Epidemiological inference from pathogen genome data

By

DJ DARWIN BANDOY

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Integrative Pathobiology
in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA
DAVIS

Approved:

Bart C. Weimer, Chair

Michael Wilkes, co-chair

C. Titus Brown

Renee Tsolis

Committee in Charge
2022

Epidemiological inference from pathogen genome data

by

DJ DARWIN BANDOY

Acknowledgements

Our Lord God for the guidance and mercy throughout my PhD studies.

My mentor Bart Weimer for having full trust in accepting me as a PhD student and allowing me to pursue my academic interests.

Lab Weimer current and past members, Carol, Nugget and Robin for providing me with a supportive environment.

Philippine California Advanced Research Institute, Republic of the Philippines provided me a tremendous opportunity in pursuing this PhD.

University of the Philippines Los Baños, College of Veterinary Medicine for allowing me to purse this endeavor.

The Graduate Group in Integrative Pathobiology, UC Davis particularly Erin Kent for the extraordinary service over and beyond and GGIP Batch 2017.

The Designated Emphasis in Biotechnology for broadening the horizons beyond my expectations.

USDA Food Safety and Inspection System for the generous support in *Salmonella* Dublin project.

My wife Shem for being always supportive of my pursuits academic and beyond. Thank you for your sacrifices and I love you.

My family in the Philippines, my parents Bituin and Cornelio and my sisters Ianna and Kay Anne who endured the difficulties and hardships of poverty but still managed to be happy.

Our brethren in MCGI Sacramento for the support and assistance while we were residing in Davis.

Abstract

The use of whole genome sequencing in infectious disease diagnostics generated an unprecedented amount and resolution of information. Large-scale sequencing of pathogens requires scalable methods in species identification, outbreak clustering, virulence phenotyping, antimicrobial resistance profiling, and epidemic modeling.

This dissertation presents a new approach in defining species membership using a pangenome framework explicitly applied to the whole genome sequences of the genus *Hungatella* which effectively identified a misclassified reference strain. Genomic epidynamics is a phylogenetic free approach in epidemiological inference, particularly the disease transmission parameter reproductive number (R). This approach offers a scalable process in elucidating heterogeneous transmission of genomic variants of SARS-CoV-2. Genomic epidynamics bridges pathogen population genomics and epidemic modeling. A genome-first approach to antimicrobial resistance definition combines automated machine learning rank resistance genes and phenotypic data thru genomic MICs. This approach was applied to a multidrug-resistant serotype of *Salmonella enterica subsp. enterica* serovar Dublin (*S.* Dublin). Machine learning-based approach to genome-wide association study revealed allelic variants of *porA* in *Campylobacter jejuni* leading to an abortive phenotype when the organism is invasive from the gut and resides in the reproductive system.

List of Tables

List of Figures

Contents

Introduction

Next-generation sequencing is transforming infectious disease diagnostics from a plethora of individual observational phenotypic tests to a complete unitary collection of phenotypic potential with an array of genes. Using this network of genes that make up a genome to examine phenotypic capability provides a method for causative agent inference coupled with suitable therapeutic options. This information can also be used to compare genome evolution and transmission globally with unprecedented resolution[1]. A pivotal dependency to make the transformation from phenotypic and biomarker schemes to the use of whole-genome sequencing (WGS) for diagnostics and epidemiology is foundational evidence that genomes accurately predict important features of pathogens from the strain to the global population. The expansion of pathogen WGS in infectious disease diagnostics must incorporate the massive variation between genomes of the same species and serotype[2]. The use of population genomics requires several shifts for a suitable and robust analysis that can be relied upon for accurate and fast infectious disease agent characterization.

Efforts to capture this diversity estimate that at least 500-1000 bacterial genomes are needed for inclusion in a diagnostic analysis independent of the rate of evolution[3-5]. Viral genome availability is rising at exponential rates with some viruses leading to an even more significant challenge than bacteria for the scale of data, but the low number of genes enables 10,000's of genomes to be compared if the infrastructure is in place. Likewise, existing phylogenetic methods for infectious disease are not designed to handle thousands of samples. In many cases, the underlying assumptions of relatedness are violated using phylogenomics and tree-based methods that are observed with the sometimes enormous genome diversity of organisms with the same name[6]. Consequently, as the number of genomes from the same species continues to grow, another consideration is required – computational capacity. The scalable computational challenge is outpacing the capacity of many diagnostic labs and local computation and beyond many clinical applications' reach. With new genomes also comes additional diversity in genes that were once thought to be directly causal but are now containing variants that are conditional on causing disease. Therefore, there is a need to bridge the gap for scalable analysis of pathogen genomes to elucidate the mechanistic basis of virulence, transmissibility, and antimicrobial resistance.

Within a species, or even a serotype, genomic diversity underpins the mechanistic basis of disease and virulence. For example, the genome distance, a pairwise measure of relatedness, between pathogenic and nonpathogenic *Escherichia coli* is estimated to be 36%. This microbial genomic diversity is best represented by the pangenome concept, consisting of the core genes, which are genes common to all isolates and accessory genes, genes found only in subgroups of the isolates. The species pangenome is shaped by a set of complex and diverse processes, including mutations, gene gain, and loss, genome reduction and rearrangement, and horizontal gene transfer. The accessory component of the pangenome, genomic islands, is also directly linked with disease, virulence, antimicrobial resistance, and metabolic functions. A common approach to define microbial species diversity is through population genetics.

Population genetics is the study of the inheritance of a gene at the scale of the population. In the context of infectious disease investigation, samples from an outbreak are collected and sequenced, and a phylogenetic tree is generated to define clusters as inference of transmission. However, this assumption is misleading as a phylogenetic tree is not a transmission tree but rather an estimation of the evolutionary relatedness based on mutation and selection. A phylogenetic tree generated from a densely sampled outbreak is not directly able to identify "who infected whom." While a phylogenetic tree can identify relationships between genomes, it is restricted to information about evolutionary relatedness. It is a poor indicator of events between individuals unless the genome diversity is coupled to additional genealogical or other information. This gap in determining disease transmission parameters can be addressed using an epidemiology first approach in analyzing pathogen WGS. While epidemiological methods are inherently designed and optimized for inferring population-level disease parameters, such methods are relatively underutilized in the genomics field[7]. Applying epidemiological methods to large-scale populations of pathogen WGS enables a quantitative approach in determining pathogen characteristics related to disease transmission, virulence, and antimicrobial resistance. This strategy allows evaluation of the impact of pathogen population genomic variation for disease transmission. A novel approach is needed to link genome content, genetic variation, and distribution with the population scale characteristics of organisms found globally and locally. This approach allows for multiple types of evolution, such as parallel and convergent selection but without epidemiological information these concepts are not valuable to determine transmission. However, they are helpful in determining the genomic evolution of pathogens to understand the disease dynamics of zoonotic pathogens.

A recent development using the epidemiological approach in determining the role of pathogen genome variation in disease and virulence is the application of bacterial genome-wide association study (GWAS)[8]. The conceptual basis of the GWAS approach in bacterial virulence discovery is the presence of common mutational variants at higher frequencies across isolates with a specific phenotype or disease[9]. The variants in the isolates without the phenotype is expected to be low or close to zero. Hence, it is possible to determine the statistical association of the mutational variants to the corresponding phenotype out of the thousands of genomic variations that directly link the population genome variation to disease transmission and persistence. Consequently, the use of bacterial GWAS for virulence gene discovery has increased in recent years[9]. However, one distinguishing element of bacterial GWAS from human GWAS is the challenge of defining an outgroup and conditional settings other for the specific question at hand. This is particularly challenging because bacteria evolve or mutate so quickly, especially compared to mammalian species that take years to fix a new mutation. Bacteria can mutate within minutes and grow in conditions that direct the evolution to 'guide' mutations that result in a new dimension for conditional alleles linked to disease but are not informative of parentage. This is compounded by horizontal gene transfer and genome rearrangements that do not occur quickly in other organisms (Table 1).

Table 1. Heritable mutation time scale among domains of life

Form of Life	Genome size (bases)	Mutation time scale	DNA source	Detailed genealogical record
Humans (mammals)	3,100 Mb	~25 years	mixture from parents to distinctly indicate parentage	Common
Livestock (cattle)	3,000 Mb	~2 years	mixture from parents to distinctly indicate parentage	Common
Plants	~5,500 Mb	Months	mixture from parents to distinctly indicate parentage	Common
Bacteria	1.8 to 8.5 Mb	Minutes to hours	Previous generation with mutations each generation	Rare to never
Viruses	~30,000 bases	Minutes	Previous generation with mutations each generation	Rare to never

For human GWAS, healthy individuals usually serve as the outgroup or controls. While certain bacterial species like *Staphylococcus aureus*, *Streptococcus pneumoniae* and *Streptococcus agalactiae* have asymptomatic carriers; therefore, can be used as outgroups, other species do not have a well delineated counterpart[10-12]. Hence, the alternative option employed to use differential virulence phenotype, such as the one used in *Helicobacter pylori* GWAS for gastric cancer with the gastritis as the outgroup phenotype[13]. Another complicated feature of bacterial

GWAS is the level of diversity within the population. These diversity results in accessory genes that are present in some isolates but not in all isolates[2]. This presents a challenge in selecting a reference species to identify the difference between bacterial sequences. Hence, alternative approaches to define genome difference via k-mers and the pangenome that are reference free enable a new method to assess relatedness. This also serves as the justification to perform a population wide approach of analysis in bacterial genomics as the presence of accessory genes are variable and cannot be applied to the entire population.

Bacterial diversity also complicates the statistical frameworks of analysis. Bacterial GWAS employs several statistical methods such as linear mixed models, Fisher exact test, and Spearman rank test[9]. While a Bonferroni correction has been used to account for multiple comparisons, this generates a relatively high level of exclusion criteria, potentially eliminating other causative variants. Another more pressing concern is the lack of ranking of the GWAS hits with the existing GWAS frameworks. Hence robust statistical frameworks are needed to rank genes associated with specific phenotypes in bacterial GWAS. On one hand, attempts have been made to model population genetic measures of selection as the process generating the GWAS hits in human studies using Fisher's geometric model[14,15]. This model explores the effect size of fitness with mutations. Ultimately, alterations in fitness by mutations will result to increase in the phenotype frequencies that will be amenable to statistical analysis. While such analytical approaches are common in microbial evolutionary experiments, they are relatively unused in microbial GWAS studies [16].

The complexities of delineating bacterial phenotypes extend beyond virulence. The prediction of bacterial antimicrobial resistance (AMR) from genome sequences suffers a similar condition due bacterial diversity, limitations in databases, and sparsity of sequence of many bacterial species[17]. A typical workflow for antimicrobial resistance prediction from sequence data compares known resistance mechanisms in curated databases[18,19]. Known resistance genes are compared and scored with the counterpart genes in the isolates. While such methods are robust for well sequenced organism and well characterized mechanisms, bacterial diversity is not static. Hence, evolutionary forces generate novel mechanisms of resistance that cannot be handled by existing databases[20]. These generate different forms of discordance[21]. A false positive indicate presence of resistance genes without the resistance phenotype and false negative shows absence of resistance genes but manifests the resistance phenotype. This limits the direct utilization of genomes for AMR prediction.

To address the limitations of the similarity index approach in predicting AMR from sequences, a population-based approach combining epidemiological techniques and machine learning has been applied to several bacterial species (Table 1). The complexity of bacterial diversity is well suited for large data analytical techniques such as machine learning. For instance, Kavvas *et al.* discovered novel mechanisms of resistance in *Mycobacterium*, one of the relatively slow evolving bacterial species using 1595 samples support vector machines[22]. Predictably, machine learning will be more valuable as an approach for species that are more rapidly evolving and hence more genetically diverse than *Mycobacterium*. Hence machine learning approaches to AMR prediction have been employed for the following bacterial species: *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Escherichia coli*, *Salmonella* and *Klebsiella* as listed in Table 1[23-27]. Most studies used support vector machine followed by Xgboost and neural networks[28]. There are also differences with the unit of analysis ranging from SNPs, k-mers to pangenome wide gene presence and absence reflecting a similar pattern of feature extraction with bacterial GWAS[25,27,29].

While machine learning approach offers a path to resolve some issues with similarity index-based assignment of resistance, fundamental issues remain-particularly genotype-phenotype discordance[30]. These discordances are designated as either very major errors (false antibiotic susceptibility) and major errors (false resistance). The errors arise from the conflicting presence of resistance genes but absence of phenotypic resistance and vice versa. There is also no universally accepted definition of resistance via minimum inhibitory concentration (MIC). The arbitrary cutoffs compound the difficulty of defining a clear resistance phenotype. There are also several antibiotics such as azithromycin which lack a well-defined MIC cutoff. MIC cutoffs are also defined without referencing the presence of resistance genes, an artefact of pre-genomic era.

Table 2. Machine learning studies in genome-based prediction of AMR

Reference	Number of samples	Unit of analysis	Machine learning Algorithm	Organism
Kavvas *et al.* (2018)	1595	SNPs	support vector machine	*Mycobacterium*
Hyun *et al.* (2020)	288 (SA) 456 (PA) 1588 (EC)	Pangenome gene presence absence, SNPs	support vector machine	*Staphylococcus aureus,* *Pseudomonas* *Escherichia coli*
Nguyen (2018)	5278	10 nucleotide units (10-mers)	Xgboost	*Salmonella*
Macesic (2020)	600 <	K-mers and SNPs	Multiple (Random forest, logistic regression)	*Klebsiella pneumoniae*
Avershina (2021)	171	K-mers	Neural networks	*E. coli and Klebsiella*
Jaillard (2021)	1665	K-mers	Support vector machine	*Klebsiella pneumoniae*
Her (2018)	59	Pangenome gene presence absence	Multiple (Support Vector Machine, Naive Bayes, Random Forest	*E. coli*
Liu (2020)	96	K-mers	Support Vector Machine and Set Covering Machine	*Actinobacillus pleuropneumoniae*

This dissertation formulated epidemiological inference from pathogen sequence data. The first two parts cover a cross-sectional approach (bacterial GWAS in virulence and AMR) and the last part focuses on time-series methods like epidemic curves. There is currently no direct disease transmission inference from pathogen genome data to place this in context. Disease transmission inference from sequences is heuristically estimated by phylodynamic methods, which combine phylogenetics and a demographic model. The conceptual basis of phylodynamics is based on the ability of the topology of phylogenies to represent immunological and evolutionary processes[31]. Extrapolation of transmission from phylogenies also requires similarity between the rate of pathogen evolution and infection spread. Hence, RNA viruses with relatively small genomes rapidly mutate, generating signals of disease transmission[32]. Aside from similarities in timescales between pathogen evolution and spread, the shape and features of phylogenies still need to be related to the disease transmission process using coalescent-based methods or birth-death methods. Coalescent-based methods link demographic values such as effective population size to the estimates of coalescence times of the phylogeny. Birth-death methods models the features of phylogenies: branches as birth and leaves as deaths, allowing a quantitative transformation. However, phylodynamic approaches do not have a framework to integrate temporal data, cannot handle large datasets due to computational complexity, and propagates uncertainties in multistep manner.

Table 3. Disease transmission inference from pathogen WGS

Publication	Unit of analysis	Epidemiological scale	Parameters Estimated	Transmission inference	Sample size of genomes	Disease
Davies *et al*. (2021)	Lineage	Population	R	Phylogenetic and statistical model	150,000	COVID-19
Pybus *et al*. (2009)	Lineage	Population	Effective population size, R for each lineage	Phylogenetic and demographic model	300 <	HCV
Zhao *et al*. (2020)	Nucleotide set	Population	Prevalence of nucleotide marker	None	47000	COVID-19
De Maio *et al*. (2017)	SNPS (genomic variants)	Infection clusters	Transmission cluster	Phylogenetic and shared variant clustering algorithm	62	Ebola 2014 outbreak
Alamil *et al*. (2019)	Genome distance	Infection clusters	Transmission pairs	Pseudo evolutionary model using distance function	21	Swine influenza virus
Jombart *et al*. (2014)	Genome distance	Transmission tree	R	Bayesian inference of transmission tree; sequence evolution model	13	SARS 2003 outbreak Singapore
Campbell *et al*. (2018)	Genome distance	Infection cluster	Transmission cluster	Transmission divergence	63-62 cases	Multiple bacterial & viral outbreaks

Stimson et al. (2019)	Genome distance	Transmission cluster	SNP threshold	Genome distance and sampling time	32	Tuberculosis
Ypma *et al.* (2012)	Genetic distance	Transmission cluster	Transmission cluster	Transmission likelihood integrated with spatiotemporal data	200 <	Avian influenza
Worby *et al.* (2014)	Pairwise genetic distance	Infection cluster	Transmission chain	geometric-Poisson distribution	35	Staphylococcus aureus

Aside from phylodynamic approaches, alternative methods have been explored to determine disease transmission (Table 2)[33-41]. The techniques utilized several scales of the genome, from mutations to genome distance. Infection clusters have been defined using shared variants. The assumption here is that sequences sharing common mutations are treated as evidence of epidemiological linkage. Multiple studies used genome distance, a metric of difference between two genomes as a tool for epidemiological linkage between clusters based on the entire genome, rather than marker genes or a small subset of gene clusters. Aside from differences in how sequences are used to establish epidemiological linkage, alternative methods also directly integrate temporal and spatial data to increase the resolution of transmission clusters. However, fundamental limitations exist as most of the methods enumerated here rely on limited samples (13-200) and only infer transmission clusters without providing a framework to estimate population-level metrics such as reproductive number.

This dissertation addresses the notable gaps in population-level inference from pathogen WGS. We used machine learning approaches in bacterial GWAS to identify alleles associated with abortive phenotype of *Campylobacter jejuni* and tested the hypothesis that genomic variants of *Campylobacter jejuni* drive variable virulence. Current methods utilized in comparative genomics, such as pairwise comparison between genomes, failed to identify causative genes or alleles out of the 8000 allelic differences between an abortive isolate and a laboratory strain. We applied a machine learning method for bacterial GWAS to address other significant limitations in virulence discovery methods associated with multiple testing of individual loci that lack the framework of interrogating the interaction of various genes or allelic variants. Furthermore, a method to rank the generated GWAS hits using classical statistical methods is absent in the literature. Statistical assumptions like the independence of units being tested like gene or allele within the genome are conceptually flawed in bacteria considering the operon configuration and horizontal gene transfer (e.g., plasmids). Hence, alternative biological and statistically compatible analyses need to be defined for bacterial population genomics.

We developed a population genomics approach combined with machine learning in

addressing gaps in predicting antimicrobial resistance using WGS in multidrug-resistant cattle adapted *Salmonella enterica subsp. enterica* serovar Dublin *(S*. Dublin*)*. We tested the hypothesis that genomic variants of *Salmonella* Dublin drive antimicrobial resistance. A critical gap in predicting resistance phenotype from pathogen WGS is the lack of correlation between genomic resistance mechanisms and resistance phenotype as defined by epidemiological cutoff values (ECVs). ECVs are antibacterial susceptibility values that distinguish wild-type bacteria from non-wild-type populations using differences in MIC (minimum inhibitory concentration). While resistance mechanisms are included in the definition of the MICs, resistance gene presence is not integrated in setting ECVs for antibiotics. Given the high resolution of WGS to identify resistance phenotypes by linking genomic mechanisms, we developed a population genomics approach to use WGS to associate resistance mechanisms with specific MICs. We also applied machine learning in WGS of *S*. Dublin to address the gaps in using a database approach to identify antimicrobial resistance, particularly the absence of framework to discover novel resistance mechanisms and susceptibility.

Genomic epidynamics addresses the gaps in disease inference using population WGS. We developed genomic epidynamics as an epidemiology first approach in measuring disease transmission of SARS-CoV-2 variants. Most existing methods to analyze viral WGS depend on associating viral evolution dynamics and disease transmission via phylogenetic analysis[33,42]. However, phylogenetics and its derivative method, phylodynamics are inadequate to handle the large volume of SARS-CoV-2 sequences with the COVID-19 pandemic. Consequently, viral lineage estimation becomes cumbersome with large sets of WGS that unfortunately leads to crude data reduction techniques like downsampling, ultimately defeating the public health investment in large-scale sequencing. Moreover, there is no universally accepted definition of a viral lineage which resulted in multiple competing classification schemes[43]. Unfortunately, it is computationally impossible to create phylogenetic trees with hundreds of thousands of genomes for meaningful epidemiological analysis[43]. Genomic epidynamics addresses the gaps in disease transmission inference by bypassing the computationally prohibitive method of phylogenetic construction between SARS-CoV-2 WGS. Using this approach, we tested the hypothesis that the genomic variants of SAR-COV-2 drive the transmissibility of COVID-19.

Thesis structure

The overarching theme in this dissertation is the development of epidemiological and machine learning approaches in pathogen population genomics to define antimicrobial resistance, virulence and disease transmission dynamics in the population.

The remainder of the dissertation is organized as follows:

Chapter 1. Biological machine learning combined with Campylobacter population genomics reveals virulence gene allelic variants cause disease (published in Microorganisms, 2020).

Chapter 2. Automated machine learning and genomic MICs as a framework for antimicrobial resistance prediction in whole genome sequences of *Salmonella enterica subsp. enterica* serovar Dublin.

Chapter 3. Analysis of SARS-CoV-2 genomic epidemiology reveals disease transmission coupled to variant emergence and allelic variation (published in Scientific Reports, 2021)

Chapter 4. Introduction of genomic epidynamics as an approach to determine disease transmission dynamics of SARS-CoV-2 variants.

Chapter 5. Demonstrates the pangenome-based species definition and clustering for bacterial population genome analysis in *Hungatella hathewayi* (preprinted in medRxiv,2020; components published in Virulence, 2021).

Chapter 1. Biological machine learning combined with *Campylobacter* population genomics reveals virulence gene allelic variants cause disease[1]

Introduction

Comparative microbial genomics has emerged from pangenome comparisons that are exclusively tied to reference genomes that define the perspective of change to a core and flexible genome perspective lacking a firm confirmation of which genes are linked to disease[44]. An alternative approach to this perspective is use of genome wide association (GWAS) methods that are common in mammalian genomics in an effort to refine the estimates of specific genes of interest. A limitation of GWAS is that it sequentially examines single loci that prevents simultaneous analysis of different allelic variants that can be interacting at different levels and population distribution between strain differentation[45]. This is a severe limitation in bacterial genomics, especially as bacterial population genomics is now possible at a scale that allows examination of non-linear and specific selective conditions for evolutionary rates of each gene and all of the alleles found in very large populations that create a big data analytical problem. A compounding limitation is the lack of appropriate statistical models that underpin this approach in bacteria since it is unknown when the populations are normally distributed or evolving in a non-linear progression. As with all large data sets, multiple comparisons require Bonferroni correction to adjust the *p*-value based on a new scale as compared to gene

[1] Bandoy, D. D. R. and B. C. Weimer (2020). "Biological Machine Learning Combined with Campylobacter Population Genomics Reveals Virulence Gene Allelic Variants Cause Disease." Microorganisms **8**(4): 549.

expression but it is on a scale that is beyond that contemplated for gene expression variation (Table 1)[46]. Further, the assumption that each gene or allele is independent is conceptually flawed in bacteria considering the operon configuration and plasmids. Hence, alternative analyses that are biological and statistically compatible need to be defined and is tractable using machine learning.

Coupling GWAS, population microbial genomics, and machine learning is poised to be a robust alternative to classical GWAS or pangenome comparison to simultaneously discover changes in microbial genomes, and genes, that span the scale of genome plasticity to alleles of a single gene. Moreover, this combination (coined here as bioML) produces a statistically underpinned comparative importance ranking for each gene and allele that are not determined from GWAS alone. These advantages combined with downstream inspection of the prioritized rankings further powers discovery to bring biologically insightful observations and solutions, especially when large genome populations are used in the analysis, from very divergent populations of alleles that are missed when the sequence is too divergent for gene calling.

An analytical strength for use of machine learning in microbiology is the ability to define functional relationships from population scale genome comparisons or genes without a priori definition of the underlying mechanism of change or specific phenotype limitations[47]. This distinctive advantage makes machine learning superior to classical statistical tests for prokaryotic systems that are highly variable, particularly bacteria wherein explanatory variables are not linearly correlated, features are dependent due to genome dislinkage, varying evolutionary rates of between genes within the same genome, and assumptions of normal distribution are violated in part due to varying selection conditions[45,48]. These biological conditions and parameters are incompatible with the assumptions of linear or correlative statistics, which is compounded with data reduction methods that provide a very small snapshot of the genome variation that yield associations that have low predictive value with highly variable genomes[49-52], such as bacteria.

We proved the concept of coupling GWAS with machine learning and population bacterial genomics (Figure 1) using previously published verified alleles of a gene that causes abortion in livestock[53-55]. We hypothesized that a specific allele of a single gene (i.e., *porA*) is linked to extraintestinal invasion and further is causative in abortion. This was done using a wet lab validated data set containing 100 genomes[53-55] combined with extreme gradient boosting (XGboost) that was previously used in biological applications[56]. In a Finnish study XGboost identified genetic variants in a human GWAS sample that integrated complex nonlinear interactions of SNPs[57]. The ability to interrogate the predictive features enables whiteboxing of parameters, which is emerging as a tool for deriving mechanistic function in biology[58]. XGboost implements adaptive optimization within the functional space by iteration of the weak learners into strong learners represented by decision trees where each new decision tree is generated by factoring the residuals generated from the difference from observed to the predicted feature (Figure 2; Supplemental Table 1).

