1 (*hybB*)	Ni/Fe-hydrogenase 2 b-type cytochrome subunit	Most common variant	LVX=260/56 CIP=606/606 NAL=45/16	LVX=3.3 CIP=2.8 NAL=1.4
Streptococcus pneumoniae	beta-lactam	WP_000449822.1	Thiaminase II	Most common variant	AMX=13/4 MEM=16/1 CXM=16/1	AMX=12.4 MEM=9.6 CXM=9.3
Streptococcus pneumoniae	beta-lactam	ADI69655.1 (*nplT*)	Neopullulanase	Most common variant	AMX=13/3 MEM=15/1 CXM=15/1	AMX=12.7 MEM=9.1 CXM=8.6
Streptococcus pneumoniae	beta-lactam	WP_000592948.1 (*ugl*)	Unsaturated chondroitin disaccharide hydrolase	-186C>T	AMX=13/6 CXM=17/2 MEM=17/3	AMX=11.9 CXM=9.6 MEM=9.2
3' flanking variants						
Campylobacter coli	quinolone	WP_002777456.1 (*hspR*)	Transcriptional repressor of DnaK operon	Most common variant	NAL=98/115 CIP=96/117	NAL=7.4 CIP=7.4
Klebsiella pneumoniae	beta-lactam	WP_000679427.1 (*qacE*Δ*1*)	Small multidrug resistance (SMR) efflux transporter	Most common variant	AMP=759/0 CEF=27/0 CRO=772/4	AMP=10.3 CEF=5.5 CRO=4.3
Salmonella enterica	quinolone	WP_012772747.1	Psp operon transcriptional activator	Most common variant	NAL=9/11 CIP=9/10	NAL=3.5 CIP=3.2
Streptococcus pneumoniae	beta-lactam	WP_000145597.1 (*nplT*)	Neopullulanase	Most common variant	AMX=13/3 MEM=15/1 CXM=15/1	AMX=12.7 MEM=9.1 CXM=8.6

Accession IDs are provided for the most common coding variant of the corresponding gene cluster (RefSeq when possible, GenBank otherwise), along with gene names when available and gene products. Mutations are defined relative to the most common 5'/3' variant for the corresponding gene. The number of resistant/susceptible genomes and $\log_2$ odds ratios (LORs) for resistance are shown for the top three drugs by LOR when data for more than three related drugs was available. Drug abbreviations and mappings to relevant sequences are available in Dataset C.7.

*Mutations are denoted relative to the start codon for 5' variants. Position -1 corresponds to the first base pair immediately adjacent on the 5' side of the gene's start codon.

**Results in the deletion of a GTG start codon and the 12 base pairs immediately adjacent on the 5' side, with respect to the most common variant. The candidate variant begins with an ATG start codon.

***Mutations: -19A>C, -23A>G, -25_24delTC, -30insA, -48G>A, -173A>T, -175C>A, -211A>T, -213C>T

C.4 Supplementary Datasets

Dataset C.1. PATRIC genome IDs for all genomes used for global pathogenomic analysis.

Dataset C.2. Consolidated SIR phenotypes derived from directly reported SIRs and inference from MICs available on PATRIC. Also includes genome MICs, MIC-SIR mappings used for SIR inference, and most common testing standard for each species-drug case determined from PATRIC annotations or manual curation of contributing BioProjects.

Dataset C.3. Distribution of unique AMR genes across 12 species and cross-species AMR gene analysis. Includes counts of gene-drug mappings for each species-drug case, AMR gene category abbreviations, drug class assignments, curation of AMR gene annotations, and assignments of AMR genes to drugs. Species distributions, localization predictions, and function classifications for re-clustered cross-species AMR genes are also included.

Dataset C.4. Distribution of complete blaTEM alleles detected across 12 pathogens and plasmid predictions for contigs containing TEM-116.

Dataset C.5. Summary of SVM model performance and top predictive features across 127 species-drug cases. Includes dataset properties, SVM mean test set MCC across 5-fold cross validation, final hyperparameter choices, known AMR genes recovered by SVM or Fisher's exact test, and lists of the top 50 features for each model.

Dataset C.6. Sequences associated with the top 50 features from each SVM model across 127 species-drug cases. Exact sequences for all such features are provided, as well as the top two most common variants of each type for all sequence clusters related to the features.

Dataset C.7. Filtering results for identifying and categorizing novel AMR gene candidates from SVM models. Also includes abbreviations for drug names.

Dataset C.8. Cell densities achieved by *cycA* and *frdD* mutants under various antibiotic stresses, base media, and supplements. Results of statistical tests between densities achieved for different strains or conditions and predicted *ampC* transcription rates for *frdD* mutants are also included.

C.5 References

[1] Jason C Hyun, Jonathan M Monk, and Bernhard O Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022.

[2] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055, July 2015.

[3] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[4] Brian P Alcock, Amogelang R Raphenya, Tammy T Y Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, Sally Y Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E Werfalli, Jalees A Nasir, Martins Oloni, David J Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N Sharma, Emily Bordeleau, Andrew C Pawlowski, Haley L Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L Winsor, Robert G Beiko, Fiona S L Brinkman, William W L Hsiao, Gary V Domselaar, and Andrew G McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, 48(D1):D517–D525, January 2020.

[5] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.*, 46(6):e35, April 2018.

[6] Brian D Ondov, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1):132, June 2016.

[7] Peter J A Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J L de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, June 2009.

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. 2012.

[9] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, 38(12):5825–5829, December 2021.

[10] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A Datsenko, Masaru Tomita, Barry L Wanner, and Hirotada Mori. Construction of escherichia coli K-12 in-frame, single-gene knockout mutants: the keio collection. *Mol. Syst. Biol.*, 2(1):2006.0008, February 2006.

Appendix D

Reconstructing the core genome of the last bacterial common ancestor - Supplementary Information

D.1 Methods

D.1.1 Pangenome construction

Genomes from the Web of Life collection [1] were filtered based on the following criteria: 1) GTDB species classification (release 202) is available [2], 2) CheckM contamination <10% and completeness >80% [3], 3) number of contigs is within three times the median number of contigs for all assemblies for the genome's species, and 4) total assembly length is within three standard deviations from the mean of all assemblies for the genome's species. 183 bacterial species defined by GTDB were identified with at least 50 genomes passing all criteria, totalling 54,085 genomes. NCBI accession IDs, GTDB phylogenetic classifications, and genome quality metrics for selected genomes are available in Dataset D.1.

Open reading frames for each genome were identified using Prodigal v2.6.3 with parameters "-c", "-m", "-g", "11", "-p", "single", "-q" [4], based on those used in by the Prokka annotation platform [5]. For each species, genes and protein sequence variants were identified using the approach previously described [6]. Briefly, protein sequences were

clustered using CD-HIT v4.6 with minimum identity 80% and minimum alignment length 80% [7] and clusters were treated as genes.

D.1.2 Species-level phylogeny construction

A species-level phylogeny for the selected 183 species was extracted from the genome-level Web of Life phylogeny as follows. For each species, the most recent common ancestor (MRCA) of its selected genomes was identified. For 19 species where the MRCA contained children from multiple GTDB species (i.e. was not monophyletic), an alternate MRCA representative was identified by computing for each MRCA child node its completeness (fraction of the species' selected genomes present among the node's children) and purity (fraction of the node's children that are of that species), and selecting the node with the highest harmonic mean between completeness and purity. Each MRCA was extended by the median distance to its species' selected genomes to model a modern representative of the species. Finally, the phylogeny was pruned such that the only leaf nodes are the 183 species representatives, and internal nodes with only 1 child were collapsed by adding distances to yield a binary phylogenetic tree. The final species-level tree and representative node assignments are available in Dataset D.2.