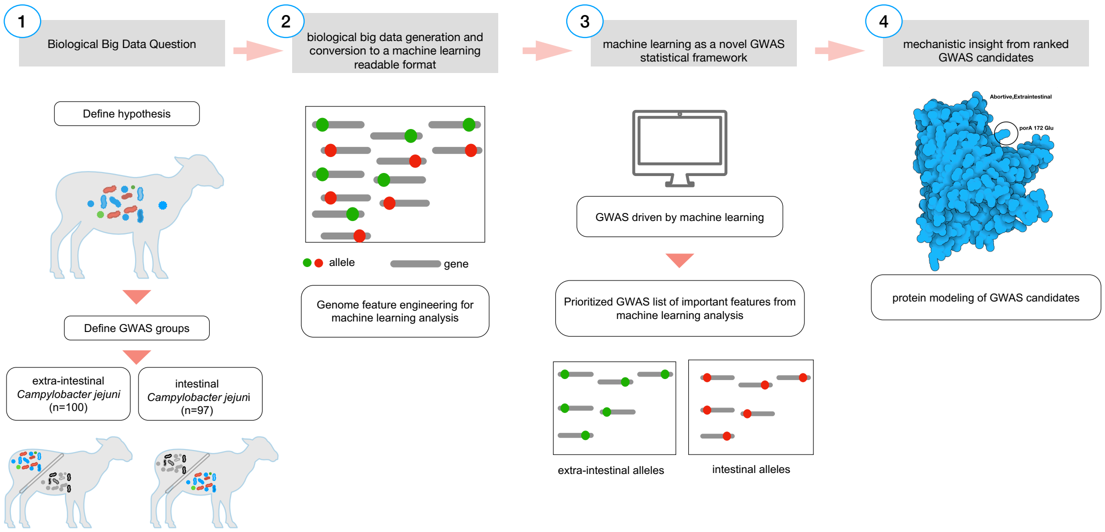

1
Biological Big Data Question
Define hypothesis
Define GWAS groups
extra-intestinal
Campylobacter jejuni
(n=100)
intestinal
Campylobacter jejuni
(n=97)
2
biological big data generation and conversion to a machine learning readable format
allele
gene
Genome feature engineering for machine learning analysis
3
machine learning as a novel GWAS statistical framework
GWAS driven by machine learning
Prioritized GWAS list of important features from machine learning analysis
extra-intestinal alleles
intestinal alleles
4
mechanistic insight from ranked GWAS candidates
Abortive,Extraintestinal
porA 172 Glu
protein modeling of GWAS candidates

Figure 1. Biological feature engineering of genomic data for machine learning analysis. A critical step in feature engineering is selection of the appropriate comparison groups to enable classification of alleles that are related to the specific phenotype of interest (i.e., intestinal (controls; diarrheal; n=108) and extraintestinal (cases; abortive; n=85) (Step 1). Population-wide allelic variants (red dot = intestinal, green dot = extraintestinal) that result from variant calling (Step 2) and are used as the input features for machine learning analysis (Step 3). The predicted model generated from the machine learning analysis is inspected for the most predictive features using biological context, input, and protein modelling (Step 4) that represents a nonsynonymous mutation from the genomic the population of allelic variants (n=193).

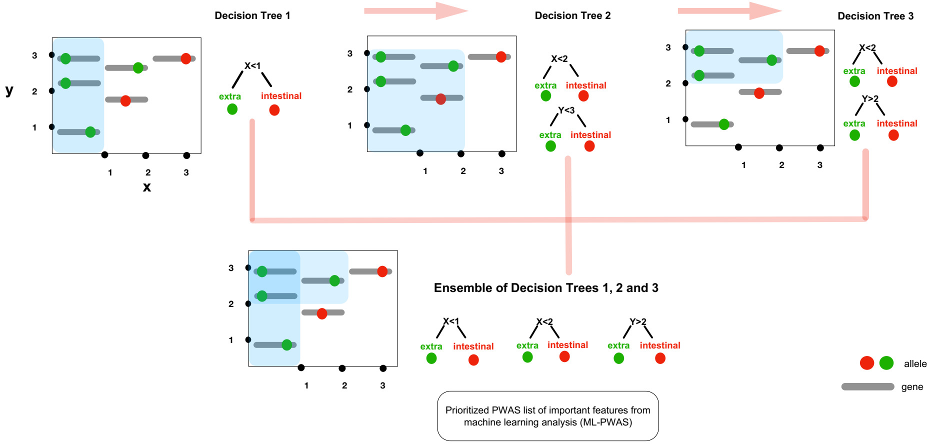


Figure 2. The conceptual framework diagram depicting machine learning in bacterial genome wide association using extreme gradient boosting (XGboost). Boosting is a technique of combining a set of weak classifiers or decision trees to increase prediction accuracy. Red dots represent an allelic variant, each grey bar represents a unique allele. Individual decision trees (1, 2, 3) fail to fully capture the allelic variants associated with the phenotype (e.g., extraintestinal abortion), but by combining the trees together results in a process called as boosting increases the discriminative power.

Results

BioML analysis identified 14 *porA* loci as the most important alleles ranging from 1.0 to 0.65 scaled importance out of the 1.2 million SNPs (Supplemental Table 1). These ranked loci were compared by body location (Figure 3), which further clarified the location of these SNPs and indels that simultaneously presented the ranked associated allelic variants within the phenotype of interest as detected with bioML as well as the non-associated alleles. This analysis procedure detected various forms of porA from more abundant versions to hybrid variation and rare variants that were not captured by gene calling (allele divergence was too great), machine learning alone, or classical statistical testing. Regions within *porA* from the cases expressing different allelic patterns were further explored for each genome and implications in biological features important in the disease. Protein structures were modeled to examine the changes in protein configuration initially yielded four distinct groups (Figure 3.) that ranged from non-abortive to variations of proteins all of which caused abortion. These alleles were directly compared to those validated *in vivo* and found to be linked to specific protein loops within alleles verified previously[53-55] – in all cases bioML found each of those to be biologically important for abortion and found new hybrid versions of the protein that were previously unrecognized.

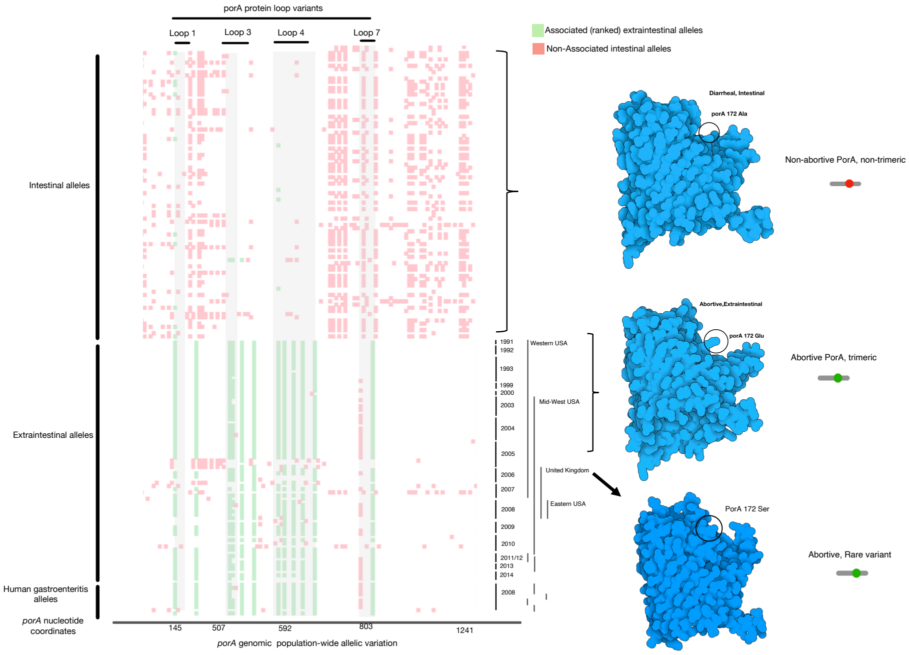
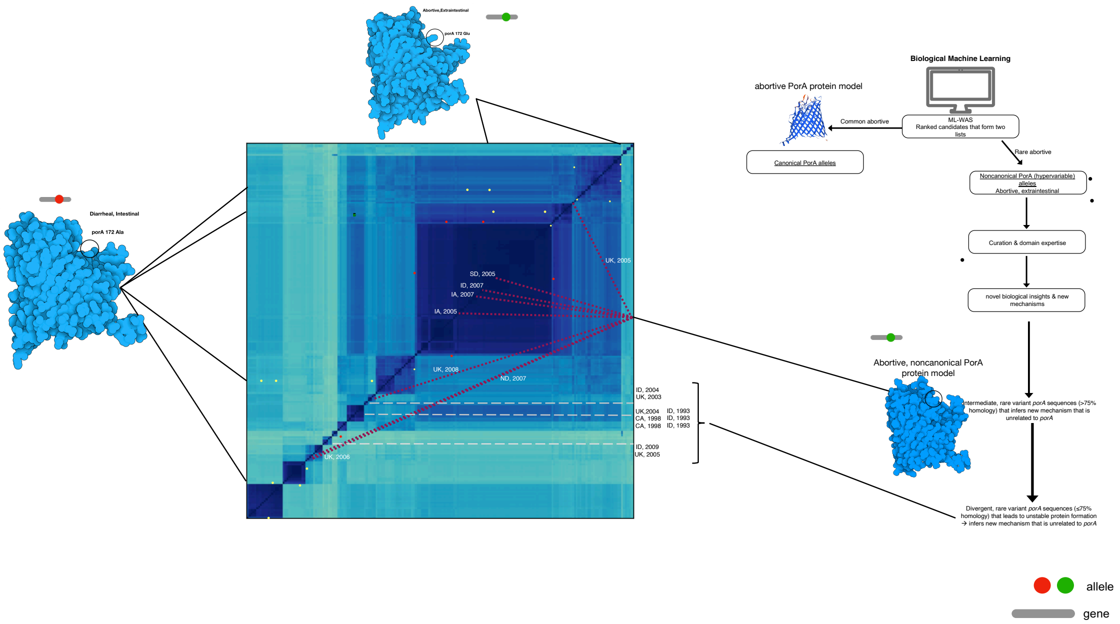


Figure 3. Comparative plot of SNP loci along the *porA* gene in all genomes. We termed this a Tetris plot as an alternative visualization of genome wide association hits because they are ranked and display only the loci that vary to produce a nonsynonymous mutation. The y-axis contains individual genomes from the cases and the controls, while the x-axis contains the GWAS SNP loci (green), the non-disease associated SNPs (red), open space (white) are loci that are identical in the gene sequence. Temporal and geographic metadata on the right side of the Tetris plot provides context for mutational enrichment over 30 years and multiple distant locations in North America and the UK. The enriched SNP variation produced different protein structures (far right in blue) as the corresponding protein model by location within the animal by SNP. Protein structural features corresponding to the ranked GWAS variants are

annotated on top and below the plot are the nucleotide coordinates. Rare variants (homology <75%) was not included by the variant caller in this visualization but manual inspection provided a method to find these variants.



Abortive,Extraintestinal
porA 172 Glu
Diarrheal, Intestinal
porA 172 Ala
UK, 2005
SD, 2005
ID, 2007
IA, 2007
IA, 2005
UK, 2008
ND, 2007
UK, 2006
ID, 2004
UK, 2003
UK,2004
CA, 1998
CA, 1998
ID, 1993
ID, 1993
ID, 1993
ID, 2009
UK, 2005
Biological Machine Learning
abortive PorA protein model
Common abortive
ML-WAS
Ranked candidates that form two lists
Canonical PorA alleles
Rare abortive
Noncanonical PorA (hypervariable) alleles
Abortive, extraintestinal
Curation & domain expertise
novel biological insights & new mechanisms
Abortive, noncanonical PorA protein model
Intermediate, rare variant porA sequences (>75% homology) that infers new mechanism that is unrelated to porA
Divergent, rare variant porA sequences (≤75% homology) that leads to unstable protein formation → infers new mechanism that is unrelated to porA
allele
gene

Figure 4. Whole genome distance matrix using minhash depicting an all against all comparison of genome diversity for all isolates used in this study overlaid with the porA variant associated with body location and disease phenotype. Genotypes and *porA* variants are connected in this depiction to examine the association between intestinal/diarrheal location (yellow dot boxes), prototypical extraintestinal/abortive (red dot boxes), non-prototypical *porA* variants in extraintestinal/abortive (maroon lines), and rare *porA* variants in extraintestinal/abortive (grey dashed lines) were co-located to their respective genomes in the genotype map. For the non-prototypical variants, the year and location of isolation was included to depict the variation over time and space in the maintenance of a minority population of *porA* variants of extraintestinal abortive *Campylobacter jejuni*. The diagram to the right depicts the process used for this analysis.

Further verification of the approach found that each of the top ranked alleles were located in loops 1, 3, 4, 7 as enriched selection loci, which perfectly verifies the published wet lab observations[53-55]. By tracking every abortive genome, we found variants that were between 90% identical with >75% protein homology that were designated as nonprototypical variants because the sequence variation was high enough to change protein structures. In a limited set of alleles, the *porA* gene was so divergent that they were not variant called but were recovered with manual curation of the bioML output. Recovery of these genes that were not initially identified created a third group of rare variant alleles that also caused abortion (Figure 4; protein homology <75%). This result provides a foundation for functional variation of a core gene from all *Campylobacter* and further provides insight into the variability of *porA* as a virulence factor.

All of the variants were mapped to the whole genome phylogeny to determine if the alleles were co-evolving with the genome (Figure 4). While some of the alleles were associated with similar genomes most of the alleles were found in >2 genotypes. Prototypical allelic variants clustered in the largest genomic group of abortive isolates, as did some of the nonprototypical *porA* variants. However, there was significant genome variation and contained the two groups that caused abortion. Rare *porA* variants were distributed within different genomic groups as well as over a 15-year span between North America and the UK. The extensive allelic variation of *porA*, as well as the different genotypes, suggests that a genome surveillance system based on SNPs or a single gene would be unsuccessful to link these genomes to a disease. In combination, these observations indicate that bioML produced a ranked list of biologically important alleles that were validated with those that were previously shown to be causal in abortion for the exact SNP and the protein loop location. Together, these observations verified that bioML was capable of accurately identifying the exact SNPs in *porA* that cause abortion.

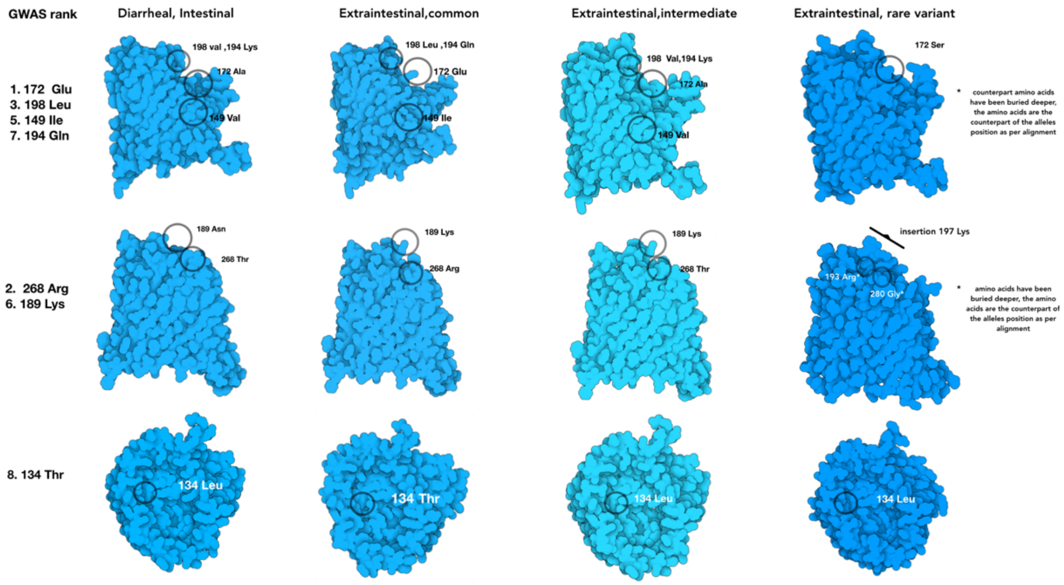


Figure 5. Protein models of the four groups of *porA* allelic variants that change the protein model structure relative to the isolate location in the host and the disease outcome. The amino acids corresponding with the bioML top ranked alleles are labelled in the common variant of *porA*, while the rest show the substituted amino acid in their respective position.

Discussion

In this study we used a previously validated wet lab data set with a tetracycline resistant strain of *Campylobacter jejuni* causing abortion in sheep[53-55] as the validation training set for bioML analysis. Previous wet lab studies used a pairwise comparison to identify 8,000 SNP difference between a reference genome and an abortive strain that subsequently utilized transformed genomes to identify specific allelic variants causing abortion. We included the validated 85 genomes that span 30 years and multiple locations as a reference set of cases and 108 control genomes of intestinal, diarrheal isolates. This approach permitted exploration of bacterial population genomic space by linking different phenotypes to validated genome variation (Figure 1). Biological feature engineering of this collection identified 1.2 million SNPs, which is not tractable using in vivo infection studies to determine the roll of all SNPs in disease. To examine this scale problem, we hypothesized that genomic changes evolved in gastrointestinal *C. jejuni* resulting in an abortive phenotype; hence, invading the intestine and progressing to other tissues – in this case the placenta resulting in abortion. Applying bioML analysis to the population of gastrointestinal, diarrheal *C. jejuni* versus extraintestinal, abortive phenotypes produced a prioritized set of alleles in a ranked order of importance to the phenotype (i.e. abortion) (Supplementary Table 1).

Since each bioML allele was validated for accuracy to wet lab results correctly, we broadened the examination of the protein changes from the ranked alleles to determine if the protein structure variation contained a specific feature or amino acid substitution that was linked to abortion (Figure 5). The first six top-ranked alleles contained various amino acid substitutions for each *porA* sequence and multiple PorA models. However, Lys_{189} was conserved among the extraintestinal variants and Asn was found in the intestinal alleles at the same position. Lys mutation changes are the most impactful in membrane pore structure and are one of the tenets of membrane topology as positive inside rule[59,60]. Positive inside rule describes the observation across membrane pores that positively charged amino acids are found within the cytoplasm and negatively charged amino acids are in the extracellular domain. Membrane topology can radically change from being oriented inside the membrane (exposed to the periplasm in this case) to outside the membrane with a single Lys mutation. Within the adjacent protein structure, Lys snorkeling effectively minimized the nonpolar chain component by burying in the hydrophobic domain and at the same time exposed the polar component to the aqueous domain is another single amino acid change that alters the topology of the membrane domain[61]. Bacterial membrane pore flipping could be a potential mechanism to avoid recognition by the immune system and enhance of ion transport for bacterial metabolism. In atypical (i.e., hypervariable alleles) this position is buried in a deeper position due to insertional mutation in rare variants, the inserted amino acids contain Lys at position 197, a new position as compared to the prototypical protein model. Additionally, insertions in the rare variants reduce the homology to <75% lead to more extensive protein structural changes that change the PorA arrangement in the membrane and were retained the ability to cause abortion. This situation is troublesome for traditional analytical approaches and would be missed completely using comparative genomics alone but bioML combined with biological tracing effectively identified this situation to successfully link multiple genotypes, protein models, and disease.

Conclusions

This study utilized a combination of GWAS, population bacterial genomics, and machine learning to identify and rank allelic variants that correspond to biologically validated alleles of porA to cause abortion. The bioML analysis were further supported by the longitudinal and spatial conservation of *porA* coupled to protein substitutions that led to biologically relevant changes in the structure to change activity. A Tetris plot visualization provided an avenue to discover divergent and rare variants that provided further insight with protein modelling that uncovered protein substitutions resulting in localization changes that affect activity and isolation localization in the host. Together these results demonstrate and validate a novel method, bioML, to discover biological variation combined with established mechanisms using population bacterial genomics. This approach provides an avenue to leverage the massive amount of bacterial genomic sequences to uncover new mechanisms of disease with potential to provide therapeutic approaches.

Methods

Biological feature engineering

Biological feature engineering entails selection of pertinent controls and cases for bioML analysis. The genomes between gastrointestinal and extraintestinal abortive isolates. *C. jejuni* controls were downloaded from Patric 3.5.28 (https://www.patricbrc.org/), June 1, 2019 (Supplemental Table 2). Abortive extraintestinal genomes *of C. jejuni* were obtained from the Sequence Read Archive (SRA; Supplemental Table 3)[55]. Fastq files were assembled using Shovill 1.0.4 (https://github.com/tseemann/shovill). Assembled files were annotated with Prokka (version 1.13.3)[62]. Variant calling was done with the reference sequence *C. jejuni* NTC11168 with Snippy 4.3.5 (https://github.com/tseemann/snippy) as previously described[63].

Gradient tree boosting as GWAS framework

GWAS variants generated from the biological feature engineering step were used as input for XGboost. The original source code for implementing gradient tree boosting is available at https://xgboost.readthedocs.io/. Confusion matrix were generated and used to assess the performance of the model (Supplemental Table 4). The relative importance of the predictive model was used as the GWAS hits

Tetris plot

Classical GWAS hits are displayed as the negative logarithm of the p-value in Manhattan plots, hence we formulated a novel visualization of the ranked alleles generated by the machine learning model to highlight the difference between approaches - we call this GWAS hit visualization a Tetris plot. We color coded the relative importance values of the associated alleles derived from the XGboost (green being associated and red being non-associated). The source genome is plotted on the y-axis and genomic coordinates on the x-axis overlaid with GWAS hits presence or absence matrix.

Population wide whole genome phylogeny

The genome distance metric was calculated using genome wide k-mer signatures to generate the population-wide phylogeny with a k-mer size of 31 scaled to 1000 with Sourmash[64]. The resulting genome wide k-mer distance was visualized as an all-against-all heatmap[64].

Protein Modelling

Assembled genomes were annotated using Prokka (V1.13.3) and PorA protein sequences were extracted for protein modelling using Swiss Model[65,66]. The most homologous protein was used as template for protein modelling. Illustrate (https://ccsb.scripps.edu/illustrate/) was used to generate the protein visualization of the alleles. Ranked bioML alleles identified by visual inspection of the Tetris plot, via the ranked variable importance were used to inspect the protein structures.

Chapter 2. Automated machine learning identifies genomic variants driving antimicrobial resistance in *Salmonella enterica subsp. enterica* serovar Dublin

Introduction

Highly dimensional bacterial whole genome sequence (WGS) data provides an unprecedented scale of biological and clinically-relevant information that requires an appropriate statistical analysis framework to determine virulence and antimicrobial resistance [67-69]. Bacterial taxonomic identification using a phylogenetic approach is the most common method to infer biological underpinnings from WGS. Antimicrobial resistance (AMR) data is increasingly generated from WGS but still needs validation and calibration with simultaneous analysis with standard phenotypic antibiotic susceptibility testing[54]. Antimicrobial resistance prediction using gene presence-absence is definitive in a select bacterial species-antibiotic combination. This highlights the distinct advantage in performing purely genome-based identification of resistance over phenotyping. WGS-based methods, once fully validated, offer faster than phenotype measurements for characterizing resistance. There are multiple methods to determine the inhibitory phenotypic resistance concentrations that require extensive time, often several days to weeks, with variable variance before clinically relevant information is available. Hence, timely therapeutic decisions are delayed. This situation leaves a gap between the appropriate antibiotic treatment and sensitivity to antibiotics that can be addressed if the proper methods are in place to define the specific genes, gene dosage, and possible interactions between genes that result in resistance.

Due to the advantages of WGS based AMR prediction, databases have been developed to infer resistance from sequence data[17-19]. This approach relies on the underlying curated resistance genes to be included in the reference database that is then used to BLAST genome content to generate a similarity score based on the underlying homology to predict resistance. For example, using CARD as the reference database, bacterial genomes are noted with the corresponding antimicrobial resistance genes based on cutoff categories (strict, perfect, loose) depending on the homology scores. The approach is robust to adequately sequenced pathogens, well-annotated in the database, and for antibiotics with well-known resistance mechanisms. If the mechanism or organism is not well studied, reliance on homology often provides variable results with uncertain confidence in the genome containing specific genes that are biologically causal to resistance.

The key challenge to using databases is bacterial evolution, which continuously generates new mutations that result in diverse strategies to evade antimicrobials; hence a need for continuous curation to integrate novel mechanisms[70]. Furthermore, a significant proportion of important bacterial pathogens are underrepresented in the public sequence repositories. This generates gaps in understanding the relevant characteristics that impact the database result's confidence. A combination of genome sparsity and underrepresentation of AMR genes leads to a discord between WGS resistance prediction and phenotypic resistance[21]. There are several additional factors leading to genotype-phenotype discrepancies including, curation of the underlying database, quality of bioinformatic pipeline used and variation in laboratory-related processes that in combination lead to errors in linking WGS to phenotype[17]. Beyond the procedure for generating the sequence, considerable variation in the pathogen genome could be an explanatory factor for this discord[3,5,71]. Complex genomic variation, such as large deletions and insertions, is missed in most variant calling pipelines, which only capture 10-20 nucleotide deletions[72]. As more genomes are sequenced, this issue can be addressed with a concerted effort to expand multiple examples of genomes from the same species, especially when specific SNP-associated resistance translates to an expressed phenotype. Allelic resistance is frequently missing from some AMR prediction platforms leading to false negatives. Specifically, for example, quinolone resistance is dependent on specific mutations in the *gyrA* gene[73]. Lack of the SNP or lack of gene coordinated expression of *gyrA* and *gyrB* leads to false-positive results using gene content predictions as present, but that is not observed as resistance in phenotypic assays. This augments the need for population WGS to be included in database content that will enhance the specific mechanisms. This is also found in beta-lactam resistance where many alleles from multiple resistance genes can be in circulation within the genomic population that may shift over time. Hence, there is a need to adequately reconcile the discrepancies in antimicrobial phenotype-genotype predictions using sequence data.

We addressed the gaps in WGS based AMR prediction by testing the hypothesis that microbial variants determine AMR plasticity in the population. Specifically, we tested the predictive capability of using multi-scale genomic features (genes, alleles, indels) to determine resistance phenotypes with MICs as a measure of antimicrobial resistance. We applied automated machine learning to identify genomic variants driving AMR using WGS and MIC data for multiple antibiotics of *Salmonella enterica subsp. enterica* serovar Dublin (*S.* Dublin). Machine learning is increasingly being used in biological domains due to its ability to detect hidden patterns in big data without explicit knowledge of the underlying process, thereby bridging the need for constantly updating databases [22,74,75]. This approach allows the discovery of novel mechanisms yet to be included in AMR databases, potentially improving the predictive capability of AMR inference from WGS data.

Automated machine learning combines multiple algorithms (random forests, gradient boosted machines) or an ensemble of neurons (neural networks) into a single function, offering the added advantage of replacing a complex coding process.