D.1.3 Gene frequency estimation

True gene frequencies were estimated by modeling the log-likelihood (LL) of an observed gene presence/absence matrix $\mathbf{X}$ as follows:

$$LL(\mathbf{X}, \mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \log(p_i q_j) + (1 - x_{ij}) \log(1 - p_i q_j)$$

where $x_{ij} = 1$ if gene i was observed in genome j or 0 otherwise, p_i is the true frequency of gene i in the genome collection, q_j is the gene recovery rate of genome j, m is the total number of genes, and n is the total number of genomes. Variables $\mathbf{p}$ and $\mathbf{q}$

were estimated by maximizing the LL using SciPy [8] with the L-BFGS-B method, with bounds $[10^{-8}, 1-10^{-8}]$ for all variables and initial guesses 0.99 for all q_j and respective observed gene frequencies (fraction of genomes with the gene) for p_i. Estimation was limited to genes with observed frequency >10%, and maximization was accelerated using exact gradients:

$$\frac{\partial LL}{\partial p_k} = \sum_{j=1}^{n} \frac{x_{kj}}{p_k} - \frac{(1-x_{kj})q_j}{1-p_k q_j}$$

$$\frac{\partial LL}{\partial q_k} = \sum_{i=1}^{m} \frac{x_{ik}}{q_k} - \frac{(1-x_{ik})p_i}{1-p_i q_k}$$

D.1.4 Benchmarking gene frequency estimation

Estimated gene recovery rates were compared to quality metrics available on GTDB. For computing modes and fitting gene frequency distributions, distributions were discretized into bins of size 1/number of genomes for each species. For fitting frequency distributions, distributions were first reduced to those with observed frequency >90% (potential core genes). Both observed and estimated distributions were fit to a power model or exponential model by fitting their corresponding cumulative distributions (Fig. D.3c), where $P(x)$ is the number of genes with frequency x and $F(x)$ is the number of genes with frequency $\leq x$. Models were fit to cumulative distributions rather than directly to frequency distributions to avoid fitting data spanning multiple orders of magnitude. Parameters were estimated using the SciPy curve_fit routine [8], with observed $F(x)$ scaled to maximum value 1 and with initial guesses for (k,c,a) of (1,1,2) for the power model and (1,1,1) for the exponential model. All parameters were bounded positive except for the power model a parameter which was bounded $a > 1$. Parameters from fitting $F(x)$ were applied directly to the corresponding $P(x)$ function and quality of fit was evaluated through R^2 and mean absolute error (MAE).

D.1.5 Core genome identification and annotation

For each species, the core genome was defined as all genes with estimated frequency >99.99%. The most common sequence variant of each gene in each species was annotated using eggNOG-emapper v2.1.6-43 [9] to map each gene to COG functional categories and orthogroup (OG), using highest depth annotations. All cross-species analyses were conducted at the OG level. OG gene names were assigned by comparing annotations from eggNOG-emapper, the 2020 COG database [10], and UniProtKB [11]. Under-characterized OGs were identified by first filtering for OGs assigned to the category "S: Function unknown" or corresponding to a gene name starting with "y", then identifying those with low UniProtKB annotation scores as of September 7, 2022.

D.1.6 LBCA core genome reconstruction, analysis, and comparison with JCVI-Syn3A

Each species' node in the species-level phylogeny was assigned the species' core genome OGs as previously identified. OG content of internal ancestral nodes in the phylogeny was estimated using Count [12] under asymmetric Wagner parsimony [13] for gain/loss penalty ratios ranging from 0.1 to 2.0 in 0.05 steps. The last bacterial common ancestor (LBCA) core genome for a given penalty ratio was taken as the set of OGs predicted for the root node of the phylogeny. COG functional categories, OG names, and under-characterized OGs were assigned the same as for modern core genomes. OG systems were assigned from the 2020 COG database [10].

For system case studies, the set of 55 bacterial 50S and 30S ribosomal proteins was taken from Yutin, et.al. 2012 [14], excluding proteins S22 and S31e which did not have COG database IDs. Aminoacyl-tRNA synthetases were limited to those corresponding to the 20 canonical amino acids and classifications were taken from Gomez and Ibba, 2020 [15]. Translation factors were taken directly from the 2020 COG databases after filtering out eukaryotic proteins. Finally, for comparison with JCVI-Syn3A, the JCVI-

Syn3A assembly with predicted ORFs (GenBank: CP016816.2) was similarly annotated with eggNOG-emapper to assign proteins to OGs.

D.1.7 LBCA core metabolism reconstruction

LBCA core OGs were mapped to KEGG orthogroups [16] through eggNOG-emapper annotations, by first identifying all sequences annotated with a given OG and taking the most common KEGG annotation among those sequences. LBCA core KEGG orthogroups were then mapped to KEGG modules by evaluating KEGG module definitions as logical expressions. KEGG modules with at least one KEGG orthogroup corresponding to an OG present in the LBCA core at $g = 1$ were visualized. Multi-enzyme reactions were treated as active if all necessary orthogroups were present. Missing reactions were manually verified for alternate pathways on KEGG, for other KEGG orthogroups capable of catalyzing the reactions and for other COG orthogroups that could be mapped to the missing KEGG orthogroup based on curating eggNOG text annotations.

D.2 Supplemental Discussion

D.2.1 Benchmarking estimation of genome-specific gene recovery rates and true gene frequencies

True gene frequencies (p_i) and genome-specific gene recovery rates (q_j) were estimated simultaneously by maximizing the log-likelihood of the observed presence/absence matrix, where $x_{ij} = 1$ if gene i was observed in genome j or 0 otherwise, and m and n are the total number of genes and genomes, respectively:

$$LL(\mathbf{X}, \mathbf{p}, \mathbf{q}) = \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \log(p_i q_j) + (1 - x_{ij}) \log(1 - p_i q_j)$$

Applied to the 183 species' pangenomes, we find that the estimated genome-specific gene recovery rates are correlated with existing metrics of genome assembly quality.

The median Spearman correlations across all species between the estimated rates and negative metrics of quality (L50, number of contigs, CheckM strain heterogeneity, CheckM contamination) were negative, while such correlations with positive metrics of quality (N50, CheckM completeness) were positive (Fig. D.2b). Particularly, the rates were consistently strongly correlated with CheckM completeness (median Spearman correlation 0.522), in line with the intuition that these rates represent catch-all probabilities for whether a given gene in a strain is recovered by each genome's assembly and annotation process. Across the 3,451 genomes with exactly one contig, CheckM completeness and gene recovery rates were strongly correlated (Pearson correlation 0.724, Fig. D.2c), and 65% of those genomes achieved gene recovery rates above 99.9999% (Fig. D.2d). These results suggest that estimated gene recovery rates derived from the pangenome matrix alone are consistent with several commonly used metrics of assembly quality.

We find that the estimated gene frequency distributions are robust to sampling and consistently yield frequency peaks at 100%, in line with the existence of core genomes. In bootstrapping experiments of six phylogenetically diverse species where 90% of genomes were randomly sampled and frequencies were re-estimated (10 samplings per species), the lowest Spearman correlation between frequencies from pairs of samplings exceeded 94% for all species and 97% in 3/6 species (Fig. D.2e). Examining potential core genes (observed frequency >90%), the modes of estimated frequencies using bins of size 1/number of genomes was 100% for all 183 species, while the modes of observed frequencies was less than 100% in 67/183 (37%) species (Fig. D.3a-b).

Estimated frequency distributions also had a more consistent functional form than observed frequency distributions. Observed and estimated frequency distributions of potential core genes were fitted to two previously proposed models for gene frequency distributions (Fig. D.3c, exponential [17] and power function [6]). Across the 183 species, the estimated frequencies more consistently fit the power model than observed frequencies fit either model, with a minimum R^2 of 0.69 compared to 0.00, 0.02, and 0.05 of the

other fitting combinations, and a median R^2 of 0.96 compared to 0.82, 0.92, and 0.74 (Fig. D.3d-e).

Finally, core genomes defined from estimated frequencies were of similar size to those defined using traditional approaches. Three core genomes were defined for each species as genes having either 1) estimated frequency >99.99%, 2) observed frequency >99%, or 3) observed frequency >95%. The sizes of estimation-based core genomes were between that of the 99% and 95% observed frequency core genomes and strongly correlated ($r = 0.79$ against 99% core genomes, $r = 0.93$ against 95% core genomes, Fig. D.4a-b). Gene content was more variable between definitions, with median Jaccard similarities between estimation-based core genomes and the 99% and 95% observed frequency core genomes of 0.73 and 0.77 respectively, and 0.61 between the 99% and 95% core genomes (Fig. D.4c). Altogether, these analyses suggest that this estimation approach yields 1) genome-specific gene recovery rates correlated with common assembly quality metrics, 2) estimated frequencies that are robust to sampling and consistent with the existence of core genomes, 3) frequency distributions with a more consistent shape than observed distributions, and 4) core genomes of similar size but different content to those defined using traditional approaches. Original values reported by these analyses are available in Dataset D.3.

D.2.2 Assessing the absence of universal core orthogroups

The absence of any universal orthogroups (OGs) present in all core genomes was not sensitive to the 99.99% frequency threshold for core genes. At more relaxed 99.9% and 99% thresholds, all OGs were still missing in at least 7 and 5 core genomes, respectively (Fig. D.7a-b). Furthermore, 128 core genomes were missing at least one of the 28 highly conserved OGs (those found in >90% of core genomes) and were from a wide range of species not limited to a specific phylogenetic group (Fig. D.7c). An alternate approach for identifying universal core OGs by looking at each OG's minimum estimated frequency

had similar results: only five OGs had frequency >90% in all species (maximum 94%) and 20 OGs had frequency >50% in all species, which were also predominantly ribosomal or other translation-associated proteins (Fig. D.7d).

D.2.3 Ambiguity in mapping between COG and KEGG orthogroups

While most metabolic COG orthogroups mapped to exactly one or zero KEGG orthogroups, there were two instances where a COG orthogroup could not be unambiguously assigned to KEGG orthogroup to determine the presence of a reaction in the LBCA core. For phosphogluconolactonase *pgl* in the pentose phosphate pathway (Fig. 5.5b), the COG orthogroup COG0363 represents both *pgl* and *nagB*, glucosamine-6-phosphate deaminase, making it impossible to determine unambiguously whether *pgl* is present in the LBCA core genome with this approach. Similarly, in methionine biosynthesis, the COG orthogroup COG0626 represents both *metB* and *metC*, making it impossible to distinguish the two steps' presence in the LBCA core genome using the current approach. In the latter case, regardless of whether *metB* and/or *metC* are active, methionine could also be produced through an acetyl-homoserine intermediate instead of through succinyl-homoserine via *metX* and *metY*.

D.3 Supplementary Figures

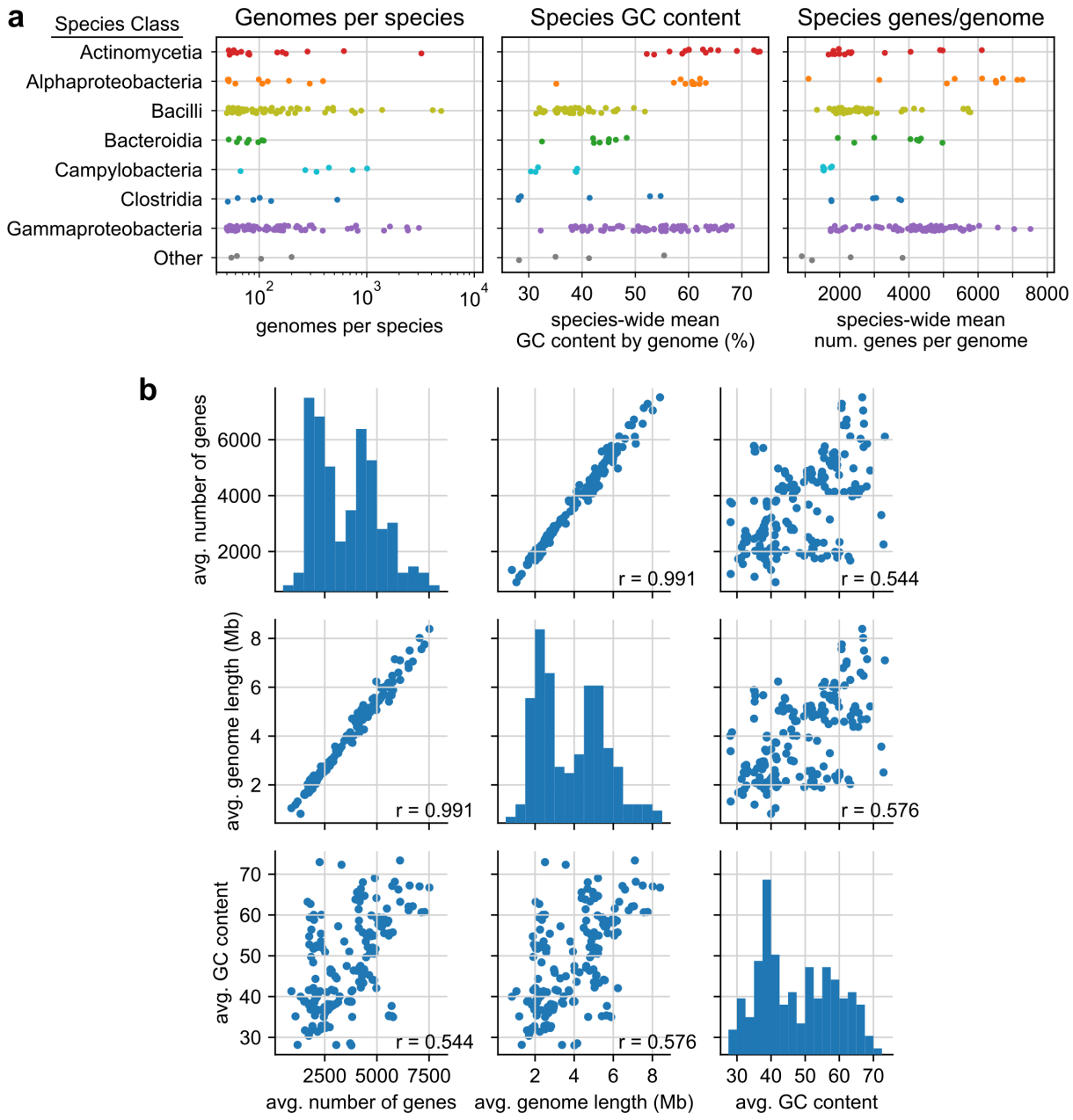


Figure D.1. Properties of selected Web of Life genomes by species and phylogenetic class. a) Genome collection sizes, mean GC content, and mean genes per genome for each species by phylogenetic class. b) Distribution of species-wide averages in number of genes per genome, genome length, and GC content. Diagonal panels show histograms for individual variables, and off-diagonal panels show scatterplots between each pair of variables. Pearson correlation coefficients are shown for each pair.