Results

We tested the hypothesis that genomic variants determine the AMR plasticity of *S.* Dublin in the population by applying automated machine learning to paired WGS and AMR phenotype data (Figure 1). First, we curated 387 *S.* Dublin isolates from USDA FSIS surveillance data of beef products with paired WGS and AMR profiles defined by NARMS MICs cutoff values. The AMR profiles covered a broad range of antibiotics: chloramphenicol, sulfonamides, trimethoprim-sulfamethoxazole, tetracycline, streptomycin, gentamicin, ampicillin, amoxicillin-clavulanic acid, ceftriaxone, cefoxitin, ceftiofur, nalidixic acid, and colistin. The accompanying

WGS data was feature engineered by AMR annotation using the CARD RGI database. Finally, automated machine learning (AutoML) was applied to each antibiotic tested using resistance and susceptible as categories and annotated resistance genes and alleles as features. AutoML scored the performance of each algorithm using measures of error (mean squared error, root mean squared error, root mean squared logarithmic error, mean absolute error) between the resistance predictions of the model and the actual observed resistance in the *S. Dublin* isolates. The resulting scores are tallied in a leaderboard based on the performance of the respective algorithm relative to the measures of error (smaller values are better in prediction). From the top-ranking algorithm in the leaderboard; variable importance quantifies the relative importance of the predictor (resistance genes) within the resistance model.

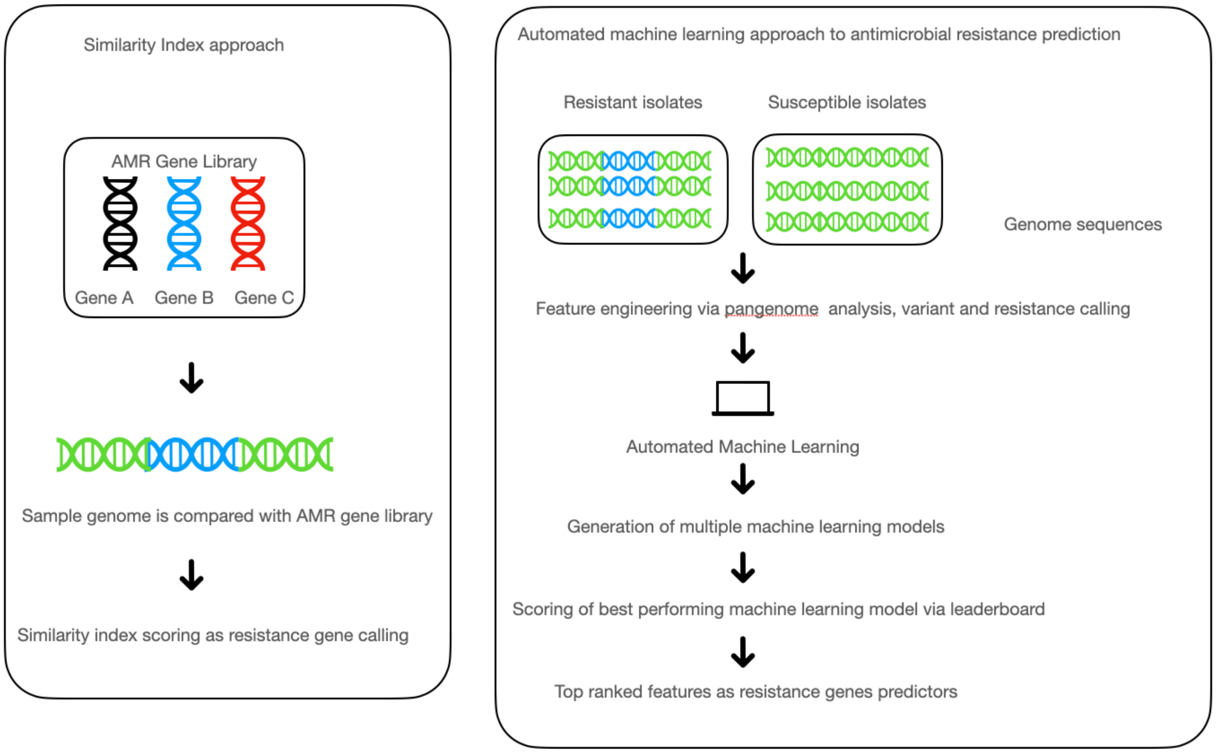

Similarity Index approach
AMR Gene Library
Gene A
Gene B
Gene C
Sample genome is compared with AMR gene library
Similarity index scoring as resistance gene calling
Automated machine learning approach to antimicrobial resistance prediction
Resistant isolates
Susceptible isolates
Genome sequences
Feature engineering via pangenome analysis, variant and resistance calling
Automated Machine Learning
Generation of multiple machine learning models
Scoring of best performing machine learning model via leaderboard
Top ranked features as resistance genes predictors

Figure 1. A conceptual framework for automated machine learning approach to antimicrobial resistance prediction contrasted with the similarity index approach. Automated machine learning is dependent on the underlying data (resistance genes) within the genomes, while similarity index is dependent on the database of resistance.

For beta-lactam antibiotic resistance in *S.* Dublin, CMY-2 is the dominant variant identified by multiple algorithms in AutoML based on variable importance ranking (Table 1). CMY-2's variable importance (100.0 %) outranked other predicted beta-lactamases (particularly homologous beta-lactamases from *E. coli).* As the predicted beta-lactamases from *E. coli* dodo not translate to Salmonella phenotypic resistance, this finding highlights the need to validate the prediction of resistance genes derived from different bacterial species. For ampicillin, deep learning was the top-ranked algorithm based on the leaderboard, although the succeeding algorithms (GBM, DRF, XGBoost) had nearly identical performance based on error measures. Beyond defining the dominant variant of beta-lactam resistance, other emerging variants were identified with the use of the order of variable importance: TEM-206 (85.0 %), TEM-214 (65.0%) and 65 % CMY-136 (65.0%). The set of variants were also ranked in the same order of variable importance for amoxicillin-clavulanic acid with a small numerical difference for (TEM-214) 66% and (CMY-136) 61 %. For ceftriaxone, the beta lactamases with high variable importance for deep learning and XGBoost algorithms were CMY-2, TEM-57 and TEM-206 with a slight difference in order. Xgboost ranked CMY-2 with 100% succeeded by TEM-57 (42.0 %), TEM-206 (29.0 %) while deep learning ranked TEM-57 higher at 73.0 % followed by CMY-2 at 52.0 % and TEM-206 (47.0%). The deep learning algorithm for cefoxitin ranked more beta lactamases for TEM-206 (82.0%), CMY-99 (80.0%), CMY-98(79.0%), CMY-17(77.0 %), CMY-2(75.0%), CMY-34 (73.0%), TEM-160 (72.0%), OXA-29 (64.0%).

Table 1. The top ranked machine learning model and top ranked variable of importance (resistance genes).

CARD	Top ranked model	Top ranked gene
beta-lactams		
Ampicillin	Deep learning	CMY-2
Co-amoxiclav	Stacked Ensemble	CMY-2
Cefoxitin	Deep learning	CMY-2
Ceftriaxone	Deep learning	CMY-2
Aminoglycosides		
Streptomycin	GBM	ANT(2'')-IA
Gentamicin	Stacked Ensemble	ANT(2”)-la
Folate pathways		
Sulfamethoxazole	Deep learning	sul2
Chloramphenicol	XGBoost	Flor
Tetracycline	GBM	tetA

For non-beta lactamases, the impact of multiple drug resistance is more pronounced in the order of ranking within each respective machine learning model. For chloramphenicol resistance, the top-ranked variable is TEM-57, a beta-lactamase (100. %), while chloramphenicol exporter (flor) is only represented with a score of (37.0) % in terms of variable importance in the model. The next top-ranked algorithm, deep learning, was able to correctly identify flor as the top-ranked variable for chloramphenicol resistance. It is worth noting that there is very little numerical difference between the top models.

Table 4. Prevalence of top-ranked gene within resistant and susceptible isolates of *Salmonella enterica subsp. enterica* serovar Dublin

CARD	Top ranked gene	Resistant phenotype with top ranked gene	Susceptible phenotype with top ranked gene
B-lactams			
Ampicillin	CMY-2	307	2
Co-amoxiclav	CMY-2	306	3
Cefoxitin	CMY-2	254	55
Ceftriaxone	CMY-2	302	7
Ceftiofur*	TEM-1	1	19
Aminoglycosides			
Streptomycin	ANT(2”)-Ia	0	21
Gentamicin	ANT(2”)-Ia	21	0

Folate pathways			
Sulfamethoxazole	sul2 floR APH(6)-Id APH(3’’)Ib tet(A)	345	4
Co-trimoxazole	CMY-2	1	4
Chloramphenicol	Flor	325	3
Tetracycline	tetA	339	1

The same effect of resistance co-occurrence is observed in sulfamethoxazole resistance, as the top-ranked algorithm, deep learning ranked sul2 (sulfonamide resistant dihydropteroate synthase) in third place (95%) in terms of variable importance following CMY-99 (100.0) and APH (3”)-Ib. This is also observed in tetracycline where tetA, an inner membrane tetracycline efflux protein, is ranked second to flor by XGBoost in terms of variable importance. Hence further curation of the underlying ranking of each variable using a white box approach allows examination of the basis of each algorithm. For non-beta lactamases, co-occurrence complicates the ranking of variable importance, but further curation based on mechanistic basis of resistance can improve the accuracy of the model.

We then analyzed the relationship of highly ranked variable importance and their presence and absence within the isolates relative to antibiotic resistance (Table 2). For beta-lactams, CMY-2 prevalence ranges from 97.0-99.0% of the resistant isolates except cefoxitin with a lower prevalence at 81.0%. For tetracycline, tetA is found in 99.0 % of the resistant isolates, sul2 in 99.0%, tetA 99.4%, gentamycin ant (2") 100.0 %, streptomycin aph (3')-1a is 78.3%. The top-ranking features generated by automated machine learning on the predicted resistance hits and AMR phenotypes consistently match known resistance mechanisms with each antibiotic class. Given the uncertainties in the predicted AMR, automated machine learning can prioritize the most likely resistance genes associated with the AMR phenotype.

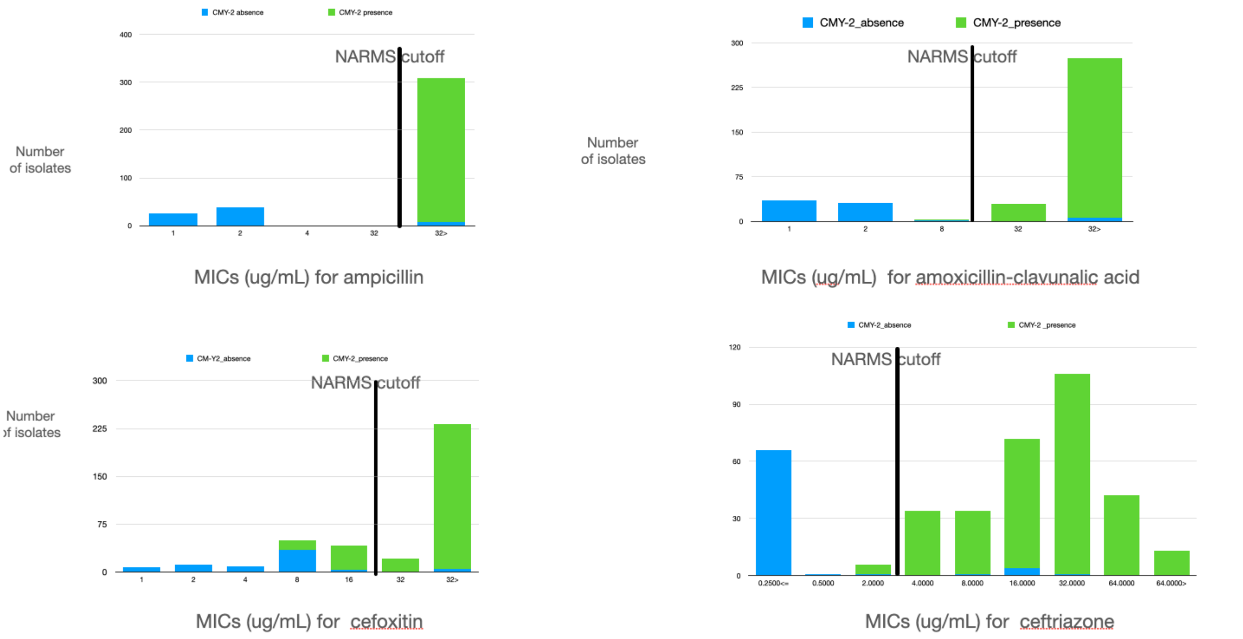


Figure 2. Genomic MICs for beta-lactam antibiotics. Green bar shows the presence of CMY-2 resistance gene relative to MICs concentrations of each antibiotic. Blue bar indicates absence of the resistance gene.

We then examined the combination of resistance genes with lower prevalence (<90.0%) in the resistance phenotype by comparing the resistance gene presence identified and ranked by

AutoML relative to MIC cutoff values. We hypothesize that resistance gene presence specifically CMY-2 determines MIC values for beta lactam resistance in *Salmonella* Dublin. Results indicate that CMY-2 presence is associated with MIC value above the resistance cutoffs like ampicillin, co-amoxiclav and ceftriaxone. Notably, cefoxitin diverges as CMY-2 is present in some of the isolates below the threshold for resistance. We compared the respective MICs relative to the presence of the dominant beta lactamase variant CMY-2 (Figure 2). Different agencies (CLSI, EUCAST) indicate a cutoff of 8.0 mg/L for susceptibles for cefoxitin in Salmonella [76]. As shown in Figure 2, this cutoff (8.0 mg/L) will delineate most isolates (284 out of 274) with CMY-59 as resistant, some isolates with CMY-59 gene (16 out 274) will be designated as susceptible hence the cause for the underlying discordance. Adjusting to cutoff based on known resistance mechanisms can mitigate the discrepancies between genotype and phenotype association. Hence a calibration of MICs cutoffs with known resistance genes could lead to a better delineation of phenotype. The gene presence and its association with resistance cutoffs is also observed in non-beta lactamase mechanisms for *S.* Dublin like tetracycline, chloramphenicol etc. except for streptomycin (Figure 3).

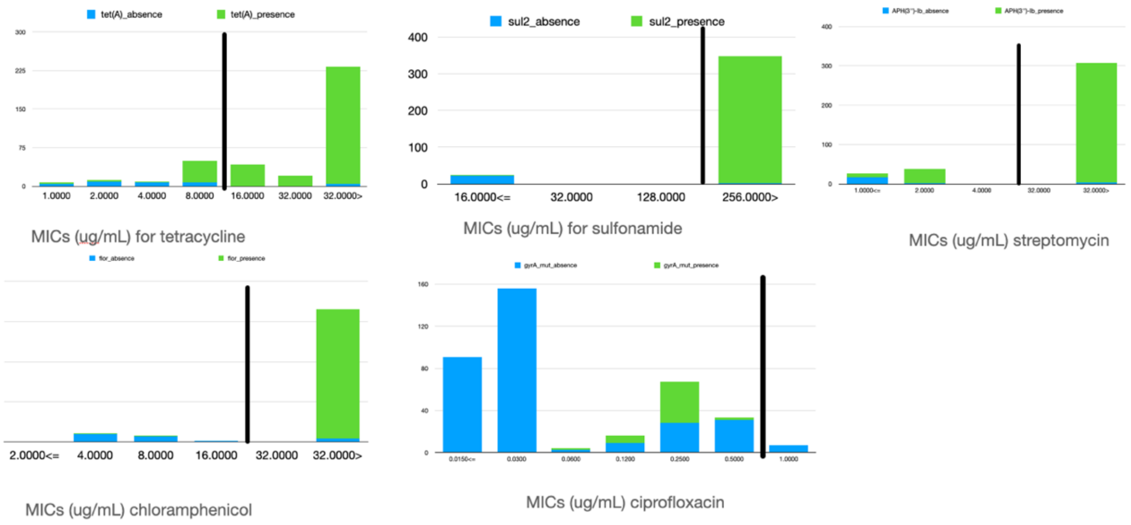


Figure 3. Green bar shows the presence of resistance gene relative to MICs concentrations of each antibiotic. Blue bar indicates absence of resistance gene.

We examined the allelic variation of *aph* (including (3'')-Ib and other variations) that is associated with streptomycin resistance. There are several genomic scales on which phenotype can be manifested from operon, gene and allele which can be appropriate unit of analysis. The current bioinformatic pipeline only resolves to the gene level, hence we perform variant calling to identify allelic variants, including SNPs and indels, followed by AutoML to determine if this approach accurately found the gene and the various alleles that may indicate specific variants cause AMR. We did not identify any alleles differentiating the resistance phenotype of streptomycin and *aph* (data not shown), indicating that no variation was associated with the phenotype/genotype discord for this antibiotic and gene, as is known for this gene. We also tried other resistance isolates as reference as to ascertain the features are sufficient extracted but still got similar negative results. We then tested another hypothesis with complex genomics variants as the driver of resistance in streptomycin in *Salmonella* Dublin. Complex variants are missed by variant callers which can handle single nucleotide polymorphisms and small indels (10-20 nucleotides) but not larger segments of the gene. The presence of 91 nucleotide deletions in (APH(3'')-Ib) results in streptomycin susceptibility (MICs > 32 mg/L) and none of the isolates with the deletions exceeded 64 mg/L MIC (Supplemental Table 1)[77].

We then applied pangenome analysis to determine the relationship between acquisition of novel phenotypes of resistance particularly with the presence of indels (Supplemental Figure 1). The overall pangenome is captured by the total gene count within the pangenome population. In the genomes of Salmonella Dublin included in this analysis, the core genes count is equivalent to 4256 with the total genes numbering 7805. The average genome count within *Salmonella* Dublin is 4743.0 while the average gene count for the isolates with the novel susceptible phenotype due to indels is 4752.0, indicating a relatively higher gene count. Hence genome diversifying events can be considered stochastic in nature but could impact manifestation of phenotypic characteristics. A gene level categorization is insufficient to resolve the resistance properties of *S.* Dublin.

The previous examples demonstrated that indels could resolve the discord between the genome content and the MIC. These mutations are not widely examined in AMR and these observations point out that the genotyping needs to be expanded. Subsequently, we used SNP analysis with nalidixic acid resistance to identify mutations associated with resistance in *gyrA* (Supplemental Table 2). This approach found 1 mutation in the nalidixic susceptible isolates and 62 mutations in the nalidixic acid resistant isolates. The nonsynonymous mutations identified in nalidixic resistant isolates (Ser83Tyr, 21 resistant isolates), (Ser83Phe ,16 resistant isolates), (Asp87Asn,18 resistant) conform with previous observations of mutations in positions 83 and 87 of gyrA resulting in quinolone resistance[78]. These mutations are potentially missed if variant analysis is not included in the methods for genomic based prediction of antimicrobial resistance.

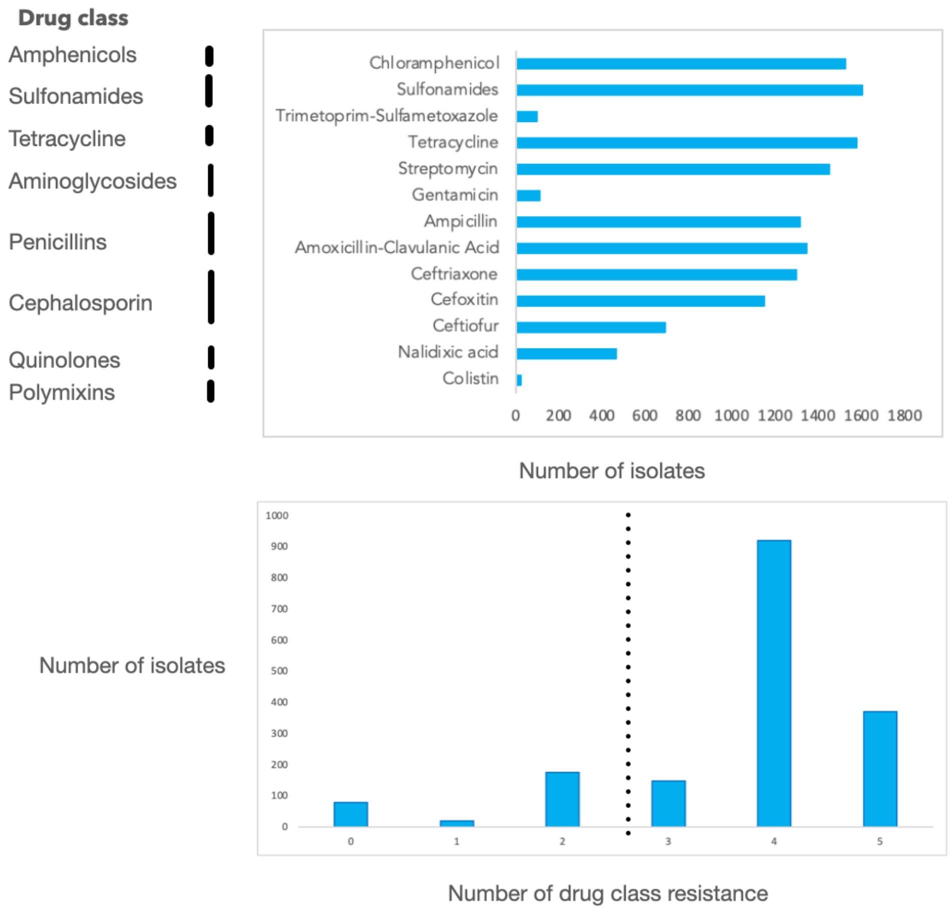


Figure 4. Antibiotic resistance phenotype from surveillance data of *Salmonella* Dublin. Bottom panel shows number of isolates with multidrug resistance defined as resistance to more than two different mechanisms.

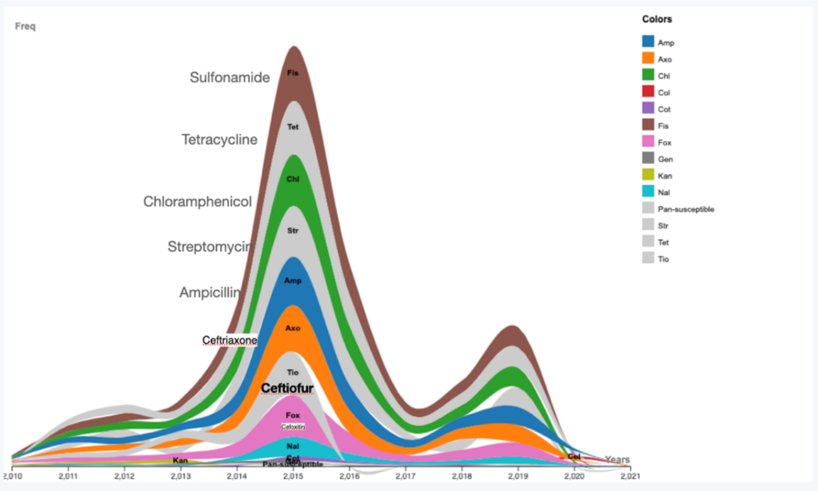


Figure 5. Longitudinal prevalence of antibiotic resistance from 2010-2020 from *Salmonella* Dublin isolates.

After confirming the resistance genotype and phenotype correlation between WGS and MICS, we examined 1707 *Salmonella* Dublin beef isolates between 2010 to 2020 from various locations in the US. A high prevalence of resistance was found among the bacterial population used in this study for sulfisoxazole (92.9%), tetracycline (91.3 %), chlortetracycline (89.2%), ampicillin (82.8%), co-amoxiclav (81.7%), ceftriaxone (80.7%), streptomycin (79.4%), streptomycin (79.4%) and cefoxitin (68.0%). Low levels of prevalence were seen with colistin (6.1%), gentamicin (5.5%), co-trimoxazole (2.9%) and intermediate levels for nalidixic acid (33.1%). A time series of AMR was examined from MICs that demonstrate a general uptrend in the number of isolated *S.* Dublin between 2013-2015 (Figure 5). The plot also demonstrates the co-occurrence of the high prevalence resistance pattern across time. There is a low-level emergence of resistance to nalidixic acid which captures resistance to quinolones in general. On the other hand, there is a decline of ceftiofur. This depicts an overall high level of resistance but also highlight specific emergence and decline of antibiotics. We then classified the co-occurrence of AMR into multidrug resistance if there is resistance to more than two drug classes (Figure 4). We observed high level of prevalence of multidrug resistance (4 drug classes, 53.8%). The high prevalence of MDR isolates is accompanied by counterpart low level of pansusceptible isolates (4.6 %).

Discussion

Predicting antimicrobial resistance from WGS is a significant challenge due to unknown resistance mechanisms, decoupled genomic and phenotype resistance characterization with MICs cutoffs determined independent of underlying genomic drivers, inadequate sampling of pathogen genomic diversity restricts the inferences, particularly of cross-species extrapolation of resistance mechanisms[79]. The application of automated machine learning to genomic surveillance data offers a scalable approach to determining dominant and emerging resistance variants. The process connects pathogen population genomics, resistance phenotyping, and the whitebox approach to machine learning algorithms. A whitebox approach to automated machine learning anchors to the concept of model explainability so that rather than focusing on prediction parameters, the constituent variables are considered by ranked variable importance[80]. We proposed this strategy to be very efficient in determining the dominant resistance variant in exploring multiple algorithms and ranking the predicted resistance genes. This complements the similarity index approach,, which could be challenging to interpret when too many homologous resistance genes are shared by evolutionary relationships that do but do not translate to functional phenotypes[81]. This dilemma is solved by ranking the predicted hits using variable importance generated from routine surveillance. Furthermore, discordance and discrepancies between top ranked gene presence and prevalence are triggers for detecting emerging variants.

One crucial element of the application of machine learning is the feature engineering step. This is a matter of defining the optimal scale of genome analysis from SNPs, k-mers, genes, and operons. While for most antimicrobial resistance, gene presence is sufficient to resolve resistance phenotypes, and by extension, mer, a subsample that efficiently represents genes, a scalable approach to cover big data for genome analysis, we have identified instances where gene presence alone is insufficient. Addressing the correct genomic scale for feature engineering is a crucial step in applying machine learning for antimicrobial resistance. Notably, most recent applications of machine learning for AMR involve k-mers and pangenome gene presence and absence, which for most antibiotic resistance is sufficient but can be inadequate for novel resistance mechanisms. These observations were commonly found in the AMR observations suggesting that unknown genetic mechanisms were not included in the reference database or that gene expression differences between the isolates account for the presence of the gene but susceptibility. As most of the AMR genotype-phenotype is shown to be concordant by applying AutoML to the resistance hits, we formulated several hypotheses for the underlying genomic mechanism of resistance that may account for the discord between the genotype and the phenotype. As the AMR hits are based on gene presence and absence, we postulated that we could recalibrate the genomic scale from gene to an allele that may not yet be captured in the reference database. If so, this may resolve the discord among many of the observations. One key finding in this study is the utility of calibration MIC cutoffs with resistance genes. The capability to define a more objective cutoff will enable a universal standard comparable worldwide. Variation in MIC cutoffs hinders comparison in surveillance data. A potential outcome from the standardization of cutoff is a better understanding of clinical outcomes because of the increase in the precision of defining resistance and susceptibility.

We note the value and importance of the said technique as s feature engineering used in the state-of-the-art AMR prediction pipeline, which is derived from comparing sequences and assigning similarity scores in the form of homology. While state of the art performs remarkably well for well-defined and sequenced species, transposing the AMR genes from one genus to another does not correlate well with beta-lactam resistance. This highlights the need to continuously sequence more samples and is a reason to perform more sequencing. Population genomics coupled with automated machine learning provides a scalable path forward of identifying organisms, defining meaningful biological clusters at different scales of analysis, enhance surveillance of AMR with a nuanced strategy to identify emerging variants.

Another important discovery in work is the synthesis of population genomics with machine learning. The ranked variable importance creates a pipeline to define dominant variants depending on how the machine learning model ranks the variable. The connection between resistance gene presence as rated by AutoML and the correlation with MIC values for each antibiotic resistance opens avenues to resolving resistance phenotype and resistance genomics at the population level. This potentially creates a streamline method to identify causative resistance genes and plot the relationship between MIC values within the population.

We demonstrate an automated machine learning approach in resolving genotype-phenotype discordance using a multidrug resistance isolate[22,69,74,75]. The system resolved two types of discrepancies (false positive, attributes resistance but is susceptible) and false negative (does not attribute resistance but is resistance) in using AMR prediction using sequence data. These discrepancies will result in overestimation or underestimation of resistance which undermine the quality of surveillance data and, in clinical settings, impact therapeutic decisions. We demonstrate the culling of false positives, which can overwhelm the accurate hits in the analysis making attribution complicated and confusing. While there might be an underlying mechanism for the false positives, like repression, a straightforward approach using machine learning can remove a significant number of false positives. Our approach also consistently identified known mechanisms of resistance and essentially separated the wheat from the chaff highlighting the resolving power of machine learning of complex biological phenomenon. We have consistently identified false negatives specially with gyrA due to limitations in the resolving power of bioinformatic pipelines (limited to genes while the underlying mechanism is allelic in nature). While this is a known limitation, published literature still propagates the discordance and results in underestimation of quinolone resistance[73]. The potential solution of using variant calling is an added step of complexity but is necessary to address proper analysis of genomes

Conclusion

This study combined automated machine learning and bacterial population genomics to identify and rank genomic variants that drive antimicrobial resistance in *Salmonella* Dublin. This approach enhances the current state-of-the-art approach based on similarity index metrics based on curated databases by ranking resistance predictions effectively reducing false positives. Furthermore, genomic MICs integrate ranked variants with resistance phenotypes in a population-wide manner providing a path to define resistance cutoffs. The detailed whitebox approach enhances the value of machine learning in discovering novel variants defining resistance mechanisms effectively updating the databases in a scalable manner. We expect the broad applicability of this approach for other serotypes of Salmonella and other species as well.

Methods

Genomes

Genomes of *Salmonella enterica subsp. enterica* serovar Dublin isolates were obtained from the Sequence Read Archive (SRA; Supplemental Table 3)[55]. The downloaded fastq files were assembled using Shovill 1.0.4 (https://github.com/tseemann/shovill). After assembly, genomes were annotated with Prokka (version 1.13.3)[62]. Snippy 4.3.5 was used to variant call the sequences (https://github.com/tseemann/snippy)[63].

Antimicrobial Resistance Profiling

A command line version of CARD Resistance Gene Identifier (5.1.1) was used to predict the resistance genes within the *Salmonella* genomes[18]. The accompanying resistance phenotype profile was provided by USDA-Food Safety Inspection System.

Automated Machine Learning (AutoML)

The resistance gene predictions from CARD RGI and the resistance phenotype was used as input for the automated machine learning using H20 AutoML graphical user interface [82]. Leaderboard rankings indicate the respective performance of difference machine learning algorithms and variable importance is inspected for each top-ranking machine learning model.

Chapter 3 Analysis of SARS-CoV-2 genomic epidemiology reveals disease transmission coupled to variant emergence and allelic variation[2]

Introduction

COVID-19 has reached global spread in all continents, except Antarctica, and was defined to be a pandemic by the World Health Organization (WHO) in March 2020[83-85]. As expected, outbreak dynamics are different among countries and regions. In part, this is due to environmental factors, contact networks, socio-cultural practices, human population characteristics, healthcare systems, the testing rate, and the public health strategies that include testing and surveillance strategies. Outbreaks are defined by the reproductive number (R)[86,87], a common measure of transmission for infectious disease spread. The probability of increased disease spread is evaluated based on the threshold when R>1; conversely a decline in spread is observed with R<1. Additionally, R can be used to estimate the proportion of the population that needs to be vaccinated in order to generate herd immunity[88], as has been discussed in a few countries as a method to control the current pandemic, as a method to measure how well population immunity is occurring in absence of a vaccine. Use of R in the context of viral mutation has yet to be examined.

Use of R for the 2020 COVID-19 pandemic was done for the initial outbreak in China as an estimate of the local epidemic expansion with the earliest estimates of R = 2.2 (95% CI, 1.4 to

[2] Bandoy, D. D. and B. C. Weimer (2021). "Analysis of SARS-CoV-2 genomic epidemiology reveals disease transmission coupled to variant emergence and allelic variation." Sci Rep **11**(1): 7380.

3.9) based on 424 cases in Wuhan, China[89]. Subsequent calculation of R, with 2033 cases from China (nationwide), slightly changed the estimate of R = 2.2 to 3.6[90]. However, estimates of R for other countries were not done routinely but rather a fixed estimate R was used based on the refined estimate based on the outbreak in China. However, even the refined estimate was inadequate in capturing spread dynamics of the pandemic and expansion within individual locations, suggesting that R was not constant at different locations of the pandemic and that a more dynamic calculation is warranted. Use of static R estimates during the epidemic spread is underestimating location and population specific outbreak dynamics during local spread[86,87]. Hence, there is a need to rapidly estimate country-specific R values during the epidemic so as to better estimate potential local hot spots that will have rapid and unexpected increases in cases. This approach can also be useful to provide global comparisons of outbreak expansion at each global location that will enable public health responses to align with the epidemiological curves across countries as well as locally.

The Wallinga and Teunis method for R estimation requires input of outbreak incidences and the period between the manifestation of symptoms in the primary case and the onset of symptoms in secondary cases to be the serial interval[91]. This approach was previously implemented in a web resource to estimate R during epidemics[92]. A key advantage of using dynamic estimates is the ease of estimating credible serial intervals, compared to other maximum likelihood estimation approaches that quickly provides valuable information to control spread of the outbreak. Additionally, integration of viral genetic variation with R estimates will provide additional information about changes in cases and indicate a change in risk. While a single report found that there was no obvious relationship between R, severity of the epidemic, SARS-CoV-2 genome diversity[93], and continual mutation of the viral genome makes this comparison an important consideration to describe outbreak dynamics so that appropriate interventions can be considered in specific locations. As the number of WGS continue to be generated, it is becoming clear that genome variation has a role in changing the epidemiological dynamics of the outbreak.

In spite of no clear path for systematic integration of viral genome variation with epidemiology, the COVID-19 pandemic is demonstrating a global unity for sharing SARS-CoV-2 whole genome sequences (WGS) with unprecedented openness. By quickly sharing genome sequences it enables investigation of the genome variation using multiple approaches to sample the virial genome space that define changes that may lead to alteration of the outbreak dynamics. WGS availability is continuing to expand and has reached a number of WGS that constitutes as a viral population for analysis, which provides additional information that cannot be gleaned from a few sequences. Population genome analysis is particularly important for SARS-CoV-2 because of the high mutation rate, which was linked by estimating transmission dynamics of rapidly evolving RNA viruses. WGS integration highlights the opportunity to infer transmission by incorporating WGS into the outbreak progression and mitigation strategies[38,94]. This approach was validated in Ebola virus (EBOV) and Middle East respiratory syndrome coronavirus (MERS-CoV) outbreaks where each virus is separated by a small number of mutations, yet these small changes produce new infection dynamics during respective outbreaks[95,96]. Rapidly evolving pathogens undergo genome sequence mutation, random drift, local selection pressure, and stochastic events that produce variant versions of the viral genomes that is likely associated with new infections [38]. Even small changes in the genome result in transmission changes that are determined by mutations between individual genomes and can be detected using WGS. SARS-CoV-2 genomes are changing over the course of outbreak but there is controversy about the impact and specifics mutations that lead to public health impacts and transmission dynamics. Viral mutations and the need for fast differentiation of changes highlights the value of systematically combining WGS with epidemiology.

Considering the lack of containment of the pandemic globally, except in Singapore, Hong Kong, and Taiwan, we hypothesized that the estimated basic R value for China do not provide reliable estimates for other countries. This is demonstrated by the observation that varies greatly by the time and location of the outbreak – highlighting the dynamic nature of R in outbreaks but more importantly in pandemics. The empirical observations of varying epidemiological curves by country, viral mutation rate, and geographically unique variation seem to accompany new cases around the world. These intertwined factors are likely individual mechanisms of change in sustaining the outbreak expansion of the pandemic. While viral sequencing is occurring quickly and the data are being made public, it is not being effectively integrated with epidemiological information because there is not an existing framework to systematically merge these different data streams. In this study, we used incidence data to estimate R and compared country specific COVID-19 transmission dynamics with viral population genome diversity. By incorporating R, the epidemic curve, and SARS-CoV-2 genome diversity we created a systematic framework that deduced how viral genome diversity can be used to describe epidemiological features of an outbreak before new cases were observed. This was done by creating a genome diversity metric that provides genome diversity context and allowed quantification of the infection dynamics globally that were divergent from the early estimates with genomic evidence. We call this approach the pathogen genome identity (GENI) scoring system. GENI scores, in combination with distinct outbreak stages, were indicative of new cases and found unrecognized local transmission.

Results

Our mutation rate calculations for SARS-CoV-2, based on the Wuhan reference genome, found the nucleotide change per month to be 1.7 (95% CI=1.4-2.0), similar to other estimates[93], with substitutions occurring at 0.9×10^{-3} (95% CI $0.5\text{-}1.4 \times 10^{-3}$) substitutions per site per year. This provided confidence that the reference genome was adequate for this study. We proceeded to determine the outbreak dynamics of COVID-19 pandemic by classifying each country's status according to epicurve stage with a framework of stages: a) index b) takeoff c) exponential d) decline as a clear method that can be used to benchmark metrics that allow a consistent integration of R and viral genome diversity measurement. First, R was determined using the instantaneous method with two different serial intervals - 2 and 7 days (Table 1). As of March 1, 2020, this framework defined global epicurves as gaining momentum globally with 52 countries in the index stage. Three countries were in the exponential stage and five countries in the takeoff stage (Figure 1). China was the only country that reached the peak of the epicurve and was characterized to be in the decline stage. There was no evidence of any other country near the decline stage, and some countries were poised to move into the takeoff and exponential phase based on the epicurve alone.

Table 1. Country-specific instantaneous reproductive number (R) estimates for SARS-CoV-2 as of March 1, 2020

		Instantaneous Reproductive Number (R) serial intervals	
Country	**Cases**	**2 days**	**7 days**
Mainland China	79251	1.6	2.1
South Korea	3150	2.8	25.6
Italy	1128	8	57.0
Iran	593	2.8	17.1
Japan	241	3.6	2.2
Singapore	102	3.3	1.6
France	100	2.9	16.9
Hong Kong	95	2.6	1.6
Germany	79	3.1	17.2
United States	70	4.3	1.7
Kuwait	45	2.6	15.3
Spain	45	3.7	10.8
Thailand	42	3.8	1.7

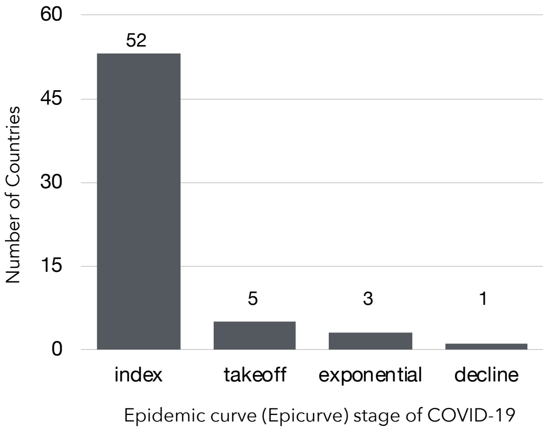


Figure 1. Distribution of country classification based on SARS-CoV-2 epicurve status.

Instantaneous R sensitively described real-time shifts of the incidence captured within each epicurve stage (Figure 2). The decline stage in China was reflected by a decrease in R estimates in the latter stages the outbreak and relative to the early estimates: 1.6 (95 % CI 0.4-2.9) and 1.8 (95 % CI 1.0-2.7) for 2- and 7-days serial interval, respectively. Superspreading events inflated R estimates seen in exponential stage that was observed in South Korea: 2.8 (95% CI 0.6-5.3) and 25.6 (95 % CI 3.0-48.2) for 2- and 7-days serial interval, respectively. Distinctive disease control was instituted in Singapore enabling it to remain in the index stage while Japan was moving to the takeoff stage characterized by increased R estimates 3.6 (95% CI 0.4-7.3) 2.2 (95% CI 1.3-3.0) for 2- and 7-days serial interval, respectively. The R estimates overlapped for all exemplar country outbreak stages in the two serial interval scenarios, suggesting that the transmission could be as short as 2 days. These estimates were relatively lower than previously reported, bringing to light the possibility of transmission during the incubation period that is associated with rapidly expanding outbreaks, which was being observed in many European counties at this time during the pandemic.

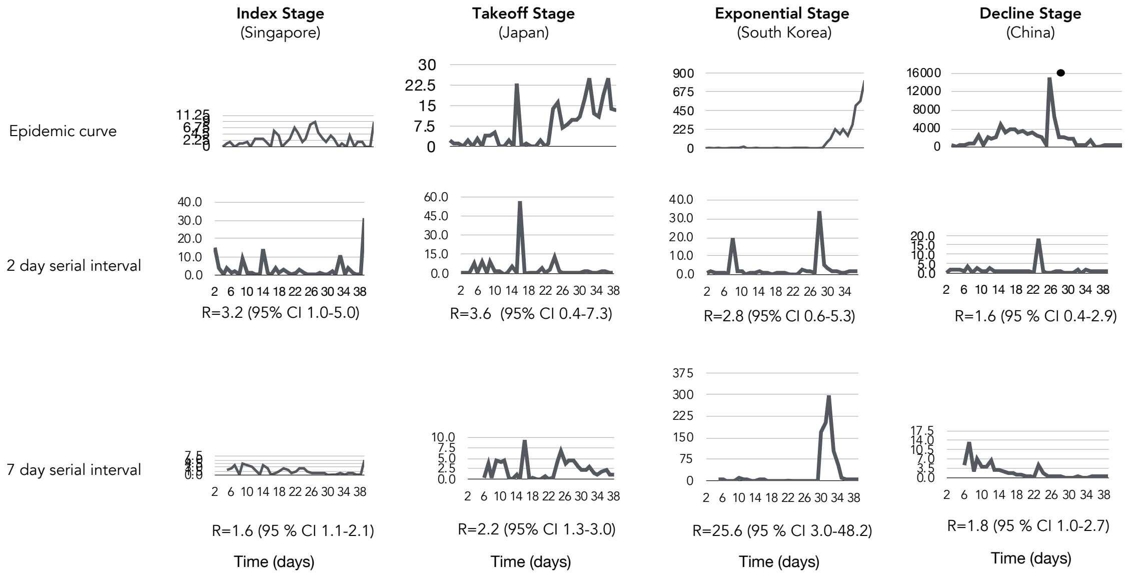


Figure 2

Figure 2. Instantaneous reproductive number estimates for different stages of the SARS-CoV-2 epidemic curve: a) index (Singapore) b) takeoff (Japan) c) Exponential (South Korea) d) decline (China) in short (2 days) and standard (7 days) serial interval. Decelerating stage of epidemic curve results to a reproductive number lower than 2 for both serial intervals, epidemic curve with multiple introductions yields 2-day serial interval with higher reproductive number and exponential serial interval yields higher reproductive number for the 7-day serial interval. The surge in the epidemic curve of China corresponds to the alteration of the case definition of SARS-CoV-2 by broadening confirmed cases with pneumonia confirmed with a computed tomography scan. South Korea's higher reproductive number is due to cryptic transmission associated with a secretive cult with altered health seeking behavior.

Low case detection of COVID-19 was observed in representative countries in the index stage with R values <2 that was attributed to effective social distancing (i.e. Hong Kong) or under detection for countries with limited testing (i.e. United States) (Figure 3a). Sustained local transmission occurred in five countries that were progressing into the takeoff stage (Japan, Germany, Spain, Kuwait and France) by R values >2 (Figure 3b). The magnitude of spread was apparent with relatively higher R estimates (>10) in Italy, Iran and South Korea, which demonstrated sudden surges in incidence due to prior undetected clusters of cases (Figure 3b). This substantially increased instantaneous R estimates relative to other estimation methods but allowed a more obvious depiction of the surge of cases that precisely differentiated the takeoff stage from the exponential stage.

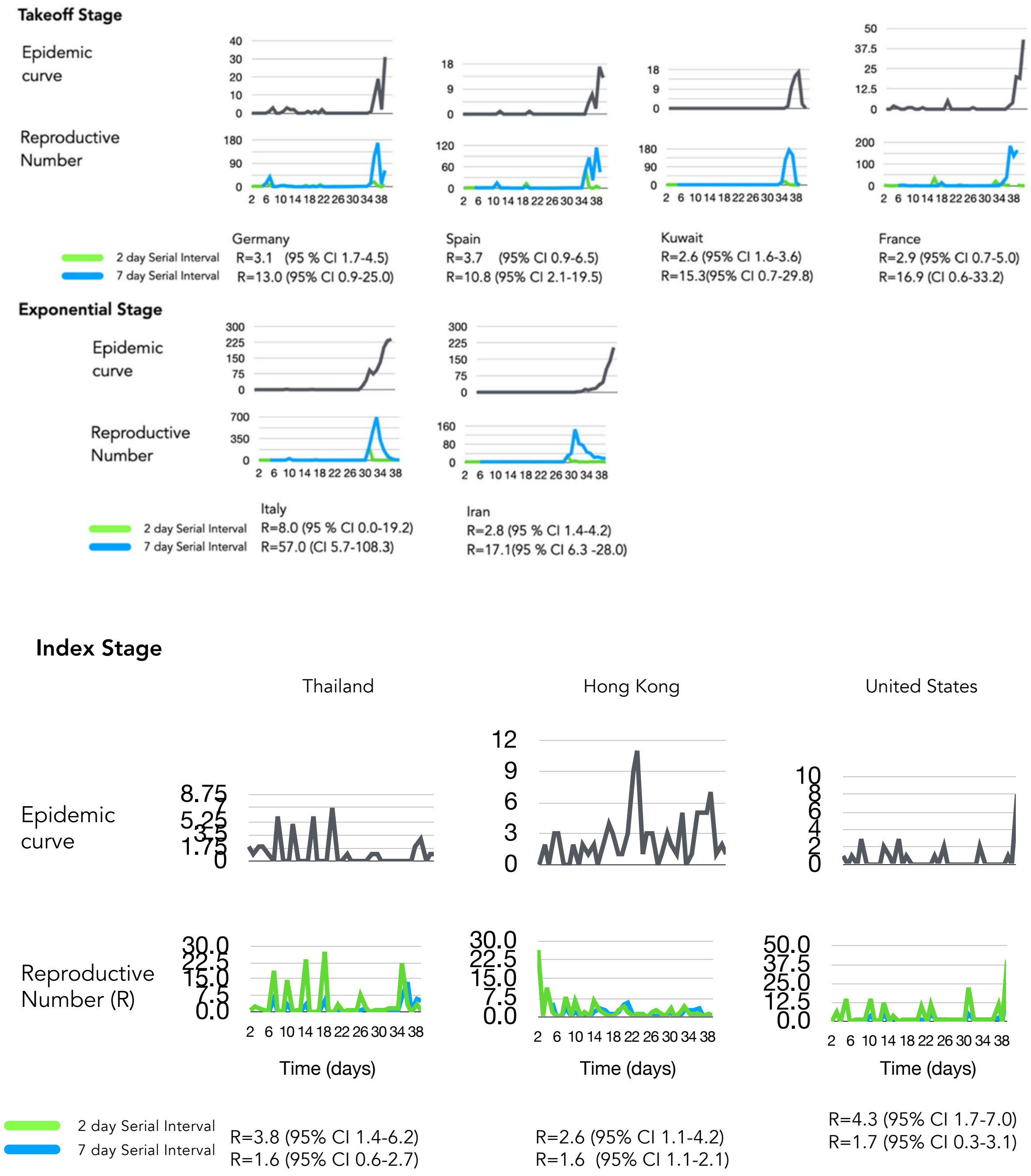


Figure 3. Epicurve estimates with different serial intervals. Panel A represents Epicurves and instantaneous R values for index stage countries using 2- and 7-day serial interval. Panel B Global dynamics of SARS-CoV-2 using instantaneous estimate of reproductive number with 2-day serial interval. Under preincubation period infectivity scenario, globally increasing R > 2. Italy's R = 8 is highest due to late detection of infection clusters. This higher R estimate is due to a huge bump in cases combined with diagnostic gap of low-level incidence. The same surge dynamics is seen in South Korea. Global dynamics of SARS-CoV-2 using instantaneous estimate of reproductive number with 7-day serial interval.

Italy's R value inflates to 57 with the 7-day serial interval assumption and overlaps with the lower threshold of 2-day serial interval R estimate. This estimation depicts a decreasing pattern for countries multiple introductions like Singapore, Hong Kong.

We further examined the association of country-specific instantaneous R estimates by comparing different local temperature ranges (tropical versus temperate) and population density of representative cities with outbreaks. The higher temperature range and population density were used for selected countries; however, no direct link was observed (Table 2). Case increases for South Korea were largely associated with an outbreak among a secretive religious group Shinsheonji (73% cases of COVID-19 in South Korea), located mainly in Daegu with a lower population density 883/km^2 as compared to the rest of the areas with an outbreak[97] and may explains the outbreak expansion early in the epicurve rather than the area's population density. While most representative countries (Table 2) have cooler temperatures (10-6˚C), Singapore's higher temperatures indicated that local transmission occurred at higher temperatures and suggests that temperature shifts will not likely change transmission. The temperature and population density did not explain changes in the epicurve. This led us to hypothesize that the viral genomic variation underpinned changes in the epicurve in each country.

Table 2. Epidemiological parameters and instantaneous R estimates.

Country	Reproductive Number (R)	Temperature (˚C) during outbreak	Population Density (people/km^2)	Interpretation in consideration of the epidemiological curve
Singapore	3.3	32	8136	Imported cases, limited local transmission
France	2.9	10	4300	Imported, Local transmission >1-2 month
Italy	8	10	7200	Imported cases, Local transmission >1 month
United States	4.3	9	8444	Imported cases, Local transmission >2 month
South Korea	2.8	6	883	Imported cases, Local transmission >1-2 month

We determined the relationship of epicurve stage with viral genetic variation using a metric that merges absolute genome variation with the rate of genome change to create the GENI score. This approach anchored viral genome diversity with the rate of evolution for SARS-CoV-2 to create an index that is comparable between countries and progression of the outbreak. To examine how the viral genome diversity was associated with the epicurve stages we first examined the index stage (Singapore) and the exponential (South Korea). Integration of GENI scores successfully distinguished the index and exponential stages (Figure 4). An increase in the GENI score was associated with the exponential stage at a median score = 4, suggesting that the viral diversity and rate of mutation was directly proportional to case increases during this stage. Singapore (index stage) had a GENI score = 2. This was found in multiple time points during the outbreak, where multiple mutation events were directly associated with an increase in cases. While China was in the decline stage the retrospective association with R, cases, and the GENI score provided longitudinal evidence of multiple case expansions with viral mutation events. This observation was especially clear early in the epicurve and indicated that SARS-CoV-2 was circulating in China at least 1 month prior to the official declaration of the outbreak (Figure 4). Merging these estimates provided evidence that repeated viral mutations indicated a change in the epicurve. These metrics were associated at each time point over 3 months, in three countries, and in three different outbreak stages. This finding is useful in integrating virus genome diversity and evolution rate into assessment of outbreak status. The approach successfully replicated the observation in viral movement between countries and within a country when the epicurve was combined into a triad with instantaneous R estimates. The proportionality of GENI scores with the epicurve stage indicated the stage of outbreak as well as determining the outbreak status.

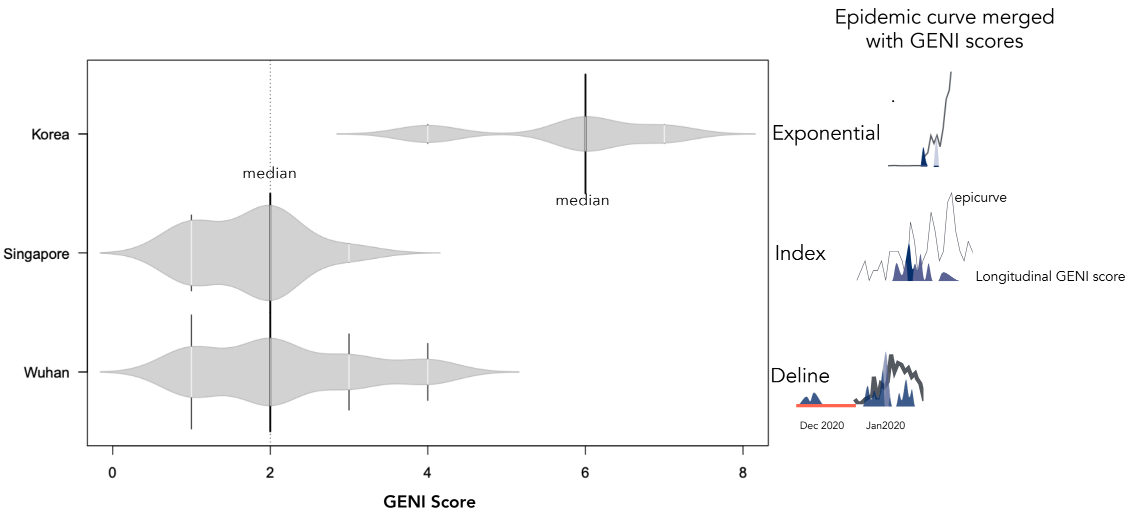


Figure 4. Relationship of pathogen genome identity (GENI) score with the temporal signal along the epidemic curve. Local transmission is captured by virus mutation as expressed in GENI score values. GENI scores of SARS-CoV-2 isolates are relative to Wuhan reference strain Wuhan-Hu-1 NC_045512.2. The red line in the China epicurve represents the time before an outbreak was determined yet genome sequences were circulating. The blue shaded curves indicate GENE scores directly overlaid with the outbreak curve. The dotted line represents the common point in time as a reference for visualization. The GENI score and epicurve show similarity except in China as the outbreak advanced to takeoff and exponential the GENI score increased while in the index stage example of Singapore the outbreak was contained and the GENI score remained <2.