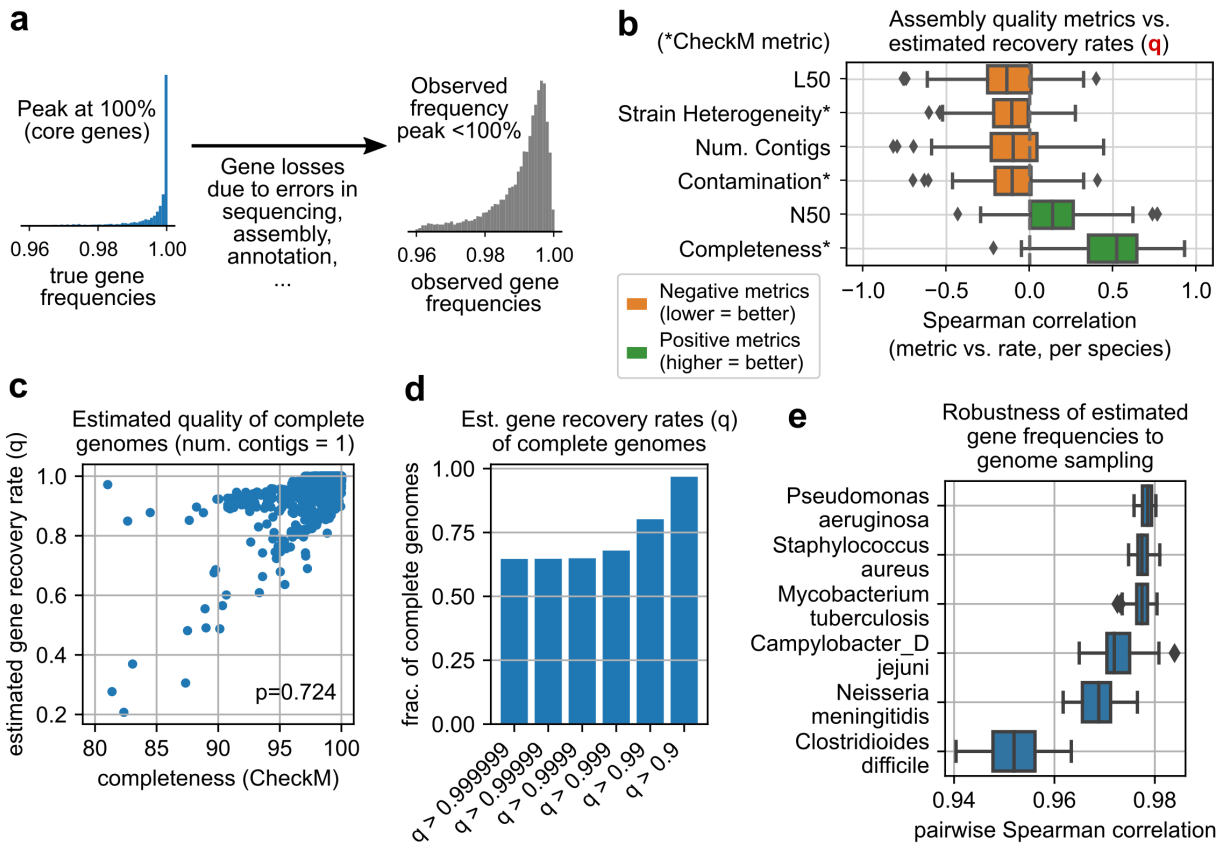


Figure D.2. Analysis of estimated gene recovery rates and gene frequencies by species with respect to genome quality and robustness to random genome sampling. (a) Gene losses due to artifacts in the genome assembly process shift the distribution of observed gene frequencies such that the most common frequency for potential core genes is not 100%. (b) Distribution of correlations between estimated recovery rates and genome assembly quality metrics across 183 species' pangenomes. (c) Comparison of CheckM completeness and estimated gene recovery rates for complete genomes (num. contigs = 1). (d) Fraction of complete genomes above various gene recovery rate thresholds. (e) Robustness of estimated frequencies to genome sampling for selected species. Pairwise Spearman correlations between estimated gene frequencies calculated from randomly sampled genome sets (90% of genomes were sampled 10 times, per species). Species are sorted by median Spearman correlation.

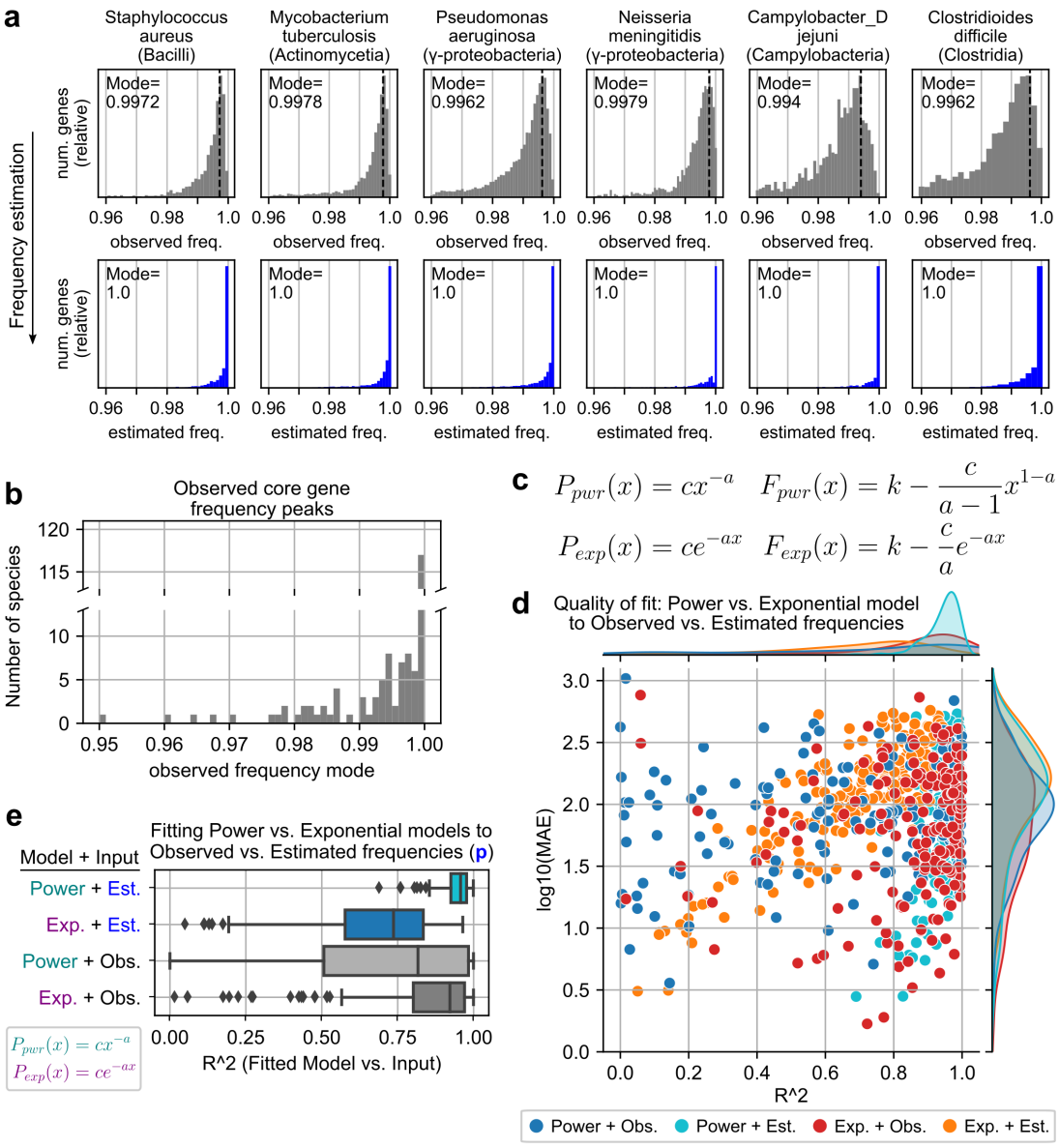


Figure D.3. Analysis of estimated gene frequency distributions by species. (a) Comparison of observed and estimated gene frequencies near 100% for selected species, sorted by number of genomes and labeled with phylogenetic class. Modes were calculated using bins of size 1/number of genomes. (b) Distribution of observed frequency modes for all 183 species across genes with frequency >50%. (c) Power and exponential functions for modeling frequency distributions $P(x)$ and corresponding cumulative distributions $F(x)$, where c, a, and k are fitted parameters. (d) Quality of fit (R^2 and log mean absolute error) between either observed or estimated frequencies and either the power or exponential model. e) Distribution of R^2 between observed or estimated frequencies and either the power or exponential model.

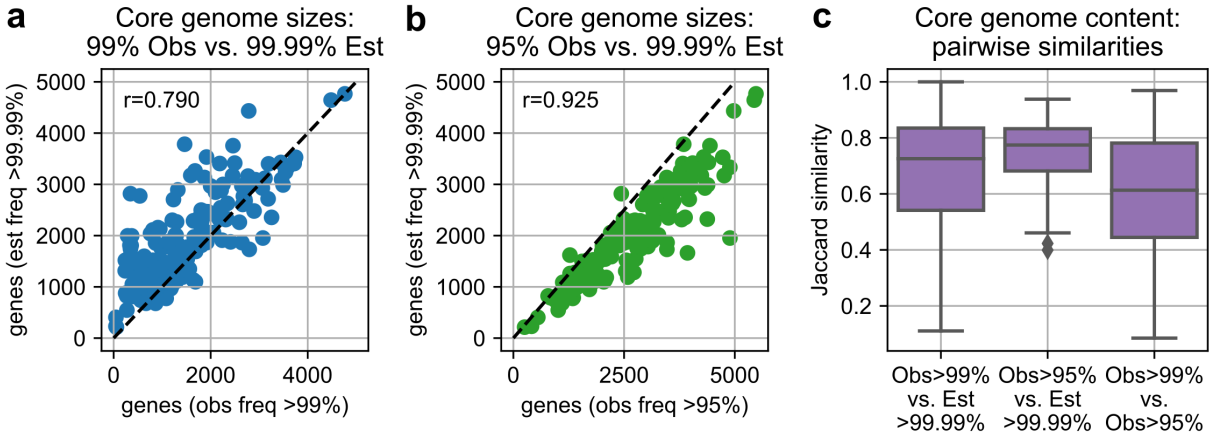


Figure D.4. Pairwise comparisons between three definitions of core genome across 183 species. (a) Pearson correlation between the number of genes with observed frequency >99% vs. estimated frequency >99.99%. (b) Pearson correlation between the number of genes with observed frequency >95% vs. estimated frequency >99.99%. (c) Pairwise Jaccard similarities between the set of genes with either estimated frequency >99.99%, observed frequency >99%, or observed frequency >95%.