Table 3. Relationship of Pathogen Genome Identity (GENI) Score derived from mutational difference from the index genome (Wuhan isolate of SARS-CoV-2 or cluster isolate reference from multiple outbreak regions outside of territory).

Equivalent Pathogen Genome Identity (GENI) score for SARS-CoV-2	Clinical Interpretation and Epidemiological Inference	Note
0-2	No difference from index case isolate genome or reference, imported case if there is no prior	Reference genome is primarily earliest isolate available.

	report, indicative of acute transmission <1 month	
3-4	recent local transmission (average 1-2 months) if there are no prior report of cases	Subsequent outbreak clusters can serve as sources of introduction hence near neighbor reference has to be selected to generate an accurate GENI score.
>4	sustained local transmission (greater than 2 months) if there is are no prior report of cases	Subsequent outbreak clusters can serve as sources of introduction hence near neighbor reference has to be selected to generate an accurate GENI score.

A framework to merge epidemiology and population genomics was derived from this study as a method to systematically integrate molecular epidemiology into public health (Figure 5). It required dynamic measurements be taken for R and surveillance efforts to determine WGS for each virus. Ideally, each case would have multiple WGS as the disease progressed, but this was not available. Using this triad of measurements accurately and quickly provided insight to measure outbreak progress but also provided an evidence-based method to judge intervention effectiveness. This study demonstrated an advancement of how to use population genomics in an infectious disease, particularly when the mutation rate is fast and the genome diversity of the population is large, such as SARS-CoV-2. GENI scores provided a missing element of evidence that defined how to estimate new cases approximately 2-5 days before they appeared. GENI score estimation accuracy increases with analysis of large numbers of genomes (i.e. populations of genomes) and from different global locations.

Further examination of this approach was done using genomes and epidemiology curves from February to April 2020, which captured documented surges in outbreaks that were aligned with the GENI score in the UK. This analysis led to further validation that genomic variation was occurring even during lockdown that was aimed at reducing the outbreak and was predictive of recurring surges in infections using >16,000 genomes (Figure 5). Low numbers of new cases were observed (Figure 5 inset) was associated with a variable GENI score (February 2020). As the cases surged in April 2020 the GENI score rose at a constant rate indicating that the genomic variation was increasing as cases were increasing. Instituting a government lockdown aimed to reduce exposure did cause variable changes in the outbreak curve it had no effect on the GENI score, which continued to rise indicating that when exposure occurred the virus was readily able to infect the person. This suggests that the underlying causes of new cases have two components – viral genome variation (evolution) and individual exposure. With this concept in mind, it can explain 'superspreading' events based on the continued genome evolution to maintain or expand host range that readily infect people that form large groups to quickly lead to new cases. Demonstration of this repeated observation using a longitudinal analysis with >16000 genomes and hundreds of cases lends extremely strong support to the notion that measuring allelic diversity is predictive of higher transmission and it will be observed when the appropriate conditions in large groups or exposure using outbreak curves. However, additional work is needed to specifically indicate the exact mutations that will initiate new cases more quickly, as demonstrated with emergence of the B.1.1.7 linage in late 2020 within the UK and quickly spread globally.

This study demonstrated an advancement of how to use population genomics in an infectious disease, particularly when the mutation rate is fast and the genome diversity of the population is large, such as SARS-CoV-2. GENI scores provided a missing element of evidence that defined how to estimate new cases approximately 2-5 days before they appeared. GENI score estimation accuracy increases with analysis of large numbers of genomes (i.e. populations of genomes) and from different global locations as demonstrated (Figure 5). Consequently, a framework to merge epidemiology and population genomics was derived from this study as a method to systematically integrate molecular epidemiology into public health (Figure 6). It required dynamic measurements be taken for R and surveillance efforts to determine WGS for each virus. Ideally, each case would have multiple WGS as the disease progressed, but this was not available. Using this triad of measurements accurately and quickly provided insight to measure outbreak progress but also provided an evidence-based method to judge intervention effectiveness.

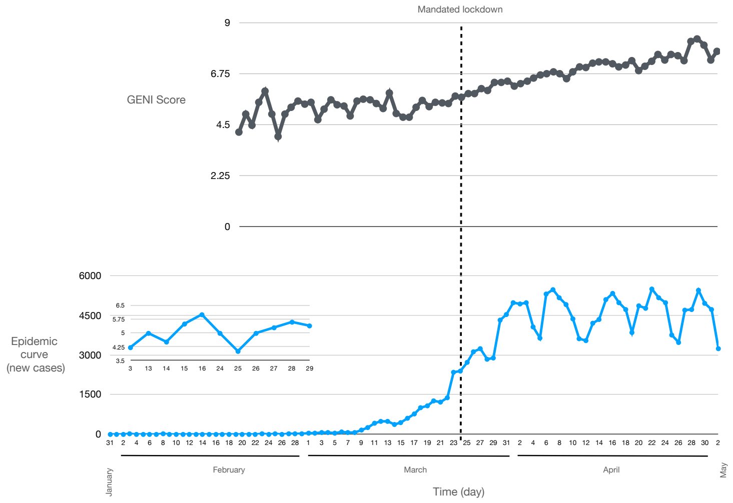


Figure 5. The GENI Score of 13419 SARS-CoV-2 sequences in United Kingdom (top) and the epidemic curve. A high initial GENI score suggests cryptic viral transmission while a consistent GENI score

indicates an increase in transmission as the pandemic progresses. This also indicates continued mutations increase viral genome diversification.

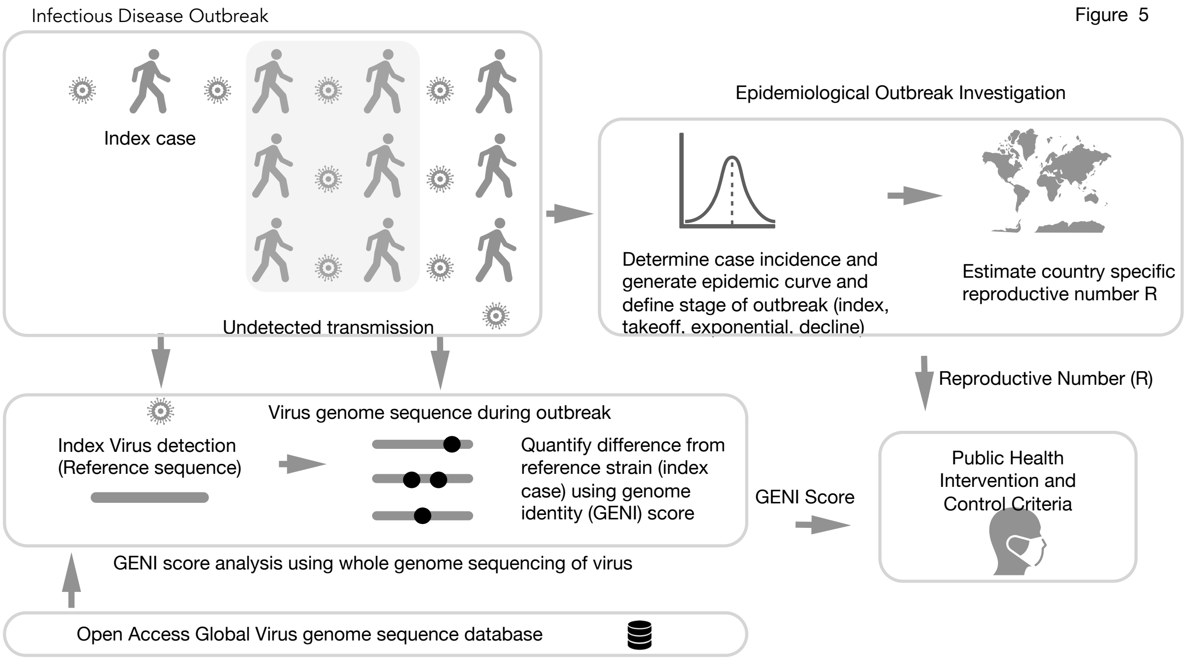

Infectious Disease Outbreak
Figure 5
Index case
Undetected transmission
Epidemiological Outbreak Investigation
Determine case incidence and generate epidemic curve and define stage of outbreak (index, takeoff, exponential, decline)
Estimate country specific reproductive number R
Reproductive Number (R)
Virus genome sequence during outbreak
Index Virus detection (Reference sequence)
Quantify difference from reference strain (index case) using genome identity (GENI) score
GENI Score
Public Health Intervention and Control Criteria
GENI score analysis using whole genome sequencing of virus
Open Access Global Virus genome sequence database

Figure 6. Integration of genomic and classical epidemiology for outbreak investigation. The foundation of epidemiology is the accurate and timely reporting of cases which enable the calculation of the number. Genomic Identity (GENI) score is formulated from genomic data of pathogens to differentiate imported cases versus local transmission and measure time of cryptic spread. Together these two epidemic values deliver insight that can be directly used for making decision criteria for public health intervention.

Discussion

Public health response is proportional to the severity and transmission dynamics of an infectious disease outbreak. This requires epidemiological metrics that can be used as decision criteria, and ideally, they can be used to assess impact of the intervention. In this work we determined that R was more dynamic in the SARS-CoV-2 pandemic than previously appreciated among the countries examined (Fig 2-3). The instantaneous R estimation with a serial interval of 2 was extremely sensitive to shifts in the epicurve during the index phase (Fig

2-3). Singapore was an excellent example of effectively controlling and containing the SARS-CoV-2 outbreak in spite of multiple mutation or multiple introduction events. They previously designated a response system called Dorscon (Disease Outbreak Response System Condition)[98] providing a systematic approach to control, which seemed to effectively control transmission so that they did not moved beyond the index phase. In contrast, other countries in this phase were poised to move into the takeoff phase (Fig 3). The transition into the takeoff phase was accompanied by a transition from a 2-day serial interval to a 7-day serial interval determine shifts in the epicurve.

Gaps in testing created a challenge in accurately defining the epicurve status. To address this diagnostic limitation, while estimates of R alone is insightful in retrospect, they alone lacked robust predictive value in this study. To overcome this limitation, we merged GENI estimates based on WGS variation and the mutation rate with the epicurve and R to provide a predictive triad of measurement that resulted in insight that accurately refined case expansion (Figure 4). Each phase of the outbreak was categorized with mutations that were associated with new cases in established outbreaks. The merged evidence indicated that China had circulating virus at least 1 month prior to the recognized outbreak. Independent of the phase framework, merging GENI scores with the epicurve found new cases in the same timeframe as new sequence variants emerged. Previous studies where the relationship of genomic diversity with epidemic severity (i.e. R) found no clear link[93]. However, by merging instantaneous R, the epicurve stage, and the GENI index we determined that a link does exist for each country examined and that this approach resulted in a direct prediction of outbreak dynamics and genomic mutations as well as the mutation rate.

The GENI index provided a basis to examine imported cases or locally spreading, both of which were addressed in this current work using established metric - R and novel integration of WGS to define changes in the sequence that were directly predictive of increases in cases. This approach leads to an epidemiological framework that is scientifically robust and at the same time can convey complex biological properties to enable an efficient characterization of an outbreak in combination. Transforming complex pathogen characteristics were accessible to the public health and medical fields using the GENI score as a complete merged information set with other characteristics of the outbreak.

Previous outbreaks, such as Ebola, employed state of the art analysis using phylodynamics that is anchored on the genetic evolution[95]. Inference, such as time to most recent common ancestor, allowed estimation of outbreak origin, population size, and R – yet this was not integrated into the outbreak dynamics and stage of advancement in the outbreak. This type of analysis is possible because genomic sequences carry temporal signals and when used in context with samples collected longitudinally, previous divergence can be determined, which has been used to do source tracking. However, the GENI score includes these signals and expands their use by merging them with the outbreak dynamic using the population genome variation as well as the mutation rate to provide an index related to the epicurve – one that was directly predictive of new cases – opposed to the genealogy of the virus.

This approach is not limited to viruses. Another recent example, in a bacterial setting, was the cholerae outbreak in Haiti wherein the phylogenetic analysis resolved the origin of the pathogen[99]. However, for this analysis to succeed, a substantial genome sequence database, of isolates collected across time and geographic location, was needed to enable placement in a phylogenetic context. As outbreaks are bound to happen in the future, investment in cataloguing the genomic space of pathogens is even more important than previously appreciated[100,101]. It is critical to obtain COVID-19 sequences from humans as well as other animals that have zoonotic potential. This was demonstrated previously with zoonotic *Campylobacter* species[102,103] that enabled disease in a variety of host species. Creating sequence repositories for pathogens is critical and underway for various pathogens[101] as well as SARS-CoV-2[104].

Prior work forewarned the flaw of being overly dependent on early estimates of R alone[105]. By having the most accurate possible information for a dynamic metric and taking into account the complex dynamics that factor in the calculation of R along with merging this the WGS and mutation rates of the pathogen a robust and insightful method to assess outbreak dynamics was created in this study. Openness and data sharing of incidence reports and sequences at unprecedented scale is being done in this pandemic and it is paying rewards[106].

Examples where information was not shared were observed in several countries and it led to cryptic spread of the disease in countries that exacerbated the outbreak. Leveraging shared resources opens unexpected collaboration and avenues for applying relevant bioinformatic and disease modelling skills across the scientific community to solve global public health problems very quickly. Establishing a systematic framework to merge epidemiology and genomics was defined in this work (Figure 5) to provide an evidence-based approach that can be used to predict locations for new cases or applied to examine intervention effectiveness to control new cases.

Conclusion

This study integrated population genomics into epidemiological methods to provide a framework for molecular epidemiology. Specifically, this study demonstrated epicurves, instantaneous R estimates, and GENI scores for SARS-CoV-2 are useful as pandemic metrics and in combination are a robust method. It was demonstrated that the pandemic is poised to become larger and that mutation will be associated with the increase in cases. Exemplar outbreaks, such as Singapore, found increases in cases with viral mutations that were effectively controlled. However, other outbreaks had expanding R estimates during the outbreak, as well as numerous viral mutation events. Use of epicurve stages, instantaneous R estimates, and GENI provided a robust and accurate framework to monitor outbreak progression to different stages with direct association between cases and increases in each metric.

Methods

Chinese CDC and WHO situations reports were used to assemble the incidence data as compiled by the Center for Systems Science and Engineering by the John Hopkins University (Baltimore, MD, USA) that was accessed on March 1, 2020[107] to construct epidemic curves (epicurves). We defined four groups along the epicurve that characterized increasing expansion and a decline phase that was used as markers of specific events for each outbreak.

The extracted time series case data were used as input for determining the instantaneous R on a daily basis to effectively capture dynamic changes in case reports. The estimates of R were selected at 2 and 7 days to examine fluctuations in reporting as between the defined phases. A parametric of uncertainty (offset gamma) and distributional estimates for the serial interval were used. A mean of 2 and 7 days, with standard deviation of 1 was used to capture short and standard serial interval assumptions using 50 sub-samples of the serial interval distribution. The Wallinga and Teunis method, as implemented by Ferguson[92], is a likelihood-based estimation procedure that captures the temporal pattern of the effective R from an observed epidemic curve. R was calculated using the web application EpiEstim App (https://shiny.dide.imperial.ac.uk/epiestim/)[92]. The descriptive statistics were used to compute the mean and confidence intervals to estimate the instantaneous R.

The GENI score was anchored on the principle of rapid pathogen evolution between transmission events. This required defining a reference sequence from the outbreak, which in this study was the Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1

NC_045512.2[108]. Publicly available virus WGS were retrieved from GISAID with whole genome variant determination using Snippy (version 4.6.0)[104,109,110]. The average mutation/isolate was divided by the total epicurve time (days) to derive a daily epidemic mutation rate that was scaled to a monthly rate that was produced. We derived a transformed value of this rate before integrating it with epidemiological information. The output from the variant calling step was then used to determine GENI score by calculating the individual nucleotide difference over the entire genome from the reference. The basis for GENI score cutoffs, to estimate transmission dates, were derived from accepted evolutionary inference of mutation rates of SARS-CoV-2 of 2 mutations/month.

We defined four epicurve stages to provide a clear method to define increases in the outbreak. First, the 'index stage' was characterized by the first report (index case) or limited local transmission indicated by intermittent zero incidence from an undulating epicurve. Second, a distinct stage, we defined to be the 'takeoff stage', wherein the troughs are approximately the same level as the previous peak but no longer reached zero. Third, the 'exponential stage' was characterized by a sharp upward increase where the outbreak was expanding quickly, and a large number of new cases emerged daily. The last stage was defined as the 'decline' and was noted when the outbreak past the peak and newly reported case counts were smaller than the previous day. Transition into the decline stage ultimately resulted in few to no new cases being reported, yet viral circulation was likely still occurring and new WGS were being found in each outbreak.

Chapter 4 Direct estimation of disease transmission from sequence data with genomic epidynamics reveal variable vaccine effectiveness with SARS-CoV-2 variants

Introduction

SARS-CoV-2, the etiologic agent COVID-19, consistently evolves at a rate of 2 mutations per month[111]. While most mutational changes are random and do not alter viral properties, selected mutations impact epidemiological parameters, including transmissibility, virulence, diagnostic performance, therapeutic and vaccine efficacy[33,112]. Mutational changes in the SARS-CoV-2 genome are monitored using whole-genome sequencing surveillance worldwide and have generated more than a million WGS within a year of the pandemic. The outcome of genomic surveillance approaches is a classification scheme based on the relative risk that designates variants of concern (VOC) for detecting signals of phenotypic changes linked with specific mutations and are associated with increasing cases[113]. However, the impact of mutations on the viral properties is uncertain, resulting to the designation of variants of interest (VOI). Consequently, determining the effect of specific mutations that alter disease transmission dynamics based on viral WGS will bring about a needed and impactful tool for public health decisions.

Various approaches have been explored to determine disease transmission using virus WGS, most of which rely on the relationship of viral evolution and disease transmission via phylogenetic analysis[33,42]. The underlying disease transmission network is either directly overlaid in the resulting phylogenetic tree to map "who infected whom" for the phylodynamic models fitted to the time series data generated from clades or lineages to determine the most probable explanation of the underlying data for the phenomenological approach[33,114]. Both approaches have produced insight into virus evolution and disease transmission early in the pandemic. Unfortunately, lineage estimation becomes more difficult with large sets of WGS, leading to data reduction and lineage breadth estimates that unintentionally hide the vast WGS diversification as the pandemic continues. This has led to using a limited number of samples that significantly hinders the use of variants to track cases and disease transmission, ultimately impacting surveillance vaccine effectiveness. However, it is particularly challenging to handle population-scale sequences with lineages creating a gap in generating WGS and bioinformatic approaches that provide impactful integration of epidemiological needs for public health. Since there is no universally accepted definition of a viral lineage that resulted in multiple competing classification schemes[43]. Generating phylogenies from population-scale WGS data has an upper limit for computation and display under 1000, which mandates a need for a new analysis and visualization platform. Unfortunately, it is computationally impossible to create phylogenetic trees with hundreds of thousands of samples for meaningful epidemiological analysis[43]. Hence, most phylodynamic analyses use a data reduction approach, limiting analyses to a few hundred WGS (i.e., cases); this approach cannot fully take advantage of the available WGS and leads to limited insight as more variants, and WGS emerge. This highlights the need to develop strategies in determining disease transmission dynamics in WGS that are scalable in computation and insight determination for public health decisions.

Previously, we designed an approach that specifically linked the mutation to increases in cases to determine that variant tracking using individual mutations accurately integrates epidemiological curves[111]. While variant tracking is an excellent metric that combines longitudinal information to estimate new outbreaks, an extension of this approach may determine variants that lead to vaccine breakthrough infections. To address this gap, we propose directly estimating disease transmission parameters from SARS-CoV-2 WGS using epidemiological dynamics (genomic epidynamics). Genomic epidynamics directly converts viral WGS into epidemiological data, effectively bridging population genomics with epidemiology in a dynamic longitudinal analysis. This contrasts with phylodynamics, wherein epidemiological inference is indirect and needs a phylogenetic tree generation step that is becoming impossible due to the WGS scale or severe data reduction that is counter to the advantage that WGS brings to disease transmission. Genomic epidynamics uses two data inputs to resolve population-scale transmission networks: temporal metadata that is derived from collection dates of the genome samples and WGS of SARS-CoV-2, which was used to generate a list of mutations for each isolate relative to the reference sequence (ex.Wuhan-Hu-1) by established bioinformatic variant calling methods. Temporal metadata and the catalog of mutations for each SARS-CoV-2 isolate are then combined into a variant epidemic curve (epicurves), based on the transmission routes inference with common variants[115]. The presence of common variants, defined as identical nucleotide mutation of virus samples from two infected individuals linked by the time of infection, is considered strong evidence for direct transmission[116]. Adding time of infection with common variants data provides directionality, effectively establishing 'who infected whom,' which is derived from the collection time of the samples. The resulting variant epicurves then provide a platform to derive vital epidemiological parameters.

One key disease transmission parameter that can be estimated from variant epicurves is the variant reproductive number (Rv) based on established epidemiological methods[111,117], except it is done with the variant WGS rather than total confirmed cases that rely on RT-PCR. Prior methods of estimating reproductive number (R) from epidemic curves have only temporal signal from onset of clinical signs to infer infection networks. The proposed approach adds a high level of precision from WGS data and direct integration with mathematical associations for every single variant and WGS. While the incident case-based R represents the average number of secondary cases from a single case, Rv represents the number of sequences generated by the index sequence with the mutation. The resulting Rv values are interpreted as either increasing transmission (Rv > 1) or decreasing transmission (Rv <1), which can be used to assess effectiveness of nonpharmaceutical interventions during the early stages of the pandemic (2020) as well as the impact of pharmaceutical use, such as vaccination (2021).

In this study, we determined the impact on disease transmission of SARS-CoV-2 mutations using genomic epidynamics via Rv calculation and longitudinal metadata integration. We focused on viral mutations with validated phenotypic changes in laboratory experiments and assessed whether those changes directly impact disease transmission and vaccination effectiveness within the population. The mutations from VOCs included in this study had: a) spike protein D614G, b) spike protein N501Y and P681H (Alpha variant, B.1.17), and c) spike protein L452R, T478K, P681R (Delta variant, B.1.617.2). These SNPs were selected for proof of concept, but the method is not limited to these mutations or a subset of SNPs. Mutations N501Y, L452R, and T478K are in the spike protein's receptor binding domain (RBD), while P681R is in the furin-mediated spike cleavage site. We hypothesized that spike protein mutations in the RBD and furin-mediated cleavage site reduce vaccine effectiveness in the population. To test the hypothesis, we directly converted SARS-CoV-2 WGS to variant epicurves and subsequently derived Rv for each mutation to measure transmission dynamics compared to actual disease in the UK. We then compared the transmission differences of SARS-CoV-2 mutations (N501Y) and (L452R, T478K, P681R) variants before and during the rollout of vaccinations and assessed the predicted prevalence of infections using a compartmental model[33,111,112]. Lastly, we determined the capability of single and full dose vaccine administration to reduce transmission of SAR-CoV-2 variants within the population to assess mutation impacts on vaccination breakthrough infection capability.

Results

We hypothesized that SARS-CoV-2 mutations in the spike protein's receptor binding domain (RBD) enhance disease transmission in the population. To test this hypothesis, we developed genomic epidynamics, an epidemiology first approach in analyzing whole-genome sequences of pathogens. Genomic epidynamics combines WGS and temporal data into a variant epidemic curve, directly converting sequence data into an epidemiologically tractable format. We used common variants between isolates as transmission inference and sampling time as temporal data to generate variant epicurves. We showed the scalability of genomic epidynamics by generating variant epicurves from 297,805 sequences during the second wave in the last quarter of 2020 from the COG-UK consortium resulting in a high density of WGS relative to COVID-19 confirmed cases (genome sequence to case ratio ~7.1 to 28.1%) (Figure 1) (Supplemental Figure 1). We used genomic epidynamics to test the differential transmission hypothesis by comparing the variant epicurves of the reference N501 and the mutational variant N501Y (Figure 2). The variant epicurve of N501Y demonstrated a strong fit (R^2=.9) to an exponential curve (Supplemental Figure 2) while the counterpart epicurve of the reference showed declining numbers, indicating a change in transmission dynamics within the population of isolates with N501Y mutation. We then used Apple mobility data to distinguish increased social interaction as alternative driver of increased transmission versus N501Y mutation. We compared the average baseline mobility prior to the emergence of N501Y consisting of driving, public transport and walking (Supplemental Figure 3). Incremental changes in mobility (driving 16%, public transport 9% and walking 2%) prior and during emergence of N501Y could not account for the change in transmission dynamics. Hence the exponential rise of N501Y is not attributable to changes in social mobility data, further supporting the hypothesis of increased transmission due to mutational changes. These exemplars capture the value of being able to convert SARS-CoV-2 WGS data to a time series allowing detection of trends at the population scale that evolve with time and outbreak dynamics. This provides the foundation of genomic epidynamics that can be extrapolated to all variants.

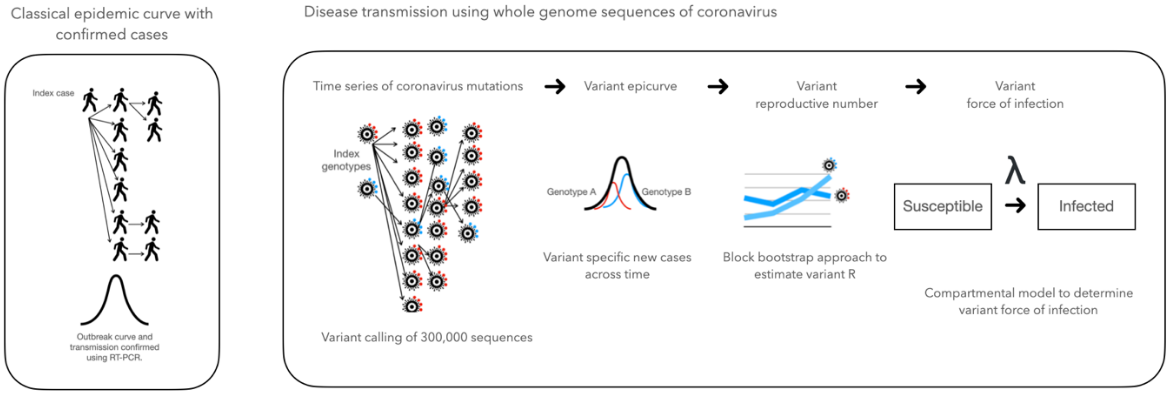

Classical epidemic curve with confirmed cases
Index case
Outbreak curve and transmission confirmed using RT-PCR.
Disease transmission using whole genome sequences of coronavirus
Time series of coronavirus mutations
Variant epicurve
Variant reproductive number
Variant force of infection
Index genotypes
Genotype A
Genotype B
λ
Susceptible
Infected
Variant specific new cases across time
Block bootstrap approach to estimate variant R
Compartmental model to determine variant force of infection
Variant calling of 300,000 sequences

Figure 1. Temporal variant framework for disease transmission inference. The estimates of disease transmission of each respective mutant are estimated using reproductive number. A compartmental model using the variant R is used to simulate the impact of different transmission properties.

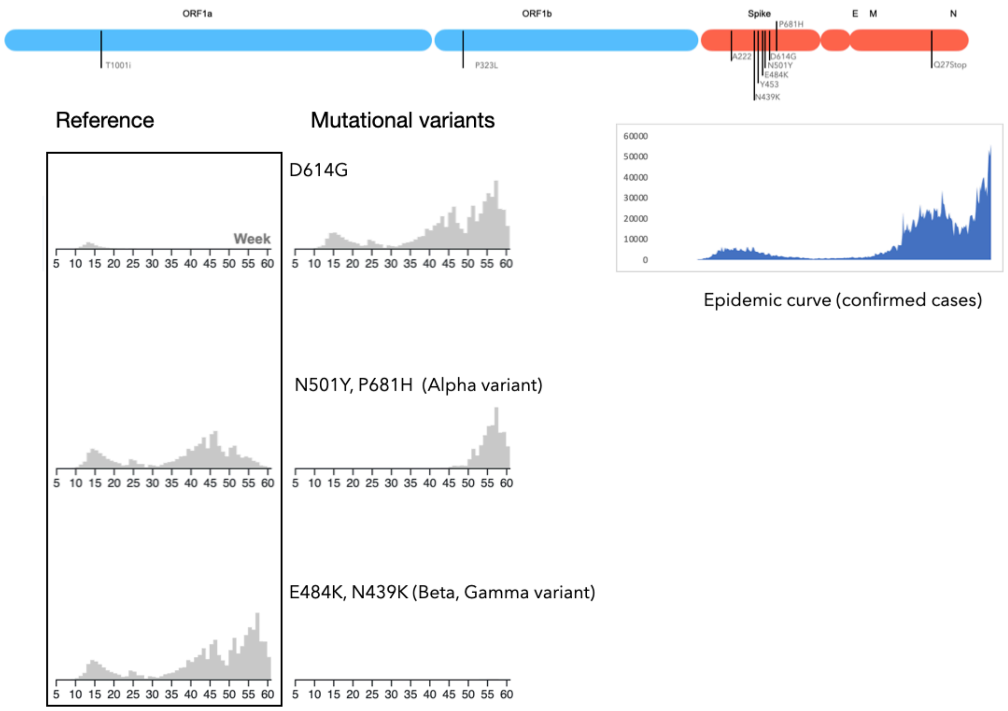


Figure 2. Epicurves, time series of new cases per day, were constructed using reference and mutational variants of coronavirus. Several patterns are observed between reference and mutants, a) extinction and dominance b) emergence c) hitchhiking with driver mutation and co-occurrence. Symmetry of occurrence allows to collapse the mutations to four main patterns. D614 first wave extinction and D614G dominance in the second wave, early dominance in the second wave and decline with the emergence of N501Y.

We then estimated Rv from the variant epicurve of the spike protein mutations to determine the difference in disease transmission in the population. Rv of N501Y is 1.0 (95 CI 0.5-1.6) compared to the reference N501 0.8 (95 CI 0.4-1.1), which indicate a difference of 25-45% transmission within the population, providing a mechanistic basis for measuring the impact of mutations in disease transmission in the population (Supplemental Figure 6). We then applied genomic epidynamics to test the hypothesis that spike protein mutation D614G enhances disease transmission in the people during the first wave of the pandemic. We compared the differences in transmission dynamics prior and during the lockdown, Rv of D614G declined from 2.1 (95 CI 1.1-3.4) to 0.9 (0.5-1.3), and Rv of the reference D614 also declined from 1.9 (95 CI 1.2-2.7) to 1.0 (95 CI 0.7-1.3) (Supplemental Figure 7). This captures the impact of nonpharmaceutical interventions in reducing disease transmission of specific variants. However, estimates of Rv values of D614G and D614 overlap, indicating very similar transmission rates, indicating that the D614G mutation did not significantly modify disease transmission in the population.

We compared the counterpart population genetic tests of selection as cross-validation for genomic epidynamics using the spike protein mutation D614G as an exemplar. Nonsynonymous to synonymous mutation ratio of 1.2 indicates selection, like the findings of other studies[33](Supplemental Figure 4 & 8). However, this test is limited in capturing temporal dynamics, particularly difference between pre and post-lockdown as it is cross-sectional in operation. Furthermore, it only captures the information relative to the mutations, ignoring any quantitative input from the reference. This limits the capability of population genetic tests to be used as inference of transmission between variants. Using a temporal-based approach like a time-series chi-square statistic for D614G mutation frequency indicates no significant difference between expected frequency ratios before lockdown but showed a significant difference in expected frequencies post lockdown (P<.01) (Supplemental Table 1). While Muller plots of SAR-CoV-2 variants allowed visualization of the temporal dynamics and identified variants showing clonal interference (Supplemental Figure 5), were associated with such plots. Hence population based genetic-based measures of selection are not suitable as inferences of disease transmissions, which highlights the value of epidemiology first approach like genomic epidynamics for pathogen WGS analysis.

We then tested the hypothesis that RBD mutations of the spike protein of SARS-CoV-2 reduce vaccine effectiveness in the population. We utilized 500,000 WGS to compute the Rv of N501Y and derived a variant-specific herd immunity threshold. The herd immunity threshold is the percentage of the population needed to be vaccinated or exposed to the disease controlled to control the epidemic. The formula for herd immunity threshold (HIT) = 1- (1/R0) and is based on R0. Hence, we propose a modified version using Rv, as variant-specific herd immunity threshold vHIT= 1-(1/Rv). We selected the timeframe of Rv during the emergence period to capture the transmission dynamics like R0. The computed vHIT for N501Y is 50 %. We then compared the disease transmission of N501Y when the vaccination levels of the population reached predicted vHIT levels (Figure 4). We observed a decline in cases associated with N501Y when vaccination coverage for two dosages got> 50 %. Regression analysis of the number of cases with N501Y mutations with the percentage of two dosage vaccine also showed a strong association (R^2=.9). Expectedly, this results in a decrease of Rv from the high estimates of 1.6 to 0.8. Overall, this points to the possibility of achieving herd immunity effect specific to a variant, which is observed as a decline in cases attributed to a variant and is a novel way to measure the impact of vaccination.

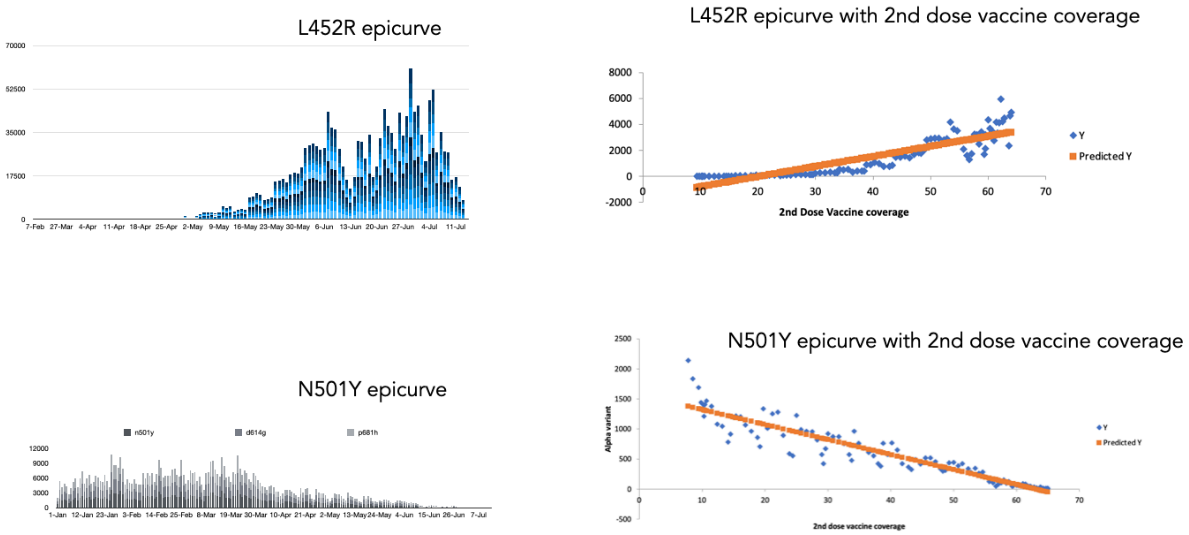


Figure 3. Epicurves of L452R and N501Y and correlation of 2^{nd} dose vaccine coverage with a number of cases.

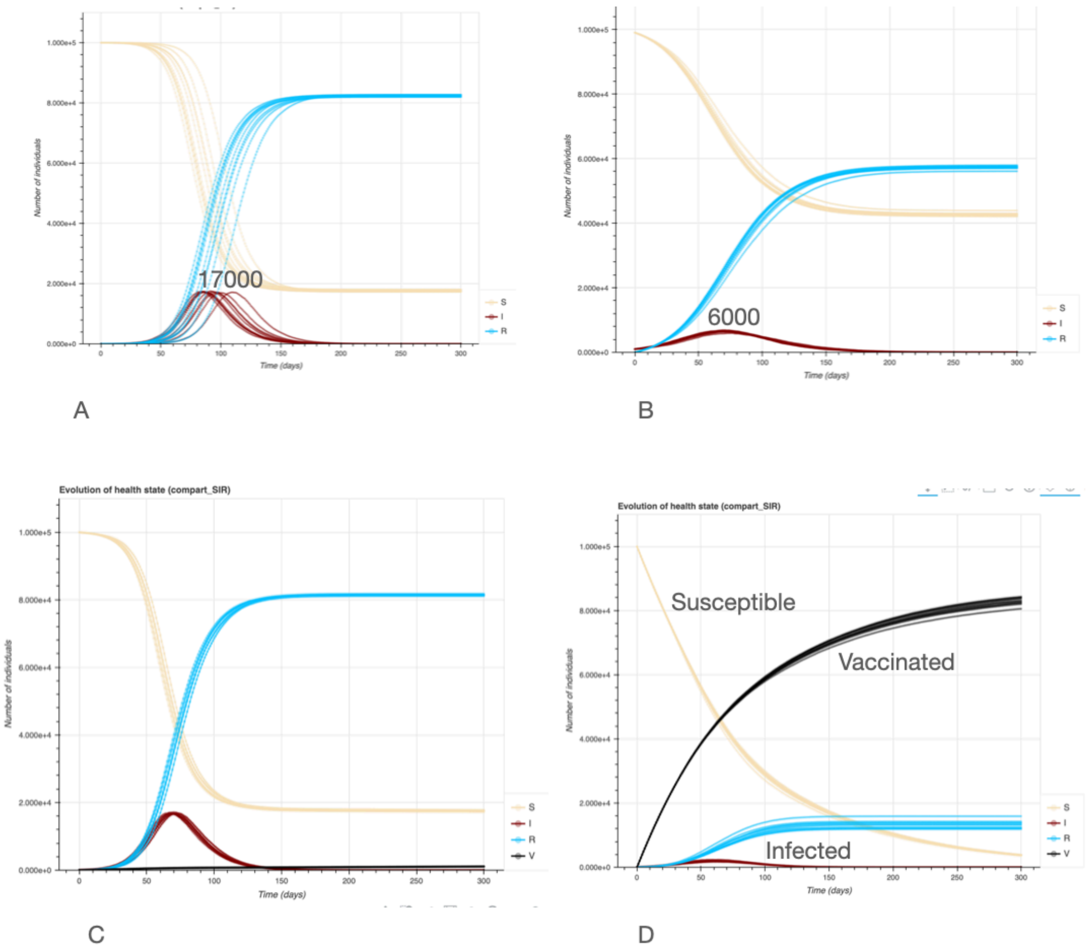


Figure 4. Epidemic models of specific variants using derived reproductive number. A. Epidemic model with a derived reproductive number of L452R B. Epidemic Model with a derived reproductive number from N501Y C. Epidemic model with slow vaccination D. Epidemic model with fast vaccination

However, we then applied the concept of variant-specific herd immunity to RBD mutation in the spike protein L452R to test the hypothesis that mutations in the RBD of the spike protein (L452R) reduce vaccine effectiveness. First, we compared the epicurves of N501Y and L452R relative to 2nd vaccination coverage (Figure 3). L452R epicurve indicated a positive increase as the vaccination coverage increased, while the opposite was observed with N501Y. This uptrend in numbers which parallels the increase in vaccination coverage, indicate a potential reduction of effectiveness of the vaccine against L452R. Furthermore, the L452R Rv = 1.6 is relatively high considering the relatively high vaccination coverage almost reached predicted herd immunity levels. These findings contrast the observed decline in the Rv of N501Y from 1.6 in the 2nd wave to 0.8. Remarkably, the regression of L345R epicurve with vaccination coverage (R^2=0.8) showed a strong relationship. This finding suggests that some method of mutational selection is in place that will produce differences between variant spread and ability to evade vaccination, making it more transmissible.

The estimation of disease transmission using Rv allows simulation of different vaccination coverage, differentiating between industrialized and developing countries. We constructed variant specific SIRV (Susceptible-Infectious-Recovered-Vaccinated) compartmental model with low and high vaccination coverage. Vaccination reduces the number of susceptibles and is predicted to result in lower number of cases. A suboptimal vaccination combined with a variant having a higher Rv will lead to more infections. As herd immunity threshold would take several months to achieve, a combination of intervention (masking and physical distancing) will likely be needed to reduce the transmission of new variants. With the foundation of Rv to precisely follow new or emerging surges in cases, we investigated variant-specific disease transmission by simulating different scenarios using compartmental models based on WGS and Rv.

To estimate the impact of a more transmissible variant, we generated an SIR (Susceptible-Infectious-Recovered) model parameterized with Rv of N501Y and the counterpart parent sequence (Figure 3). The upper estimates range was used for the parent strain (1.1) and the variant (1.6). The difference of reproductive number between the two is 45%, slightly lower than previously estimated using phenomenological methods. The initial condition for the compartmental model is ten infectious individuals in a total of 100,000 population. In an unmitigated scenario (Figure 4A), the peak of infection (7000 cases) will be reached around 90 to 100 days from the introduction, with infection numbers being higher, around 250% for the mutational variant. While 50% will be asymptomatic, peak infection levels indicate hospitalization of 750 individuals at 10 % severe cases. A simulation of a slow vaccination scenario (Figure 4C) indicates similar figures to the unmitigated scenario, while a fast vaccination ratio (Figure 4D) showed a reduction of peak infection from 7000 cases to 2000 cases. Hence even with an available vaccine, it is still essential to combine other measures to decrease the overall number of issues and minimize the introduction of variants with higher transmissibility. These findings indicate that Rv and variant tracking accurately estimate the epidemiological dynamics, which led to investigating mutations in other genes with pharmaceutical interventions implemented via multiple vaccines.

Discussion

We present the conceptual framework of genomic epidynamics, which bridge the population genomics and epidemiology gap. This was executed by converting sequence data of SARS-CoV-2 combined with sampling date into variant epicurves which allow estimation of disease transmission via Rv. The approach addresses the limitations of using pathogen WGS primarily with phylodynamics, which needs a phylogenetic generation step plus a demographic model[42].

The need to be able to perform large-scale analysis of epidemiological insight from sequences will continue to increase, and hence alternative methods of research is important. We demonstrate that bypassing lineage assignment and using mutations to generate variant epicurve could lead to a faster and more scalable disease inference at the population level. We leveraged previous work demonstrating that transmission networks can be derived with common variants combined with sampling dates[115]. We also integrated temporal signals from sampling dates which are not usually incorporated in coalescent based approaches.

Existing methods of estimation of R from epidemic curves lack the information to track "who infected whom" or infection networks and hence either rely on fitting incidence data with the statistical model or calculating the likelihood of infection pairs given temporal proximity[117]. The inclusion of pathogen sequence data with the construction of variant epicurves provides a high-resolution method to track infection networks. This addition enables direct integration of pathogen sequence data with epidemiological methods making it possible to compute disease parameters. While previous studies already demonstrated the possibility of tracking infection networks in small-scale community clusters with shared variants between pathogen isolates, our approach expanded the scale into a significant population-level analysis.

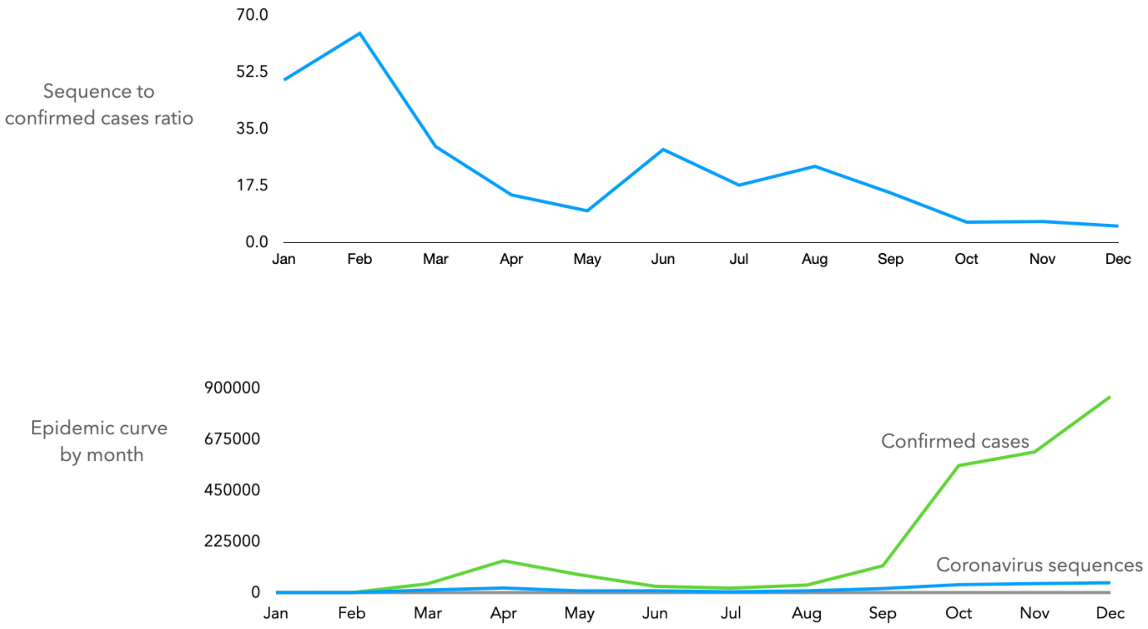
We demonstrated the value of genomic epidynamics by evaluating vaccine effectiveness considering the emergence of spike protein mutations in SARS-CoV-2. While we focused on these sets of mutations as the vaccine design is based on the nucleotide sequence of the spike protein, the co-occurrence of other mutations in other loci can still contribute to overall transmission of the variants. Calculation of Rv of each specific mutation allows a straightforward evaluation of vaccine effectiveness in the population. Using Rv, variability in vaccine effectiveness is observed with simultaneous decline and increase observed with different variants within the population. This work demonstrates genomic epidynamics in monitoring vaccine effectiveness relative to the emergence of variants.

As the vaccine effectiveness demonstrates heterogeneities depending on the variant, a higher vaccination coverage requirement is also the consequence of a more transmissible variant. As there are pockets of virus transmission providing for a reserve pool of mutations, continued genomic surveillance will be pivotal to monitor the rise of new variants. It would need to be integrated into vaccine design. The glaring challenge with this approach is under-investment of most countries in pathogen sequencing for disease detection. The ability to model precision disease transmission is still dependent on the amount of sequencing data generated. Hence developing countries are underrepresented in the public databases while the UK is the primary generator of sequencing data. An interim solution for low sequencing can be addressed partially by generating models to create a simulation of the epidemic spread of specific variants, and our study provides a method for how to implement such an approach.

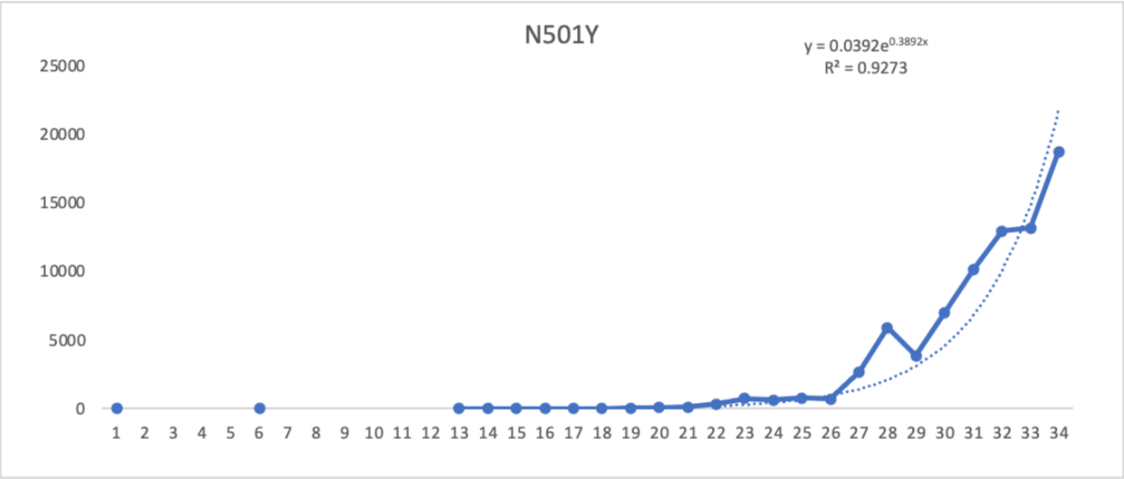
Conclusion

The COVID-19 pandemic generated an unprecedented amount of WGS of a single pathogen. Making sense of large-scale sequencing data is now becoming a role for epidemiologists who need to distill the information to policymakers. Striving disease transmission insights from sequencing data will be pivotal in monitoring the pandemic, generating decision-making criteria for lockdowns and circuit breakers, and monitoring vaccination campaigns. Our approach of deriving epidemiological metrics using extensive scale sequencing data provides a bridge between pathogen evolution and epidemic modeling. It addresses the gap with the use of the reproductive number as a measure of disease spread which is the most used parameter by public health agencies worldwide.

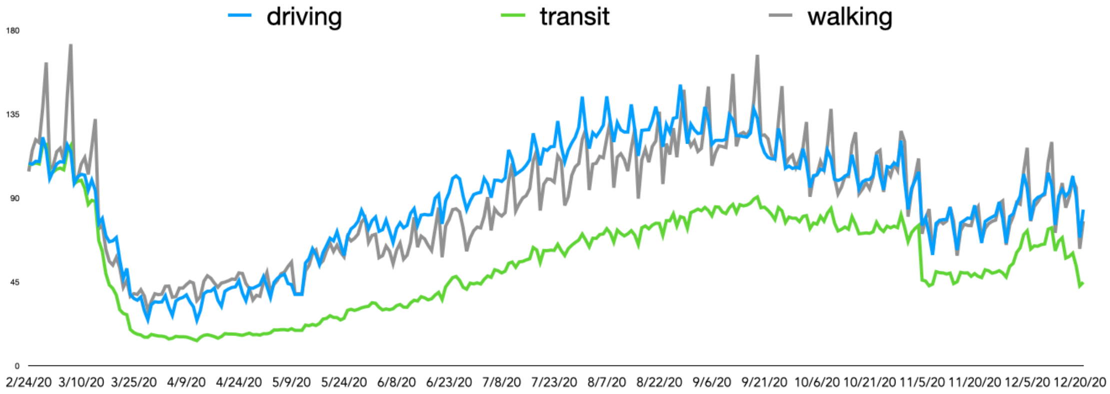
Supplemental Figures



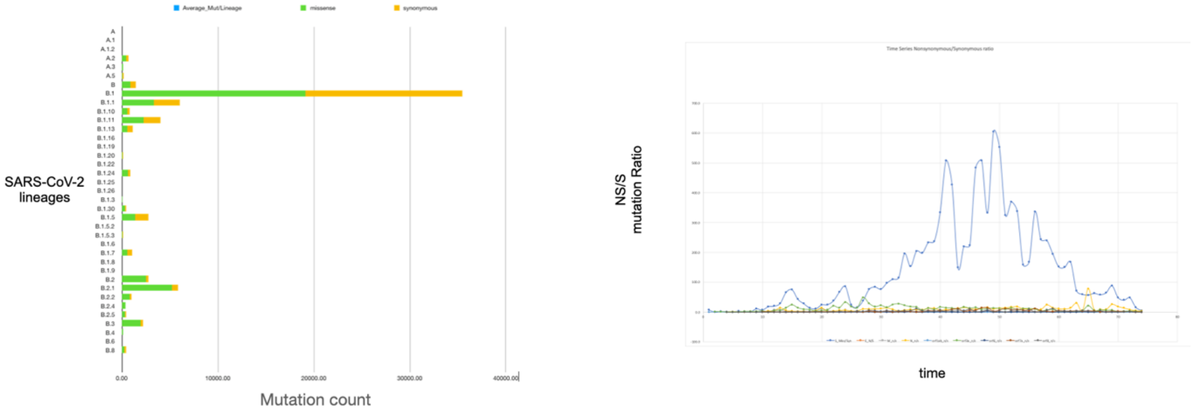
Supplemental Figure 1. The ratio of SARS-CoV-2 whole genome sequences to confirmed cases ratio included in the analysis. The lower figure shows the absolute number of confirmed cases, and the total number of SARS-CoV-2 confirmed cases.