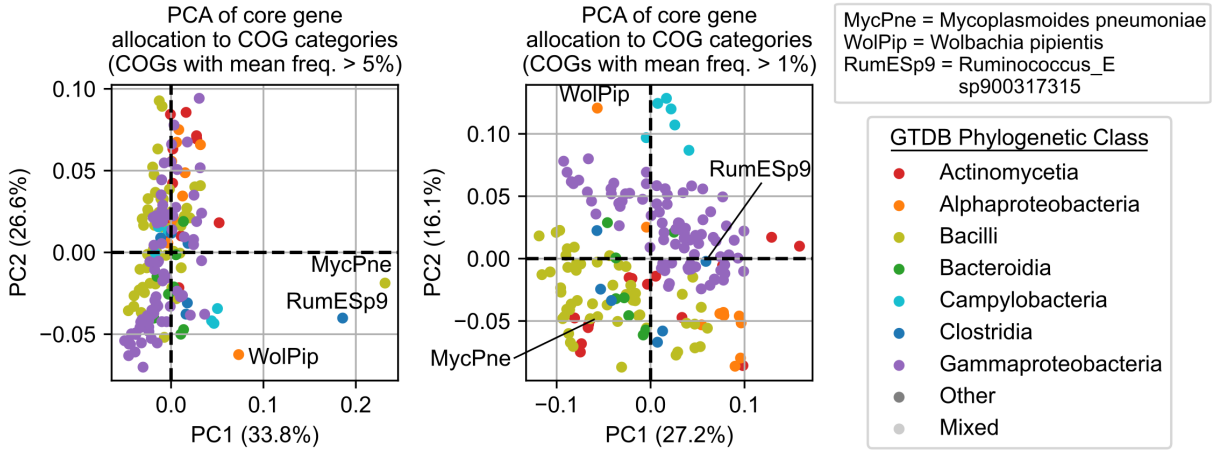


Figure D.5. Principal component analysis of core genome allocation of genes to functional categories. PCA was applied to a matrix where each species was represented as a row vector and each element corresponds to the fraction of the species' core genes that was assigned to a specific COG functional category. On the left, COG categories with frequency <5% were grouped together and on the right COG categories with frequency <1% before applying PCA, and projections onto the first two components and percent of variance explained are shown. Outliers from the 5% plot are labeled.

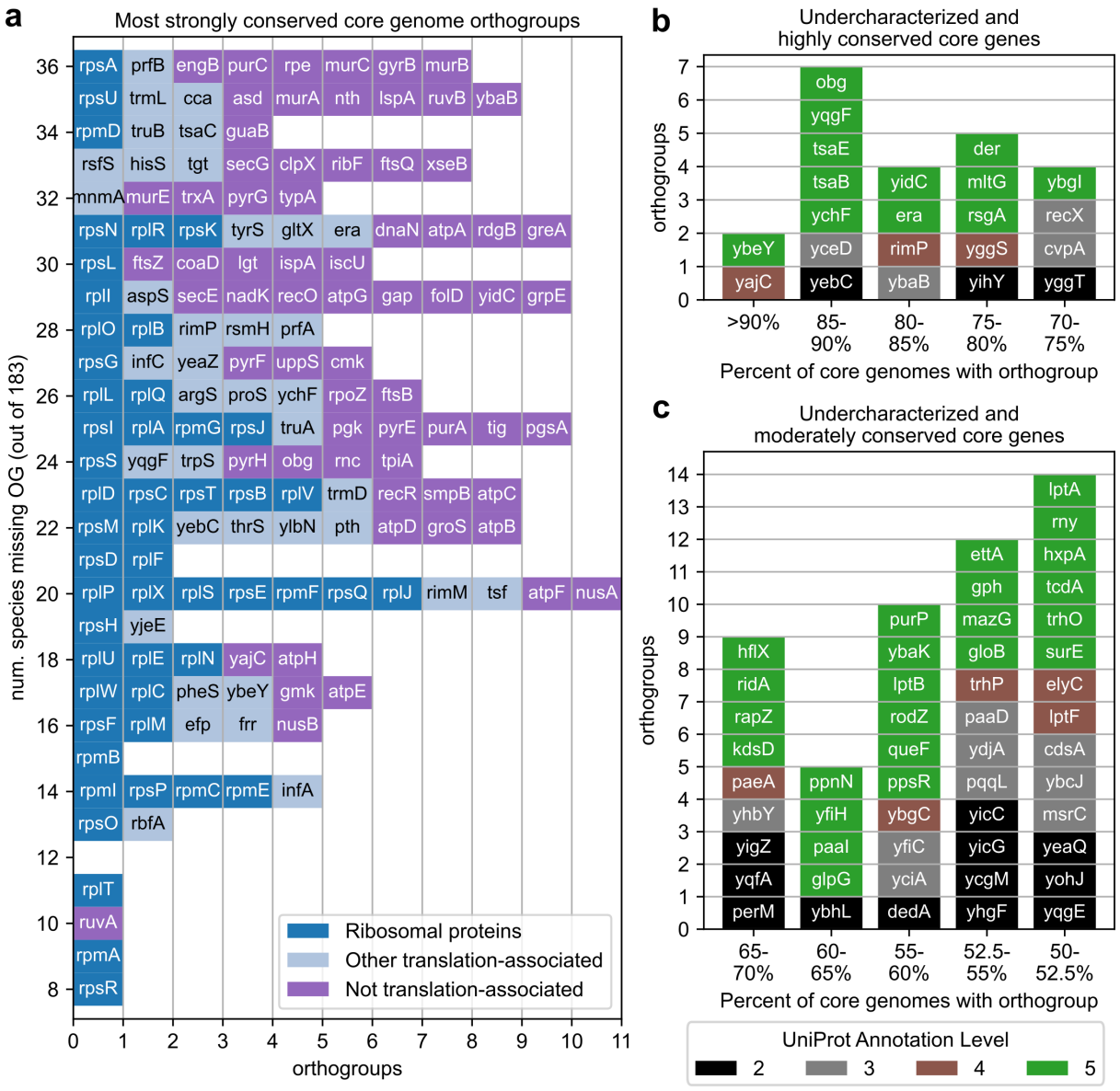


Figure D.6. Survey of orthogroups frequently observed across 183 core genomes. (a) Orthogroups observed in at least 80% of core genomes, sorted by frequency of observation and function. No orthogroup was observed in more than 175 core genomes (i.e. missing in fewer than 8). (b-c) Under-characterized orthogroups observed in at least 50% of core genomes. Orthogroups were identified as potentially under-characterized if assigned to COG category "S: Function unknown" or associated with a y-gene. Orthogroup gene name and level of annotation was corroborated with UniProtKB.

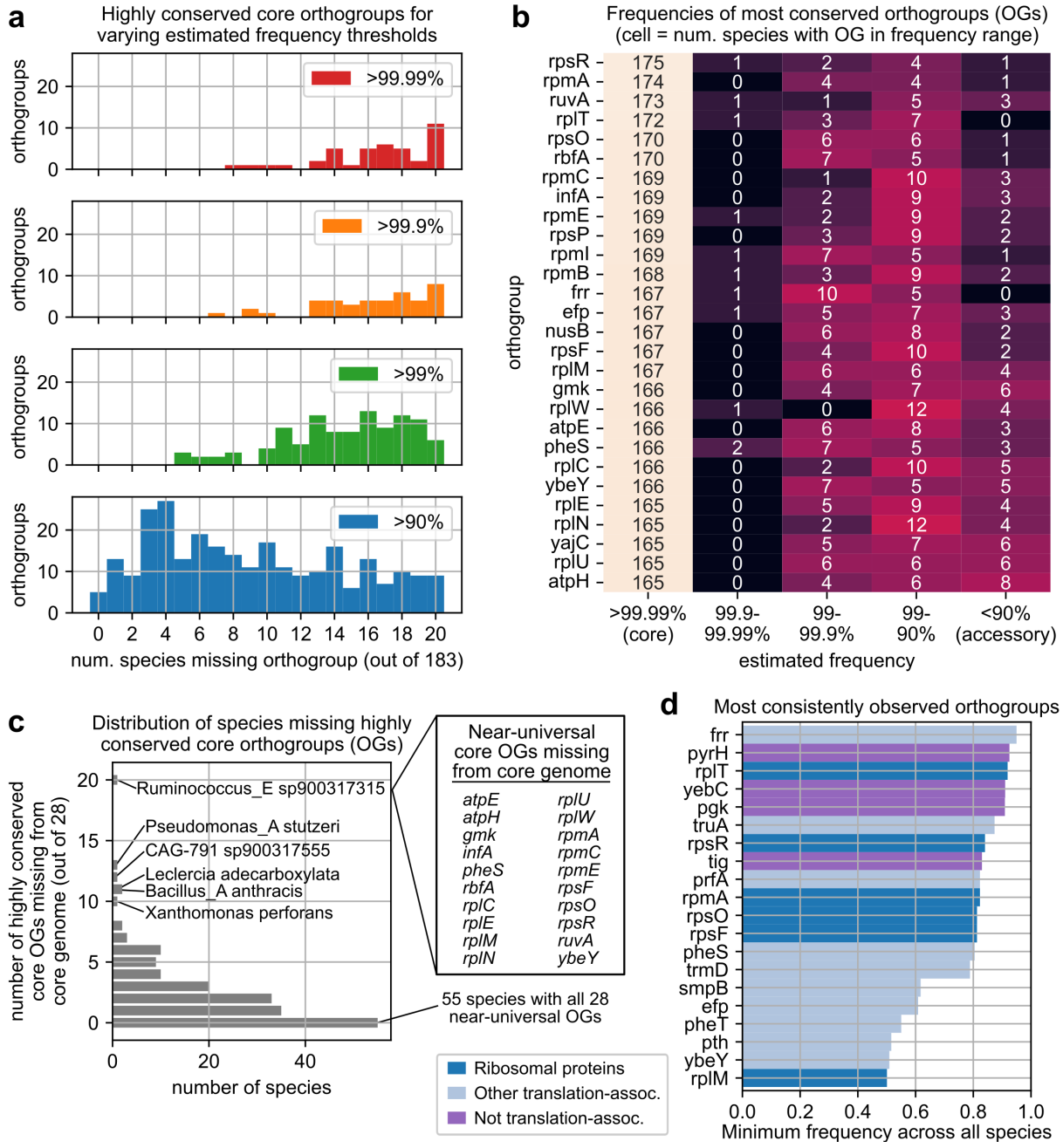


Figure D.7. Sensitivity of conserved orthogroups to core genome frequency threshold. (a) Distribution of orthogroups frequently observed in core genomes for different frequency thresholds for defining core genes. (b) Distribution of estimated frequencies of most conserved orthogroups. (c) Distribution of the number of missing, highly conserved core orthogroups across species. Species missing the most such orthogroups are labeled, and the 20 orthogroups missing from *Ruminococcus_E sp900317315* are shown. (d) Orthogroups with the highest minimum estimated frequency across all 183 species.

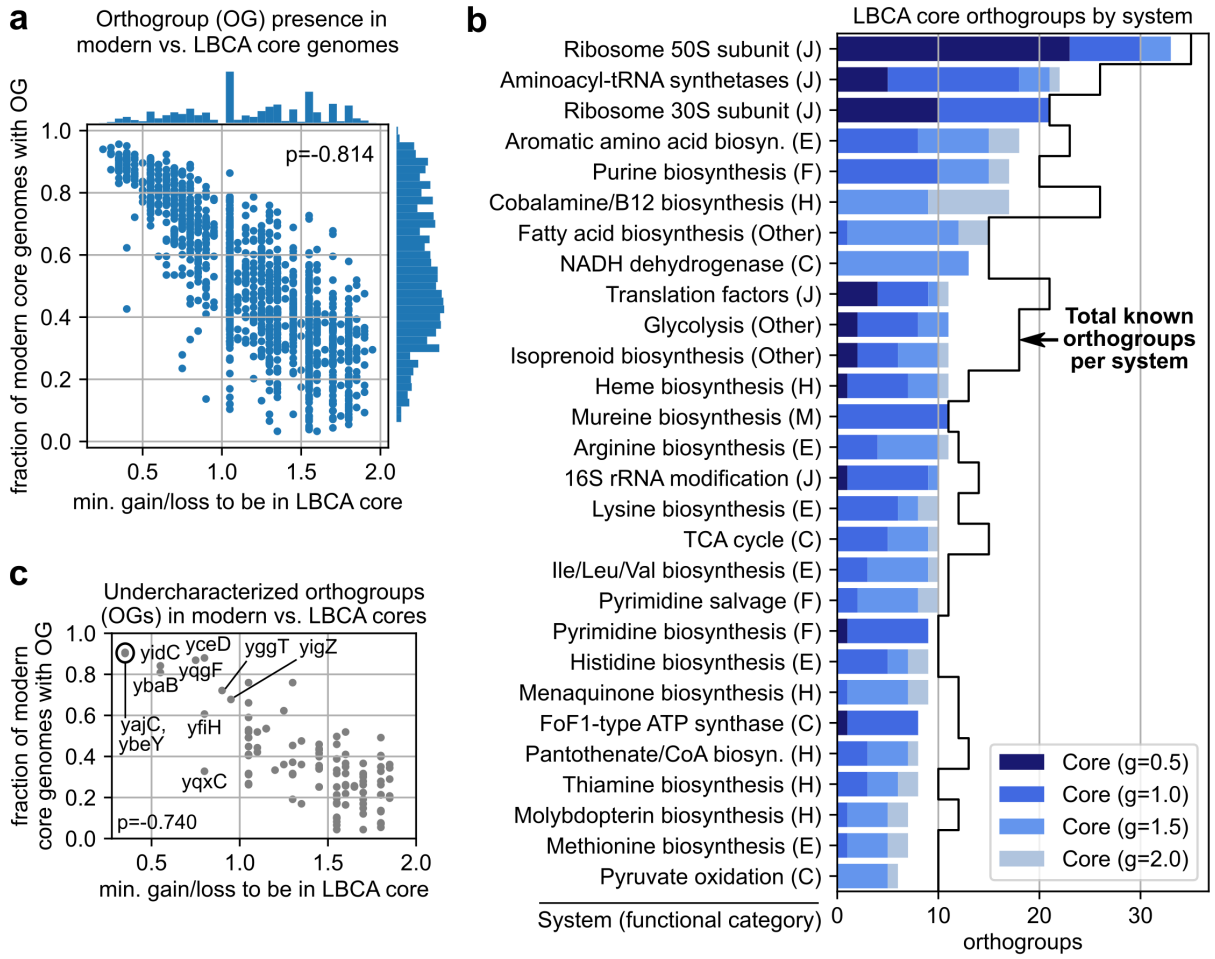


Figure D.8. Additional analyses of the LBCA core genome gene content. (a) Relationship between the minimum gain/loss penalty during ancestor reconstruction (g) for an orthogroup to be predicted to be in the LBCA core genome vs. the frequency at which it is observed in modern core genomes. Pearson correlation coefficient is shown. (b) Expanded distribution of LBCA core orthogroups by biological system and minimum gain/loss penalty to be observed. Systems with at least 10 orthogroups of which at least 50% are present in the LBCA core genome at $g = 2.0$ are shown. (c) Under-characterized LBCA core orthogroups, by minimum gain/loss penalty and fraction of 183 modern bacterial core genomes with the orthogroup. Orthogroups observed for $g < 1.0$ are labeled, and the Pearson correlation coefficient is shown.