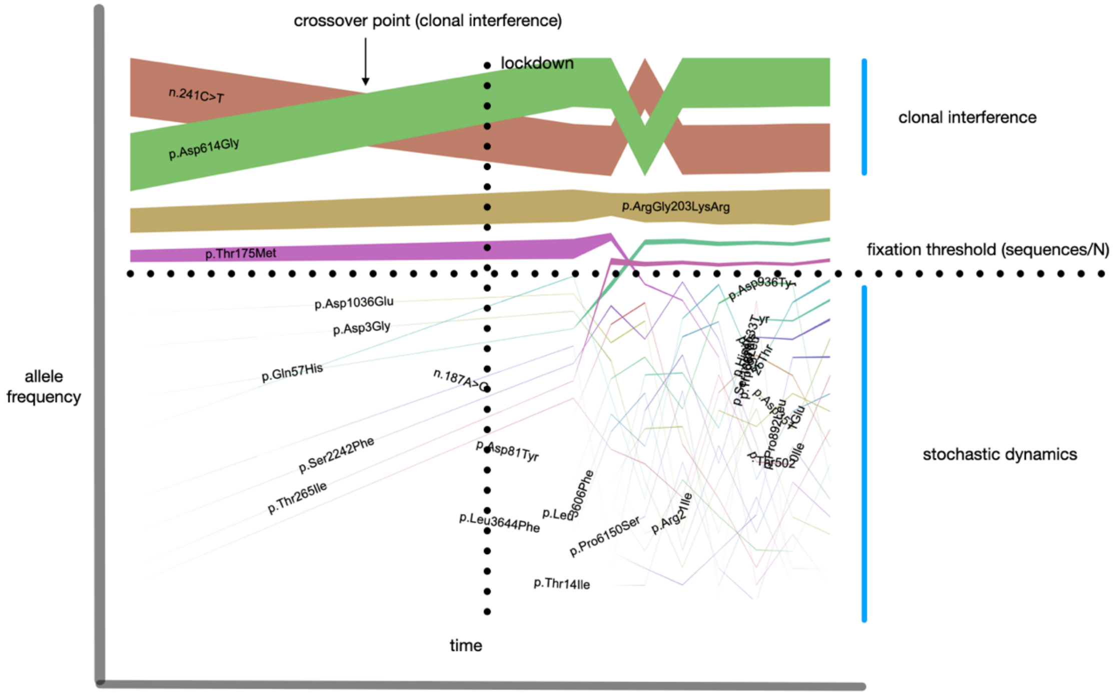
Supplemental Figure 2. Exponential curve fitted to N501Y.



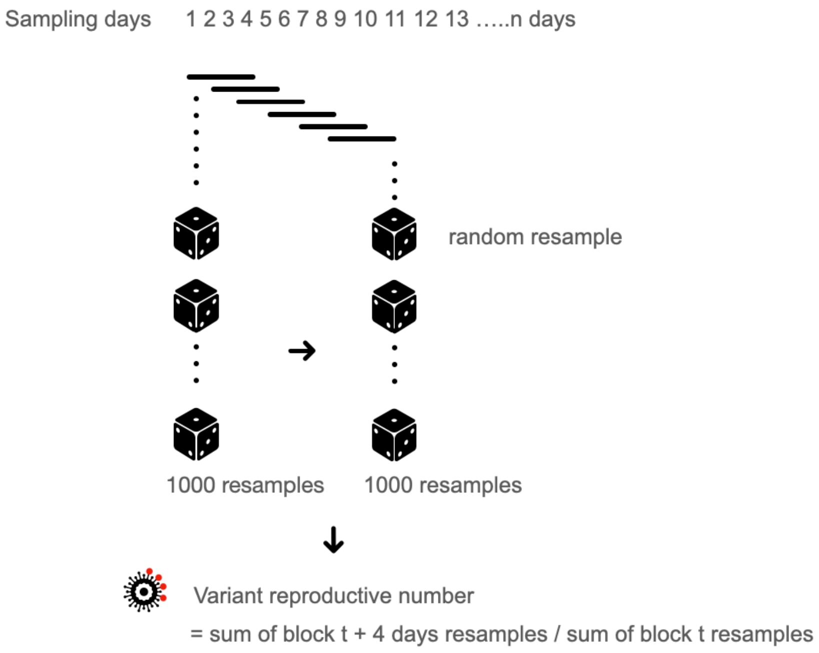
Supplemental Figure 3. Mobility data as a proxy measure for human interaction



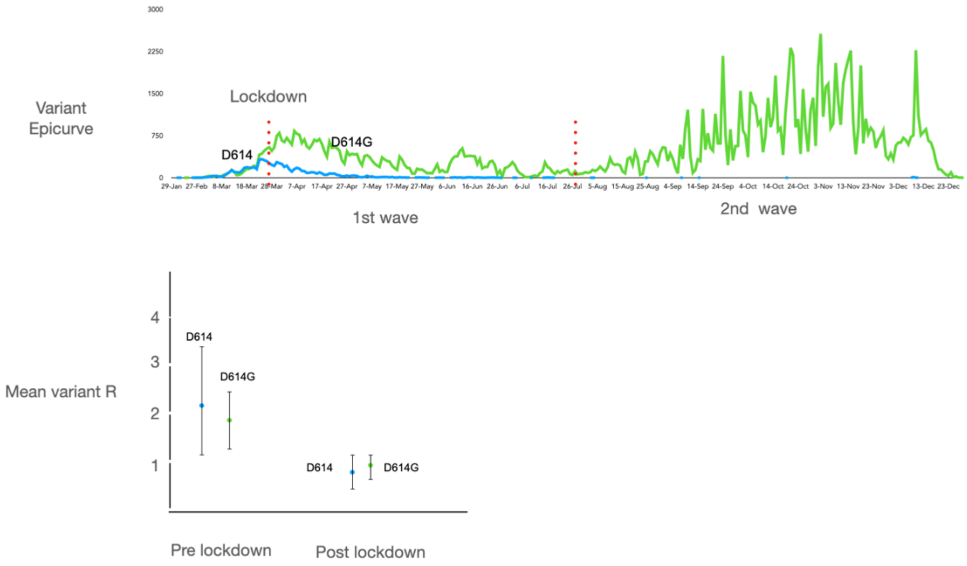
Supplemental Figure 4 Ratio of nonsynonymous and synonymous mutations of the different lineages of Sars-CoV-2.1B Nonsynonymous/synonymous mutations ratio across time. Figure 1C Mutational rank plot of D614G and wildtype mutation across time showing the cross over point



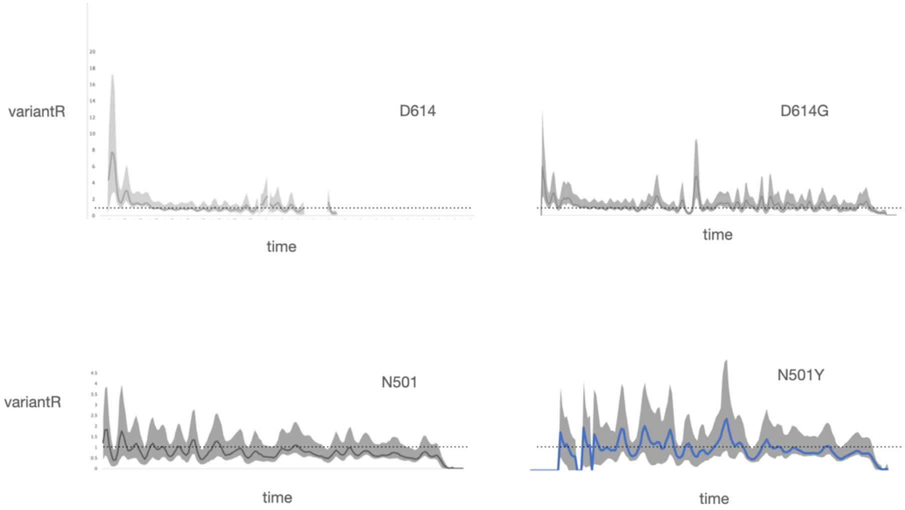
Supplemental Figure 5 Mutation rank allele plot of the dominant genomic lineage of Sars-CoV-2 B1.


Sampling days
1 2 3 4 5 6 7 8 9 10 11 12 13 …..n days
random resample
1000 resamples
1000 resamples
Variant reproductive number
= sum of block t + 4 days resamples / sum of block t resamples

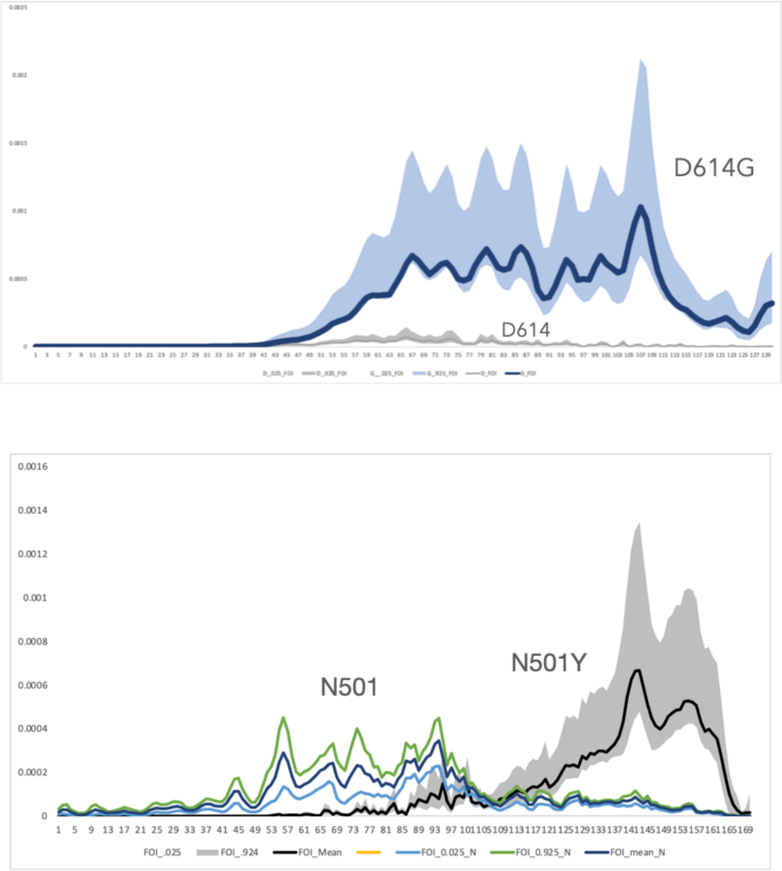
Supplemental Figure 6. Monte Carlo approach to computing the variant reproductive number



Supplemental Figure 7 Variant epicurves between 1st and 2nd wave of infections of COVID-19, variant R comparison between D614 and D614G.



Supplemental Figure 8. Variant R values across time between reference and mutational variant versions of SARS-CoV-2.



Supplemental Figure 9. Force of infection specific to reference and mutational variants of SARS-CoV-2.

Supplemental Table 1. Significance of mutations detected in whole genome sequences of SARS-CoV-2

Mutation	Gene	Significance
D614G	Spike	Increased transmission
E484K	Spike	Reduction in neutralization by monoclonal antibody
N501Y	Spike	Increased transmission and mortality

Y453F	Spike	Mink associated mutation, higher affinity to ACE2 receptor
N439K	Spike	Resistance to antibodies
A222	Spike	Increased transmission
P681H	Spike	Near the furin site
Q27Stop	Orf8	Increased transmission
P323L	RdRp	Associated with disease severity
T1001i	Orf1A	Increased transmission

Supplemental Table 2. Time-Series chi-square statistics for D614G mutation frequency

Month	D_Frequency	G_Frequency	p < 0.01
Feb	0.56	0.44	
March	0.41	0.59	***Not* significant at *p* < 0.01.**
April	0.15	0.85	The chi-square statistic is 36.7071. The *p*-value is < 0.00001. Significant at *p* < .01.The chi-square statistic with Yates correction is 34.9383. The *p*-value is < 0.00001. **Significant at *p* < 0.01**

May	0.05	0.95	The chi-square statistic is 61.3516. The p-value is < 0.00001. Significant at $p < .01$. The chi-square statistic with Yates correction is 58.9692. The p-value is < 0.00001. **Significant at $p < 0.01$**

Supplemental Table 3. Mutation prevalence between 1st and 2nd waves of the COVID-19 Pandemic

	Prevalence	Prevalence_2ndWave
A222_	99.9	63.8
A222V	0.1	36.2
D614_	16.0	0.1
D614G	84.0	99.9
E484_	100.0	99.8
E484K	0.0	0.2
N439_	98.8	98.3
N439K	1.2	1.7
N501_	100.0	54.5
N501Y	0.0	45.5
P323_	16.2	0.1
P323L	83.8	99.9
P681_	100.0	54.9
P681H	0.0	45.1
q27_	99.9	54.7
Q27Stop	0.1	45.3
T1001_	100.0	55.2
T1001I	0.0	44.8
ref_21765_	100.0	54.1
Del_21765_6_	0.0	45.9

Methods

Genomes

SARS-CoV-2 whole genome sequences were downloaded from the COG-UK Consortium site. Variant calling was performed using SNIPPY [109].

Variant Reproductive Number

Mutational variant and temporal metadata were combined to generate variant epicurves. Variant epicurves are then used to estimate the reproductive number of each respective variant by using Monte Carlo approach of random sampling within the four-day moving window. The moving sum of the four-day window was divided the preceding four-day sum to generate variant reproductive number.

Epidemic Modelling

The reproductive number of each variant was used as the model parameter for Susceptible-Infectious-Recovered epidemic model. Infectious period used was 7 days and recovery period were 14 days.

Chapter 5 Pangenome based bacterial species identification and clustering for bacterial population genome analysis in *Hungatella hathewayi* [3-4]

Introduction

Clostridia are a very diverse group of organisms. The taxonomy is in constant revision in light of new whole genome sequence production and genomic flux[118]. While organism classification can be reassigned, the identified isolates within the same species retain their relatedness. In the analysis of 13,151 microbial genomes, the misclassification (18%) was determined by binning into cliques and singletons with ANI data using the Bron-Kerbosch algorithm, which resulted in the misclassification of 31 out of the 445 type strains[119]. The different causes of the type strain misclassification include poor DNA-DNA hybridization (e.g. high genomic diversity), low DNA-DNA hybridization values, naming without referencing to another type strain, and lack of 16s rRNA data. *Hungatella hathewayi,* or its prior designation *Clostridium hathewayi,* was not included in the previous as there were very few *Hungatella* genomes in the time of that publication. As more metagenomes are published increasing claims of finding new organisms are mounting. To this point, Almeida et al. reported an increase of 1952 uncultured organisms

[3-4] Bandoy, D. D. R., B. C. Huang and B. C. Weimer (2019). "Misclassification of a whole genome sequence reference defined by the Human Microbiome Project: a detrimental carryover effect to microbiome studies." medRxiv: 19000489 and Hernández-Juárez LE, Camorlinga M, Méndez-Tenorio A, Calderón JF, Huang BC, Bandoy DDR, Weimer BC, Torres J. Analyses of publicly available *Hungatella hathewayi* genomes revealed genetic distances indicating they belong to more than one species. Virulence. 2021 Dec;12(1):1950-1964. doi: 10.1080/21505594.2021.1950955. PMID: 34304696; PMCID: PMC8312603.

that are not represented in well-studied human populations, where they presented data to support that rare species will be difficult to accurately identify and do not match existing references[120].

Public repositories of genomic data have experienced tremendous expansion beyond human curatorial capacities, which is an ever increasing issue with the high rate of WGS production[121,122]. Recently, it was estimated that ~18% of the organisms are misclassified in microbial genome databases[119]. This high rate of error led to investigation of misclassification of specific organisms, including *Aeromonas*[123] *Fusobacterium*[124], and ultimately entire reference databases[119]. These studies found misclassified type strains, which calls into question the foundation of the taxonomy and inferred relatedness when population genomes are being used for epidemiological purposes, especially with rare organisms that are not well represented in the reference database. The work presented here uniquely identified a misclassified reference species and found propagation of incorrectly labelled genomes in several highly cited microbiome studies[125,126,127,128].

Results

Based on this species delineation notion, we discovered that the Human Microbiome Project reference genome for *Hungatella hathewayi* (WAL18680) was misidentified while building a phylogeny of *Hungatella* species using a population of whole genome sequences[63]. Both 16s rRNA and average nucleotide identity (ANI[119]) analysis indicated that WAL18680 was not a member of the *Hungatella* genus based on genome assessment (Table 1). Population genome comparison analysis was instrumental in discovering that WAL18680 was misclassified and the impact for genomic epidemiology purposes would be important.

Table 1. Average nucleotide identity of *Hungatella* isolates. BCW 8888 is highly similar to other *hathewayi*. WAL 18680 is 71.17 and falls into the cutoff limits for another genus. 2789STDY5834916 ANI 85.62 (BCW 8888) and is a novel species of *Hungatella*.

	BCW_8888	effluvii_DSM_24995	WAL_18680	ve202_11	2789STDY5834916	2789STDY5608850	12489931
BCW_8888	100	94.34	70.85	96.66	85.41	98.05	96.8
effluvii_DSM_24995	94.38	100	70.49	94.51	85.45	94.41	94.93
WAL_18680	**71.17**	**70.78**	**100**	**71.03**	**72.01**	**71.07**	**71.07**
ve202_11	96.63	94.51	70.66	100	85.19	96.51	98.73
2789STDY5834916	85.62	85.59	71.98	85.39	100	85.74	85.83
2789STDY5608850	98.05	94.3	70.65	96.38	85.41	100	96.48
12489931	96.77	94.95	70.82	98.82	85.65	96.68	100

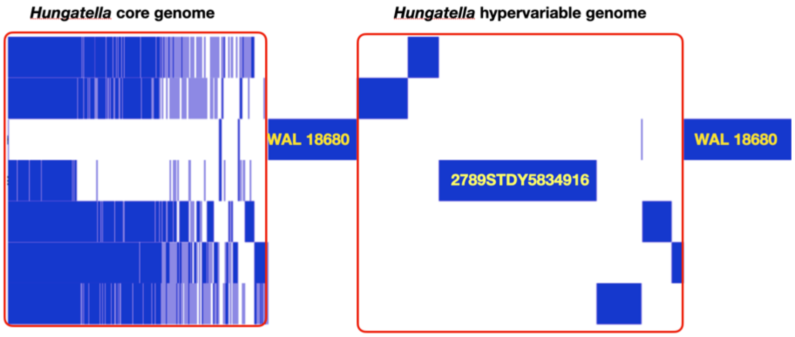


Figure 1. Pangenome of *Hungatella*. WAL18680 was originally identified as ***Clostridium hathewayi.*** After a recent taxonomic reclassification, it was renamed as *Hungatella hathewayi*. **(WAL 18680) does not have the core genome of other *Hungatella* species (***hathewayi* ***or*** *efluvii)* and possess very few core genes common to the other *Hungatella* species. The bulk of its genome is not found in other *Hungatella* species, indicating it belongs to another genus. Strain 2789STDY5834916 is a novel *Hungatella* species.

The misclassified *H. hathewayi* WAL18680 has been used to generate phylogenomic analysis, reference WGS for metagenome analysis, and web server identification platforms utilizing the metagenomic classifiers[127,129,130]. Epidemiologically, association with clinical disease will be discordant with genomic data and result in inaccurate conclusions on the microbiome ecology or therapies based on the microbiome membership to mitigate disease leading to the wrong causal relationship to be concluded[126]. As more microbiome studies are linking rare microbes to biological outcomes, a need exists to quickly identify inaccurate assignment when only a few WGS of individual organisms are available for use as a reference. This creates an issue with low sampling of the genome space for rare organisms and may result in mis-naming based on a small set of phenotypic assays that do not represent the genome content or flux[71].

Discussion

H. hathewayi was first described as an isolate was from human feces[131] and was subsequently reported in a patient with acute cholecystitis, hepatic abscess, and bacteremia[132,133]. It was also later reported in a case of appendicitis[134]. *H. hathewayi* is (WAL18680) one of the designated reference strains in Human Microbiome Project and is used extensively for binning and classification of microbiome related studies, which confounds analysis of the genus *Hungatella*. This organism can be isolated from the microbiome depending on the enrichment conditions[126]. Having a reference species misclassified is detrimental to microbiome research and in epidemiological investigations. To solve this issue, we developed a heurist to minimizing misclassification for rare reference species as a result of cross-validation of the genomic information for name assignment.

The standard procedure of the 100K Pathogen Genome Sequencing Project[4,122,135-137] determines the identity of bacterial pathogen isolates in clinical samples using WGS and the genome distance (ANI[138,139]) before proceeding with additional comparisons. This analysis was done with a group of isolates from suspected *Clostridioides difficile* infection cases. We identified a species of *H. hathewayi* using genome distance using the entire genome sequence that was implemented for high dimensional comparison using MASH[140] (with the maximum sketch size). This was coupled to comparison of all of the available WGS to represent the entire genome diversity to build a whole genome phylogeny [141] to determine the naming accuracy of the clinical isolates . Unexpectedly, one particular sequence was well beyond the species ANI threshold for *C. difficile*. We found that based on ANI, is a putative new species of *Hungatella* (strain 2789STDY5834916). Weis et al.[103,142] used this method with *Campylobacter* species to demonstrate that genome distance accurately estimates host-specific genotypes, zoonotic genotypes, and disease within livestock disease with validated reference genomes. While ANI was the first estimate to raise questions for the accurate identification of this organism, we proceeded with a cross-validation strategy to verify the potential misclassification of the reference species.

We advanced with the initial mis-identification by determining the pangenome analysis with the hypothesis that outbreak isolates would cluster together based on the isolate origin (i.e. an individual or location)[63] as well as contain the same core genome. We found that WAL18680 did not contain any of the core genome relative to all of the other *Hungatella* genomes (Figure 1). Together, these genomic metrics prove that this reference genome was misclassified, which has extensive implications as reference sequences are commonly used for genomic identity for outbreak investigations. Additionally, metagenome studies require reference genome databases to identify bacterial community members. This result indicates that if the epidemiological workflow did not include specific whole genome alignment, inaccurate conclusions and misleading deductions will be made – as was observed by Kaufman et al.[71] – where they found that genome diversity is unexpectedly large and expands based on a power law with each new WGS that is added to the database. Combining the fact that this is a reference genome from a rare organism from a very diverse group, that the genome evolution rate is a power law, and that this is a reference genome from the Human Genome Project the implications for the misidentification have far reaching implications.

Conclusion

Conflicts of taxonomic classification based on traditional methods, such as phenotypic assays, metabolism, with genomic based parameters will likely increase as more genomes are produced and use of the entire genetic potential (i.e. the entire genome). The need for heuristical indicators of misclassification are needed as is the need to expand WGS that adequately represent bacterial diversity among and within taxonomy to represent the genetic diversity of any single organism.

Methods

Whole genome sequences of the genus *Hungatella* were downloaded from NCBI GenBank on May 1, 2019 [143]. Whole genome sequences were annotated using Prokka 1.13.3 and pangenome analysis was performed with Roary 3.12.0 , visualized with Phandango and manually curated [44,144]. Average nucleotide identity (ANI) was estimated using a digital DNA-DNA hybridization approach.

Chapter 6. Conclusion

This dissertation aimed to determine genomic variants that drive antimicrobial resistance, virulence, and transmissibility of pathogens using population wide WGS (Table 11.). With the application of automated machine learning, population genomics and epidemic modeling, the genomic variants have been identified and ranked and further disease parameters defined. Genomic epidynamics is a phylogenetic free approach in measuring disease transmissibility scaled to several hundreds of thousands of WGS, providing a method to measure disease transmission of genomics variants of SARS-CoV-2. The COVID-19 pandemic ushered a new era in infectious disease genomics characterized by data driven demand for public health intervention. Pivotal to the public health decisions is assimilation of predictive models integrating scenarios generated by epidemic models. Linking pathogen genome data with epidemic models opens opportunities to bridge population genomics with genomic epidemiology. While phylogenetics attempts to connect these two fields, scalability is a huge impediment. Furthermore, phylogenetics lack an intuitive input for important epidemiological parameters particularly temporal signature. With genomic epidynamics, WGS data is effectively combined with temporal metadata which enables epidemic modeling of specific genomic variants. A consequence for such a strategy is the ability to analyze the impact of vaccination relative to specific variants. This approach provided a way to determine the relationship of herd immunity levels specific to genomic variants in a scalable manner. Therefore, the proliferation of novel variants would need constant surveillance of changes in transmissibility and hence highlights the value of this approach in measuring disease transmission dynamics using population WGS of pathogens.

Another discovery in this dissertation is the formulation of genomic MICs, which combined automated machine learning prioritized variants and correlated resistance gene presence with MIC. This approach validated a set of resistance genes using variable importance ranking, providing a path to whitebox machine learning models. Hence, genomic variants important for antimicrobial resistance have been identified in *Salmonella* Dublin but can be expanded to other bacterial species. One important feature of this approach is the overall improvement in resistance prediction using databases. This is manifested with the resolution of several discordant phenotype-genotype combination between antibiotic resistance phenotype and presence of resistance genes.

Fundamental limitations serve as challenges in extrapolating important disease parameters from WGS, including limitations in testing AMR resistance particularly the range of antibiotics, inadequate resolving capabilities of variant calling, which impacted feature engineering and the complexity of multidrug resistance. Further broadening the genomic space by further sequencing to cover more of the underlying genomic diversity could improve the predictive capabilities within the AMR databases. Another issue is the lack of transparency with metadata sharing as WGS are published with very limited access to pertinent clinical information. This significantly hindered further downstream analysis which could have been very useful for public health decisions.

Table 1. Dissertation chapter hypotheses and conclusions.

	Hypothesis	Conclusion
Chapter 1	Genomic variants drive virulence in abortive phenotype in *Campylobacter jejuni.*	Machine learning approach with pathogen genome wide association identifies genomic variants of porA driving abortion in *Campylobacter jejuni.*
Chapter 2	Genomic variants drive antimicrobial resistant in *Salmonella enterica subsp. enterica serovar* Dublin.	Automated machine learning combined with genomic MICs define antimicrobial resistance and susceptibility in *Salmonella* Dublin.
Chapter 3	Genomic variation expands with disease transmission of SARS-CoV-2.	Genome identity measures pathogen variation as function of temporal transmission SARS-CoV-2.
Chapter 4	Genomic variants drive transmission dynamics of SARS-CoV-2.	Genomic epidynamics is a scalable approach to measuring variant specific disease transmission dynamics of SARS-CoV-2.
Chapter 5	Pangenome defines species membership of *Hungatella hathewayi.*	*Hungatella* reference species are misidentified due to systemic errors in phylogenetic approach in species clustering.