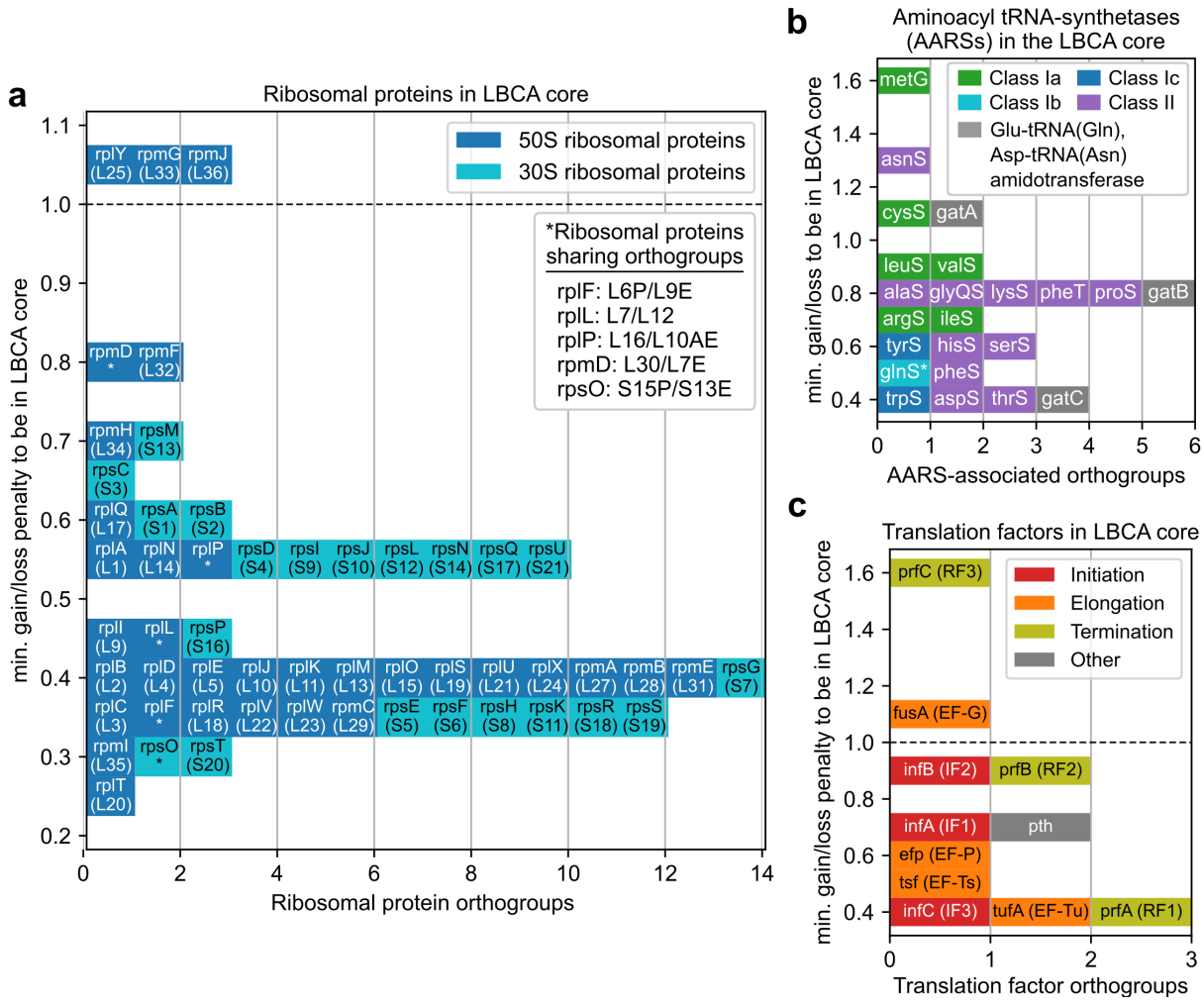


Figure D.9. Distribution of LBCA core orthogroups related to ribosomal proteins, aminoacyl-tRNA synthetases, and translation factors. The minimum gain/loss penalty (during ancestor reconstruction via asymmetric Wagner parsimony) at which each orthogroup is predicted to be in the LBCA core genome is shown for (a) ribosomal proteins, (b) aminoacyl-tRNA synthetases, and (c) translation factors. Predicted gene names and corresponding protein names are shown. Starred gene *gln*S* refers to a single orthogroup containing both GluRS and GlnRS.

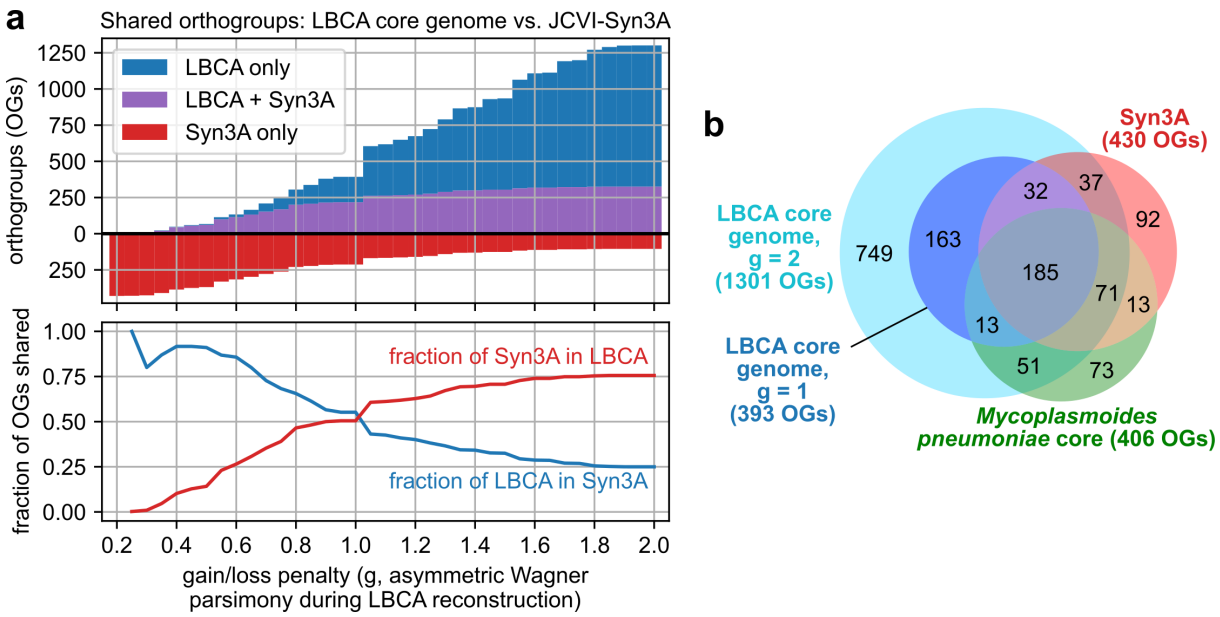


Figure D.10. Breakdown of gene overlap between the LBCA core genome and genome of minimal organism JCVI-Syn3A. (a) Shared orthogroups (OGs) between the LBCA core genome and Syn3A as the gain/loss penalty ratio (g) is increased during ancestral reconstruction through asymmetric Wagner parsimony (resulting in larger, less stringently defined LBCA core genomes). (b) Shared OGs between Syn3A, the LBCA core genome (at $g = 1$ and $g = 2$), and the core genome of *Mycoplasmoides pneumoniae*.

D.4 Supplementary Tables

Table D.1. 28 under-characterized orthogroups frequently observed across the core genomes of 183 species. Columns left-to-right correspond to predicted gene name, fraction of core genomes with the orthogroup, COG ID, UniProt ID (*E. coli*) when available, COG functional category, UniProt annotation level (Sept. 2022), and orthogroup annotation.