References

1 Black, A., MacCannell, D. R., Sibley, T. R. & Bedford, T. Ten recommendations for supporting open pathogen genomic analysis in public health. *Nat Med* **26**, 832-841, doi:10.1038/s41591-020-0935-z (2020).

2 Tettelin, H. & Medini, D. The pangenome: Diversity, dynamics and evolution of genomes. (2020).

3 Kaufman, J. H., Christopher A. Elkins, Matthew Davis, Allison M. Weis, Bihua C. Huang, Mark K. Mammel, Isha R. Patel, Kristen L. Beck, Stefan Edlund, David Chambliss, Judith Douglas, Simone Bianco, Mark Kunitomi, Bart C. Weimer. in *Microbial Ecology: Current Advances from Genomics, Metagenomics and Other Omics* (ed Diana Marco) Ch. pp. 45-64., (Caister Academic Press, 2019).

4 Kaufman, J., Ed Seabolt, Mark Kunitomi, Akshay Agarwal, Kristen Beck, Harsha Krishnareddy and **Bart C. Weimer** . Exploiting Functional Context in Biology: Reconsidering Classification of Bacterial Life. *IEEE Computer Science* 17-20, doi:DOI 10.1109/ICDEW.2018.00009 (2018).

5 Weimer, B. C. *et al.* Defining the food microbiome for authentication, safety, and process management. *IBM Journal of Research and Development* **60** (2016).

6 Young, A. D. & Gillung, J. P. Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology* **45**, 225-247, doi:https://doi.org/10.1111/syen.12406 (2020).

7 Khoury, M. J., Millikan, R., Little, J. & Gwinn, M. The emergence of epidemiology in the genomics age. *Int J Epidemiol* **33**, 936-944, doi:10.1093/ije/dyh278 (2004).

8 San, J. E. *et al.* Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol* **10**, 3119, doi:10.3389/fmicb.2019.03119 (2019).

9 Allen, J. P., Snitkin, E., Pincus, N. B. & Hauser, A. R. Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning. *Trends Microbiol* **29**, 621-633, doi:10.1016/j.tim.2020.12.002 (2021).

10 Young, B. C. *et al.* Panton-Valentine leucocidin is the key determinant of Staphylococcus aureus pyomyositis in a bacterial GWAS. *Elife* **8**, doi:10.7554/eLife.42486 (2019).

11 Lees, J. A. *et al.* Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nat Commun* **10**, 2176, doi:10.1038/s41467-019-09976-3 (2019).

12 Gori, A. *et al.* Pan-GWAS of Streptococcus agalactiae Highlights Lineage-Specific Genes Associated with Virulence and Niche Adaptation. *mBio* **11**, doi:10.1128/mBio.00728-20 (2020).

13 Berthenet, E. *et al.* A GWAS on Helicobacter pylori strains points to genetic variants associated with gastric cancer risk. *BMC Biol* **16**, 84, doi:10.1186/s12915-018-0550-3 (2018).

14 Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol* **16**, e2002985, doi:10.1371/journal.pbio.2002985 (2018).

15 Fisher, R. A. *The genetical theory of natural selection.* (The Clarendon press, 1930).

16 Couce, A. & Tenaillon, O. A. The rule of declining adaptability in microbial evolution experiments. *Front Genet* **6**, 99, doi:10.3389/fgene.2015.00099 (2015).

17 Hendriksen, R. S. *et al.* Using Genomics to Track Global Antimicrobial Resistance. *Front Public Health* **7**, 242, doi:10.3389/fpubh.2019.00242 (2019).

18 Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* **48**, D517-D525, doi:10.1093/nar/gkz935 (2020).

19 Bortolaia, V. *et al.* ResFinder 4.0 for predictions of phenotypes from genotypes. *J Antimicrob Chemother* **75**, 3491-3500, doi:10.1093/jac/dkaa345 (2020).

20 Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* **74**, 417-433, doi:10.1128/MMBR.00016-10 (2010).

21 Doyle, R. M. *et al.* Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study. *Microb Genom* **6**, doi:10.1099/mgen.0.000335 (2020).

22 Kavvas, E. S. *et al.* Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nat Commun* **9**, 4306, doi:10.1038/s41467-018-06634-y (2018).

23 Hyun, J. C., Kavvas, E. S., Monk, J. M. & Palsson, B. O. Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens. *PLoS Comput Biol* **16**, e1007608, doi:10.1371/journal.pcbi.1007608 (2020).

24 Nguyen, M. *et al.* Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal <i>Salmonella</i>. *Journal of Clinical Microbiology* **57**, e01260-01218, doi:doi:10.1128/JCM.01260-18 (2019).

25 Macesic, N. *et al.* Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data. *mSystems* **5**, e00656-00619, doi:doi:10.1128/mSystems.00656-19 (2020).

26 Avershina, E. *et al.* AMR-Diag: Neural network based genotype-to-phenotype prediction of resistance towards beta-lactams in Escherichia coli and Klebsiella pneumoniae. *Comput Struct Biotechnol J* **19**, 1896-1906, doi:10.1016/j.csbj.2021.03.027 (2021).

27 Jaillard, M., Palmieri, M., van Belkum, A. & Mahe, P. Interpreting k-mer-based signatures for antibiotic resistance prediction. *Gigascience* **9**, doi:10.1093/gigascience/giaa110 (2020).

28 Her, H. L. & Wu, Y. W. A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains. *Bioinformatics* **34**, i89-i95, doi:10.1093/bioinformatics/bty276 (2018).

29 Liu, Z. *et al.* Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. *Frontiers in Microbiology* **11**, doi:10.3389/fmicb.2020.00048 (2020).

30 van Belkum, A. *et al.* Developmental roadmap for antimicrobial susceptibility testing systems. *Nat Rev Microbiol* **17**, 51-62, doi:10.1038/s41579-018-0098-9 (2019).

31 Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327-332, doi:10.1126/science.1090727 (2004).

32 Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet* **10**, 540-550, doi:10.1038/nrg2583 (2009).

33 Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, doi:10.1126/science.abg3055 (2021).

34 Zhao, Z., Sokhansanj, B. A., Malhotra, C., Zheng, K. & Rosen, G. L. Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. *PLoS Comput Biol* **16**, e1008269, doi:10.1371/journal.pcbi.1008269 (2020).

35 De Maio, N., Worby, C. J., Wilson, D. J. & Stoesser, N. Bayesian reconstruction of transmission within outbreaks using genomic variants. *PLoS Comput Biol* **14**, e1006117, doi:10.1371/journal.pcbi.1006117 (2018).

36 Alamil, M. *et al.* Inferring epidemiological links from deep sequencing data: a statistical learning approach for human, animal and plant diseases. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180258, doi:10.1098/rstb.2018.0258 (2019).

37 Jombart, T. *et al.* Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol* **10**, e1003457, doi:10.1371/journal.pcbi.1003457 (2014).

38 Campbell, F., Strang, C., Ferguson, N., Cori, A. & Jombart, T. When are pathogen genome sequences informative of transmission events? *PLoS Pathog* **14**, e1006885, doi:10.1371/journal.ppat.1006885 (2018).

39 Stimson, J. *et al.* Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Molecular Biology and Evolution* **36**, 587-603, doi:10.1093/molbev/msy242 (2019).

40 Ypma, R. J. *et al.* Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *J Infect Dis* **207**, 730-735, doi:10.1093/infdis/jis757 (2013).

41 Worby, C. J., Chang, H. H., Hanage, W. P. & Lipsitch, M. The distribution of pairwise genetic distances: a tool for investigating disease transmission. *Genetics* **198**, 1395-1404, doi:10.1534/genetics.114.171538 (2014).

42 Pybus, O. G. *et al.* The epidemic behavior of the hepatitis C virus. *Science* **292**, 2323-2325, doi:DOI 10.1126/science.1058321 (2001).

43 Szarvas, J. *et al.* Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform. *Commun Biol* **3**, 137, doi:10.1038/s42003-020-0869-5 (2020).

44 Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693, doi:10.1093/bioinformatics/btv421 (2015).

45 Power, R. A., Parkhill, J. & de Oliveira, T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet* **18**, 41-50, doi:10.1038/nrg.2016.132 (2017).

46 Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724, doi:10.1186/1471-2164-11-724 (2010).

47 Breiman, L. Statistical Modeling: The Two Cultures. *Statistical Science* **16**, doi:10.1214/ss/1009213726 (2001).

48 Read, T. D. & Massey, R. C. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Med* **6**, doi:ARTN 109. 10.1186/s13073-014-0109-z (2014).

49 Shapiro, B. J. *et al.* Population genomics of early events in the ecological differentiation of bacteria. *Science* **336**, 48-51, doi:10.1126/science.1218198 (2012).

50 Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141-161, doi:10.1007/s10142-015-0433-4 (2015).

51 Bobay, L. M. & Ochman, H. The Evolution of Bacterial Genome Architecture. *Front Genet* **8**, 72, doi:10.3389/fgene.2017.00072 (2017).

52 Martinez-Carranza, E. *et al.* Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case of Escherichia coli. *Front Microbiol* **9**, 1059, doi:10.3389/fmicb.2018.01059 (2018).

53 Weis, A. M., Clothier, K. A., Huang, B. C., Kong, N. & Weimer, B. C. Draft Genome Sequences of Campylobacter jejuni Strains That Cause Abortion in Livestock. *Genome Announc* **4**, doi:10.1128/genomeA.01324-16 (2016).

54 Weis, A. M. *et al.* Genomic Comparison of Campylobacter spp. and Their Potential for Zoonotic Transmission between Birds, Primates, and Livestock. *Appl Environ Microbiol* **82**, 7165-7175, doi:10.1128/AEM.01746-16 (2016).

55 Wu, Z. *et al.* Point mutations in the major outer membrane protein drive hypervirulence of a rapidly expanding clone of Campylobacter jejuni. *Proc Natl Acad Sci U S A* **113**, 10690-10695, doi:10.1073/pnas.1605869113 (2016).

56 Chen, T. & Guestrin, C. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* 785-794 (2016).

57 Behravan, H. *et al.* Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Sci Rep* **8**, 13149, doi:10.1038/s41598-018-31573-5 (2018).

58 Jason H. Yang, S. N. W., Meagan Hamblin, Douglas McCloskey, Miguel A. Alcantar, Lars Schrübbers, Allison J. Lopatkin, Sangeeta Satish, Amir Nili, Bernhard O. Palsson, Graham C. Walker, James J. Collins. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell* **177**, 1649-1661, doi:https://doi.org/10.1016/j.cell.2019.04.016 (2019).

59 Heijne, I. N. a. G. Fine-tuning the topology of a polytopic membrane protein: Role of positively and negatively charged amino acids. *Cell* **62**, 1135-1141, doi:https://doi.org/10.1016/0092-8674(90)90390-Z (1990).

60 Elazar, A., Weinstein, J. J., Prilusky, J. & Fleishman, S. J. Interplay between hydrophobicity and the positive-inside rule in determining membrane-protein topology. *Proc Natl Acad Sci U S A* **113**, 10340-10345, doi:10.1073/pnas.1605888113 (2016).

61 Kim, C. *et al.* Basic amino-acid side chains regulate transmembrane integrin signalling. *Nature* **481**, 209-213, doi:10.1038/nature10697 (2011).

62 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

63 Bandoy, D. Pangenome guided pharmacophore modelling of enterohemorrhagic Escherichia coli sdiA. *F1000Research* doi:https://doi.org/10.12688/f1000research.17620.1 (2019).

64 Brown, T. a. I., Luis. sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software* **1**, 27, doi:10.21105/joss.00027 (2016).

65 Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).

66 Bienert, S. *et al.* The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res* **45**, D313-D319, doi:10.1093/nar/gkw1132 (2017).

67 Miller, J. J. *et al.* Phylogenetic and Biogeographic Patterns of Vibrio parahaemolyticus Strains from North America Inferred from Whole-Genome Sequence Data. *Appl Environ Microbiol* **87**, doi:10.1128/AEM.01403-20 (2021).

68 Bandoy, D. D. Large scale enterohemorrhagic E coli population genomic analysis using whole genome typing reveals recombination clusters and potential drug target. *F1000Res* **8**, 33, doi:10.12688/f1000research.17620.3 (2019).

69 Bandoy, D. D. R. & Weimer, B. C. Biological Machine Learning Combined with Campylobacter Population Genomics Reveals Virulence Gene Allelic Variants Cause Disease. *Microorganisms* **8**, doi:10.3390/microorganisms8040549 (2020).

70 Bottery, M. J., Pitchford, J. W. & Friman, V. P. Ecology and evolution of antimicrobial resistance in bacterial communities. *ISME J* **15**, 939-948, doi:10.1038/s41396-020-00832-7 (2021).

71 Kaufman, J. H., Christopher A. Elkins, Matthew Davis, Allison M Weis, Bihua C. Huang, Mark K Mammel, Isha R. Patel, Kristen L. Beck, Stefan Edlund, David Chambliss, Simone Bianco, Mark Kunitomi, Bart C. Weimer. Microbiogeography and microbial genome evolution. *arXiv*, 1703.07454 (2017). <https://arxiv.org/abs/1703.07454>.

72 Andreu-Sanchez, S. *et al.* A Benchmark of Genetic Variant Calling Pipelines Using Metagenomic Short-Read Sequencing. *Front Genet* **12**, 648229, doi:10.3389/fgene.2021.648229 (2021).

73 Fenske, G. J., Thachil, A., McDonough, P. L., Glaser, A. & Scaria, J. Geography Shapes the Population Genomics of Salmonella enterica Dublin. *Genome Biol Evol* **11**, 2220-2231, doi:10.1093/gbe/evz158 (2019).

74 Liu, Z. *et al.* Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. *Front Microbiol* **11**, 48, doi:10.3389/fmicb.2020.00048 (2020).

75 Lewin-Epstein, O., Baruch, S., Hadany, L., Stein, G. Y. & Obolski, U. Predicting antibiotic resistance in hospitalized patients by applying machine learning to electronic medical records. *Clin Infect Dis*, doi:10.1093/cid/ciaa1576 (2020).

76 Tyson, G. H. *et al.* Establishing Genotypic Cutoff Values To Measure Antimicrobial Resistance in Salmonella. *Antimicrob Agents Chemother* **61**, doi:10.1128/AAC.02140-16 (2017).

77 Mangat, C. S. *et al.* Genomic Investigation of the Emergence of Invasive Multidrug-Resistant Salmonella enterica Serovar Dublin in Humans and Animals in Canada. *Antimicrob Agents Chemother* **63**, doi:10.1128/AAC.00108-19 (2019).

78 Weigel LM, S. C., Tenover FC. gyrA mutations associated with fluoroquinolone resistance in eight species of Enterobacteriaceae. *Antimicrob Agents Chemother*, doi:doi:10.1128/AAC.42 (1998).

79 Fricke, W. F. & Rasko, D. A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet* **15**, 49-55, doi:10.1038/nrg3624 (2014).

80 Yang, J. H. *et al.* A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell* **177**, 1649-1661 e1649, doi:10.1016/j.cell.2019.04.016 (2019).

81 Kankainen, M., Ojala, T. & Holm, L. BLANNOTATOR: enhanced homology-based function prediction of bacterial proteins. *BMC Bioinformatics* **13**, 33, doi:10.1186/1471-2105-13-33 (2012).

82 LeDelll, E. H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning (AutoML)* (2020).

83 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565-574, doi:10.1016/s0140-6736(20)30251-8 (2020).

84 Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **382**, 727-733, doi:10.1056/NEJMoa2001017 (2020).

85 Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *The Lancet* **395**, 689-697, doi:10.1016/s0140-6736(20)30260-9 (2020).

86 Heesterbeek, J. A. P. & Dietz, K. The concept of Ro in epidemic theory. *Statistica Neerlandica* **50**, doi:doi:10.1111/j.1467-9574.1996.tb01482.x.

87 Delamater, P. L., Street, E. J., Leslie, T. F., Yang, Y. T. & Jacobsen, K. H. Complexity of the Basic Reproduction Number (R0). *Emerg Infect Dis* **25**, 1-4, doi:10.3201/eid2501.171901 (2019).

88 Fine, P., Eames, K. & Heymann, D. L. "Herd immunity": a rough guide. *Clin Infect Dis* **52**, 911-916, doi:10.1093/cid/cir007 (2011).

89 Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med*, doi:10.1056/NEJMoa2001316 (2020).

90 Zhao, S. *et al.* The basic reproduction number of novel coronavirus (2019-nCoV) estimation based on exponential growth in the early outbreak in China from 2019 to 2020: A reply to Dhungana. *Int J Infect Dis*, doi:10.1016/j.ijid.2020.02.025 (2020).

91 Wallinga, J. & Teunis, P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology* **Volume 160**, Pages 509–516, doi: https://doi.org/10.1093/aje/kwh255.

92 Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am J Epidemiol* **178**, 1505-1512, doi:10.1093/aje/kwt133 (2013).

93 Bedford, T. *Phylodynamic estimation of incidence and prevalence of novel coronavirus (nCoV) infections through time*, <https://github.com/blab/ncov-phylodynamics> (2020).

94 Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol* **31**, 1869-1879, doi:10.1093/molbev/msu121 (2014).

95 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, doi:10.1126/science.1259657 (2014).

96 Cotten, M. *et al.* Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *The Lancet* **382**, 1993-2002, doi:10.1016/s0140-6736(13)61887-5 (2013).

97 Control, K. C. f. D. *The update of COVID-19 in ROK*, <https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030> (2020).

98 Niehus, R., De Salazar, P. M., Taylor, A. & Lipsitch, M. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers *medrkiv*, doi:10.1101/2020.02.13.20022707 (2020).

99 Eppinger, M. *et al.* Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *mBio* **5**, e01721, doi:10.1128/mBio.01721-14 (2014).

100 Weis, A. M. *et al.* Large-Scale Release of Campylobacter Draft Genomes: Resources for Food Safety and Public Health from the 100K Pathogen Genome Project. *Genome Announc* **5**, doi:10.1128/genomeA.00925-16 (2017).

101 Weimer, B. C. 100K Pathogen Genome Project. *Genome Announc* **5**, doi:10.1128/genomeA.00594-17 (2017).

102 Taff, C. C. *et al.* Influence of Host Ecology and Behavior on Campylobacter jejuni Prevalence and Environmental Contamination Risk in a Synanthropic Wild Bird Species. *Applied and Environmental Microbiology* **82**, 4811-4820, doi:10.1128/aem.01456-16 (2016).

103 Weis, A. M. *et al.* Genomic Comparisons and Zoonotic Potential of Campylobacter Between Birds, Primates, and Livestock. *Applied and environmental microbiology*, 7165-7175, doi:10.1128/AEM.01746-16 (2016).

104 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22**, doi:10.2807/1560-7917.ES.2017.22.13.30494 (2017).

105 Wilder-Smith, A. & Freedman, D. O. Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak. *J Travel Med*, doi:10.1093/jtm/taaa020 (2020).

106 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123, doi:10.1093/bioinformatics/bty407 (2018).

107 Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, doi:10.1016/s1473-3099(20)30120-1 (2020).

108 Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**, D8-D13, doi:10.1093/nar/gkx1095 (2018).

109 Seemann, T. *Rapid haploid variant calling and core genome alignment*, <https://github.com/tseemann/snippy> (2020).

110 Bandoy, D. Large scale enterohemorrhagic E coli population genomic analysis using whole genome typing reveals recombination clusters and potential drug target. *F1000Research* **8**, doi:10.12688/f1000research.17620.1 (2019).

111 Bandoy, D. D. & Weimer, B. C. Analysis of SARS-CoV-2 genomic epidemiology reveals disease transmission coupled to variant emergence and allelic variation. *Sci Rep* **11**, 7380, doi:10.1038/s41598-021-86265-4 (2021).

112 Boehm, E. *et al.* Novel SARS-CoV-2 variants: the pandemics within the pandemic. *Clin Microbiol Infect*, doi:10.1016/j.cmi.2021.05.022 (2021).

113 Abdool Karim, S. S. & de Oliveira, T. New SARS-CoV-2 Variants - Clinical, Public Health, and Vaccine Implications. *N Engl J Med* **384**, 1866-1868, doi:10.1056/NEJMc2100362 (2021).

114 Vaughan, T. G. *et al.* Estimating Epidemic Incidence and Prevalence from Genomic Data. *Mol Biol Evol* **36**, 1804-1816, doi:10.1093/molbev/msz106 (2019).

115 Worby, C. J., Lipsitch, M. & Hanage, W. P. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *Am J Epidemiol* **186**, 1209-1216, doi:10.1093/aje/kwx182 (2017).

116 Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369-1372, doi:10.1126/science.1259657 (2014).

117 Wallinga, J. & Teunis, P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am J Epidemiol* **160**, 509-516, doi:10.1093/aje/kwh255 (2004).

118 Yutin, N. & Galperin, M. Y. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. *Environ Microbiol* **15**, 2631-2641, doi:10.1111/1462-2920.12173 (2013).

119 Varghese NJ, M. S., Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **Aug 18;43(14):6761-71**, doi:10.1093/nar/gkv657 (2015).

120 Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499-504, doi:10.1038/s41586-019-0965-1 (2019).

121 Weimer, B. C. 100K Pathogen Genome Project. *Genome Announcements*, genomeA.00594-00517, doi:DOI: 10.1128/genomeA.00594-17 (2017).

122 Kong, N. *et al.* Draft Genome Sequences of 1,183 Salmonella Strains from the 100K Pathogen Genome Project. *Genome Announc* **5**, e00518-00537, doi:10.1128/genomeA.00518-17 (2017).

123 Awan, F. *et al.* Comparative genome analysis provides deep insights into Aeromonas hydrophila taxonomy and virulence-related factors. *BMC Genomics* **19**, 712, doi:10.1186/s12864-018-5100-4 (2018).

124 Kook, J., Park, SN., Lim, Y.K. Genome-Based Reclassification of Fusobacterium nucleatum Subspecies at the Species Level. *Current Microbiology* **74**, 1137-1147, doi:https://doi.org/10.1007/s00284-017-1296-9 (29 June 2017).

125 I, R. D. a. T. Comparative Genomic Analysis of the Human Gut Microbiome Reveals a Broad Distribution of Metabolic Pathways for the Degradation of Host-Synthetized Mucin Glycans and Utilization of Mucin-Derived Monosaccharides. *Front. Genet.* **8**, doi:10.3389/fgene.2017.00111 (2017).

126 Atarashi, K. *et al.* Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* **500**, 232-236, doi:10.1038/nature12331 (2013).

127 Sabag-Daigle A, W. J., Borton MA,, Sengupta A, G. V., Wrighton KC, & Wysocki VH, A. B. Identification of bacterial species that can utilize fructoseasparagine. *Appl Environ Microbiol* **84:e01957-17**, doi: 10.1128/AEM.01957-17 (2018).

128 Yu, L. *et al.* Grammar of protein domain architectures. *Proc Natl Acad Sci U S A* **116**, 3636-3645, doi:10.1073/pnas.1814684116 (2019).

129 Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41-49, doi:10.1016/j.ymeth.2013.06.027 (2013).

130 Carrico, J. A., Rossi, M., Moran-Gilad, J., Van Domselaar, G. & Ramirez, M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect* **24**, 342-349, doi:10.1016/j.cmi.2017.12.015 (2018).

131 Steer, T., Collins, M. D., Gibson, G. R., Hippe, H. & Lawson, P. A. Clostridium hathewayi sp. nov., from human faeces. *Syst Appl Microbiol* **24**, 353-357, doi:10.1078/0723-2020-00044 (2001).

132 Kaur, S., Yawar, M., Kumar, P. A. & Suresh, K. Hungatella effluvii gen. nov., sp. nov., an obligately anaerobic bacterium isolated from an effluent treatment plant, and reclassification of Clostridium hathewayi as Hungatella hathewayi gen. nov., comb. nov. *Int J Syst Evol Microbiol* **64**, 710-718, doi:10.1099/ijs.0.056986-0 (2014).

133 Elsayed, S. & Zhang, K. Human infection caused by Clostridium hathewayi. *Emerg Infect Dis* **10**, 1950-1952, doi:10.3201/eid1011.040006 (2004).

134 Woo, P. C. *et al.* Bacteremia due to Clostridium hathewayi in a patient with acute appendicitis. *J Clin Microbiol* **42**, 5947-5949, doi:10.1128/JCM.42.12.5947-5949.2004 (2004).

135 Weis, A. M. *et al.* Large-Scale Release of Campylobacter Draft Genomes: Resources for Food Safety and Public Health from the 100K Pathogen Genome Project. *Genome Announc* **5**, e00925-00916, doi:10.1128/genomeA.00925-16 (2017).

136 Weis, A. M., Bihua C. Huang, Dylan B. Storey, Nguyet Kong, Poyin Chen, Narine Arabyan, Brent Gilpin, Carl Mason, Andrea K. Townsend, Woutrina A. Miller, Barbara A. Byrne, Conor C. Taff, Bart C. Weimer. Large-scale release of Campylobacter draft genomes; resources for food safety and public health from the 100K Pathogen Genome Project. *Genome Announcements* **5**, e00925-00916 (2016).

137 Chen, P. *et al.* 100K Pathogen Genome Project: 306 Listeria Draft Genome Sequences for Food Safety and Public Health. *Genome Announc* **5**, e00967-00916, doi:10.1128/genomeA.00967-16 (2017).

138 Auch, A. F., von Jan, M., Klenk, H. P. & Goker, M. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* **2**, 117-134, doi:10.4056/sigs.531120 (2010).

139 Auch, A. F., Klenk, H. P. & Goker, M. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci* **2**, 142-148, doi:10.4056/sigs.541628 (2010).

140 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).

141 Bandoy, D. Large scale enterohemorrhagic E coli population genomic analysis using whole genome typing reveals recombination clusters and potential drug target. *F1000Research* **8**, doi:10.12688/f1000research.17620.1 (2019).

142 Lawton, S. J. *et al.* Comparative analysis of Campylobacter isolates from wild birds and chickens using MALDI-TOF MS, biochemical testing, and DNA sequencing. *J Vet Diagn Invest* **30**, 354-361, doi:10.1177/1040638718762562 (2018).

143 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **44**, D67-72, doi:10.1093/nar/gkv1276 (2016).

144 Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, doi:10.1093/bioinformatics/btx610 (2017).