Gene Name	Frac. Core	Orthogroup	UniProt ID	COG cat.	Annot. Level	Annotation
yebC	0.8798	COG0217	P0A8A0	K	2	Transcriptional and/or translational regulatory protein YebC/TACO1
yceD	0.8798	COG1399	P0AB28	S	3	23S rRNA accumulation protein YceD (essential in plants, uncharacterized in bacteria)
ybaB	0.8087	COG0718	P0A8B5	S	3	DNA-binding nucleoid-associated protein YbaB/EfbC
yihY	0.7596	COG1295	P0A8K8	S	2	Unchar. membrane protein, BrkB/YihY/UPF0761 family (not an RNase)
cvpA	0.7377	COG1286	P08550	S	3	Colicin V production accessory protein CvpA, regulator of purF expression and biofilm formation
yggT	0.7213	COG0762	P64564	S	2	Cytochrome b6 maturation protein CCB3/Ycf19 and related maturases, YggT family
recX	0.7104	COG2137	P33596	S	3	SOS response regulatory protein OraA/RecX, interacts with RecA
yhbY	0.6940	COG1534	P0AGK4	J	3	RNA-binding protein YhbY
yigZ	0.6776	COG1739	P27862	S	2	Putative translation regulator, IMPACT (imprinted ancient) protein family
perM	0.6667	COG0628	P0AFI9	S	2	Predicted PurR-regulated permease PerM
yqfA	0.6612	COG1272	P67153	S	2	Predicted membrane channel-forming protein YqfA, hemolysin III family
ybhL	0.6230	COG0670	P0AAC4	S	2	Integral membrane protein YbhL, putative Ca2+ regulator, Bax inhibitor (BI-1)/TMBIM family
yciA	0.5902	COG1607	P0A8Z0	I	3	Acyl-CoA hydrolase
dedA	0.5847	COG0586	P0ABP6	S	2	Membrane integrity protein DedA, putative transporter, DedA/Tvp38 family
yfiC	0.5574	COG4123	P31825	S	3	tRNA1(Val) A37 N6-methylase TrmN6
pqqL	0.5410	COG0612	P31828	S	3	Predicted Zn-dependent peptidase, M16 family
yhgF	0.5355	COG2183	P46837	K	2	Transcriptional accessory protein Tex/SPT6
ycgM	0.5355	COG0179	P76004	Q	2	2-keto-4-pentenoate hydratase/2-oxohepta-3-ene-1,7-dioic acid hydratase (catechol pathway)
yicG	0.5355	COG2860	P0AGM2	S	2	Unchar. membrane protein
yicC	0.5301	COG1561	P23839	S	2	Unchar. stationary-phase protein YicC, UPF0701
ydjA	0.5301	COG0778	P0ACY1	C	3	Nitroreductase
paaD	0.5301	COG2151	P76080	S	3	Metal-sulfur cluster biosynthetic enzyme
yohJ	0.5246	COG1380	P60632	S	2	Putative effector of murein hydrolase LrgA, UPF0299
yeaQ	0.5191	COG2261	P64485	S	2	Unchar. membrane protein YeaQ/YmgE, transglycosylase-associated protein family
msrC	0.5191	COG1956	P76270	T	3	GAF domain-containing protein, putative methionine-R-sulfoxide reductase
cdsA	0.5191	COG4589	O31752*	S	3	Predicted CDP-diglyceride synthetase/phosphatidate cytidylyltransferase
yqgE	0.5137	COG1678	P0A8W5	K	2	Putative transcriptional regulator, AlgH/UPF0301
ybcJ	0.5027	COG2501	P0AAS7	S	3	Ribosome-associated protein YbcJ, S4-like RNA-binding protein

*Best match was to *B. subiltis* gene shown here, nearest *E. coli* gene is *ynbB* with UniProt ID P76091.

Table D.2. Translation-associated orthogroups from three systems only present in the LBCA core genome when reconstructed with gain/loss penalties (g) above 1.0.

orthogroup	g	system	gene	comments
COG1825	1.05	Ribosome 50S subunit	*rplY* (L25)	Missing across several bacterial phylogenetic classes [18]
COG0267	1.05	Ribosome 50S subunit	*rpmG* (L33)	Missing in several bacterial species, compensatory paralogs such as *rpmGB* have been observed [18]
COG0257	1.05	Ribosome 50S subunit	*rpmJ* (L36)	Missing in several bacterial species, compensatory paralogs such as *ykgO* have been observed [18]
COG1358	>2.00	Ribosome 50S subunit	*rpl7Ae* (L7ae)	Essential protein in archaea only, homologs observed in only a limited set of bacteria [19]
COG0215	1.05	Aminoacyl-tRNA synthetases	*cysS* (CysRS)	Missing in some methanogenic archaea, Cys-tRNA generated through an o-phosphoseryl-tRNA intermediate via SepRS [20]
COG0017	1.25	Aminoacyl-tRNA synthetases	*asnS* (AsnRS)	Missing in several bacteria and archaea, Asn-tRNA generated by modifying misacylated Asp-tRNA-(Asn) via GatABC [15]
COG0143	1.55	Aminoacyl-tRNA synthetases	*metG* (MetRS)	Complex evolutionary history, may result in incomplete coverage by a single orthogroup [21]
COG0480	1.05	Translation factors	*fusA* (EF-G)	EF-G subfamilies are highly diverse, may result in incomplete coverage by a single orthogroup [22]
COG4108	1.60	Translation factors	*prfC* (RF-3)	Predicted to have emerged post-LBCA as an offshoot of EF-G [23]

D.5 Supplementary Datasets

Dataset D.1. Genomes selected for LBCA analysis, including GTDB phylogenetic classifications and assembly quality metrics.

Dataset D.2. Species-level phylogenetic tree used for LBCA reconstruction. Includes labels for each species' representative node.

Dataset D.3. Results of benchmarking experiments for estimating true gene frequencies and genome-specific gene recovery rates from pangenomes.

Dataset D.4. Frequencies and annotations of highly conserved core orthogroups, LBCA core genome orthogroups, and orthogroups annotated within JCVI-Syn3A.

Dataset D.5. Mappings between COG and KEGG orthogroups, active KEGG modules in the LBCA core genome, and metabolite abbreviations for LBCA metabolic pathways.

D.6 References

[1] Qiyun Zhu, Uyen Mai, Wayne Pfeiffer, Stefan Janssen, Francesco Asnicar, Jon G Sanders, Pedro Belda-Ferre, Gabriel A Al-Ghalith, Evguenia Kopylova, Daniel McDonald, Tomasz Kosciolek, John B Yin, Shi Huang, Nimaichand Salam, Jian-Yu Jiao, Zijun Wu, Zhenjiang Z Xu, Kalen Cantrell, Yimeng Yang, Erfan Sayyari, Maryam Rabiee, James T Morton, Sheila Podell, Dan Knights, Wen-Jun Li, Curtis Huttenhower, Nicola Segata, Larry Smarr, Siavash Mirarab, and Rob Knight. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nat. Commun.*, 10(1):5477, December 2019.

[2] Donovan H Parks, Maria Chuvochina, David W Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004, November 2018.

[3] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055, July 2015.

[4] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, March 2010.

[5] Torsten Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069, July 2014.

[6] Jason C Hyun, Jonathan M Monk, and Bernhard O Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022.

[7] W Li, L Jaroszewski, and A Godzik. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283, March 2001.

[8] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods*, 17(3):261–272, March 2020.

[9] Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, 38(12):5825–5829, December 2021.

[10] Michael Y Galperin, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, 49(D1):D274–D281, January 2021.

[11] UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, 49(D1):D480–D489, January 2021.

[12] Miklós Csurös. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912, August 2010.

[13] Miklós Csűrös. Ancestral reconstruction by asymmetric wagner parsimony over continuous characters and squared parsimony over distributions. In *Comparative Genomics*, Lecture notes in computer science, pages 72–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[14] Natalya Yutin, Pere Puigbò, Eugene V Koonin, and Yuri I Wolf. Phylogenomics of prokaryotic ribosomal proteins. *PLoS One*, 7(5):e36972, May 2012.

[15] Miguel Angel Rubio Gomez and Michael Ibba. Aminoacyl-tRNA synthetases. *RNA*, 26(8):910–936, August 2020.

[16] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000.

[17] Eugene V Koonin and Yuri I Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.*, 36(21):6688–6719, December 2008.

[18] Michael Y Galperin, Yuri I Wolf, Sofya K Garushyants, Roberto Vera Alvarez, and Eugene V Koonin. Nonessential ribosomal proteins in bacteria and archaea identified using clusters of orthologous genes. *J. Bacteriol.*, 203(11), May 2021.

[19] Nathan J Baird, Jinwei Zhang, Tomoko Hamma, and Adrian R Ferré-D'Amaré. YbxF and YlxQ are bacterial homologs of L7Ae and bind k-turns but not k-loops. *RNA*, 18(4):759–770, April 2012.

[20] C S Hamann, K R Sowers, R S Lipman, and Y M Hou. An archaeal aminoacyl-tRNA synthetase missing from genomic analysis. *J. Bacteriol.*, 181(18):5880–5884, September 1999.

[21] Y I Wolf, L Aravind, N V Grishin, and E V Koonin. Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals

a complex history of horizontal gene transfer events. *Genome Res.*, 9(8):689–710, August 1999.

[22] Tõnu Margus, Maido Remm, and Tanel Tenson. A computational study of elongation factor G (EFG) duplicated genes: diverged nature underlying the innovation on the same structural template. *PLoS One*, 6(8):e22789, August 2011.

[23] A Maxwell Burroughs and L Aravind. The origin and evolution of release factors: Implications for translation termination, ribosome rescue, and quality control pathways. *Int. J. Mol. Sci.*, 20(8), April 2019.