Downloaded from orbit.dtu.dk on: Apr 20, 2024

DTU Library

Evergreen methods for phylogeny

Szarvas, Judit

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Szarvas, J. (2021). *Evergreen methods for phylogeny*. Technical University of Denmark.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Ph.D. thesis in Bioinformatics

DTU Food
National Food Institute

Evergreen methods for phylogeny

Judit Szarvas

DTU

Supervisor: Prof. Ole Lund

Co-supervisor: Prof. Frank Aarestrup

DTU Food
National Food Institute
Technical University of Denmark
Kemitorvet
Building 202
2800 Kongens Lyngby, Denmark

Resumé

Fremkomsten og tilpasningen af nye teknologier i DNA-sekventering i det sidste årti førte til en bredere anvendelse af sekventering. Helgenom-sekventering af patogener for fødevaresikkerhed, og menneskers og dyrs sundhed har fået trækkraft, og myndighederne i forskellige lande har taget teknologien til sig og gjort den til en del af deres rutineprocedurer. Genomisk epidemiologi, brugen af genomiske data i udbrudsundersøgelser og sygdomsovervågning, har øget antallet af løste udbrud og reduceret klyngestørrelserne og forbedret folkesundheden på nationalt niveau. Desuden kunne overfloden af WGS-data i offentlige data-lagre også bruges. Imidlertid har der været behov for nye softwareløsninger for at give et overblik over disse data, og ud over det, for at sammenligne og analysere prøverne af de samme subtyper for at opdage potentielle udbrudsklynger. Min ph.d. består af projekter med det formål at udvikle skalerbare og kontinuerlige, dvs. "evergreen" beregningsmetoder til genomisk epidemiologi.

Denne afhandling introducerer først subtype-bestemmelse inden for mikrobiologi, DNA-sekventeringsteknologier og fylogeni, der skal tjene som baggrund for projekterne inden for genomisk epidemiologi.

De første to inkluderede manuskripter centreret omkring en pipeline til automatiseret fylogenomisk analyse af bakterielle isolater (PAPABAC). Da helgenom-sekventeringsdata for bakterieisolater fra kliniske prøver og miljøprøver indsamlet over hele verden frigives i tusindvis, er sammenligning af disse isolater med det formål at detektere udbrud blevet mulig ved at anvende en skalerbar og autonom metode. PAPABAC blev brugt i **Paper I** til at nøjagtigt sammenligne udbrudsisolater af *Escherichia coli*, *Campylobacter jejuni* og *Listeria monocytogenes* benchmark-datasæt, selv med tilstedeværelsen af tusind ikke-udbrudsrelaterede prøver. Derfor blev det brugt i Evergreen Online, en platform til overvågning af fødevarebårne udbrud. Dens ydeevne ved klyngedannelse blev sammenlignet med NCBIs Pathogen Detection pipeline og viste sig at være ens. I **Paper II** analyserede vi WGS-data fra 5.665 methicillin-resistente *Staphylococcus aureus* (MRSA) og 2.572 *Enterococcus faecium* patientisolater opsam-

let i Danmark over en 5-årig periode. PAPABAC gennemgik opgraderinger for at håndtere ukendte og manglende baser mere fleksibelt. De SNP-baserede træer var i overensstemmelse med dem, der blev opnået med cgMLST. For MRSA var epidemiologiske data også tilgængelige, og PAPABAC-klyngerne indeholdt kendte hospitalsudbrud og andre epidemiologisk forbundne isolater.

Konceptet bag PAPABAC kunne også anvendes til andre sekvensdata, ikke kun WGS-data fra bakterier. **Paper III** indeholder Krummholz, en modificeret version af PAPABAC, der er egnet til subtype og klyngekonsensussekvenser opnået fra vira. For at demonstrere gennemførligheden af metoden blev Krummholz-træer genereret for norovirus VP1-sekvenser downloadet fra Genbank. Ved 85% lighedstærskel adskilte pipelinen VP1-sekvenserne langs deres genogrupper, hvilket skaber et overblik over den genetiske mangfoldighed af de deponerede sekvenser i Genbank. I Microreact viser Krummholz-træerne for genogrupperne GII.2, GII.3, GII.4 og GII.17 den tidsmæssige og rumlige spredning af prøver indsamlet under gastroenteritisudbrud i mennesker. Således blev Krummholz påvist at hjælpe og forenkle overvågningen af virale patogener, hvilket er meget vigtigt for vaccine- og diagnostikudvikling.

Det afsluttende projekt, der er beskrevet i **Paper IV**, analyserede helgenomsekventerede patientisolater af svamp indsamlet til en point-prævalensundersøgelse fra Danske Kliniske Mikrobiologiske Laboratorier i 2018. De 51 prøver blev type-bestemt, og derefter blev tredive *Candida albicans* og otte *Candida glabrata* prøver udsat for fylogenetisk analyse. Der blev ikke påvist nogen tydelig sammenhæng mellem fylogenetisk placering og prøveudtagningsinstitut eller kilde i de danske prøver. Prøverne blev også testet for modtagelighed for svampedræbende midler og i sidste ende var alle *C. albicans* prøver fænotypisk modtagelige for alle testede svampedræbende midler, og kun to *C. glabrata* prøver var fænotypisk resistente over for azoler. Den fylogenetiske analyse blev udført med en pipeline skrevet til en populær workflowmanager, hvilket sikrer reproducerbarhed og lettere analyse af yderligere prøver. Således inkluderede vi også prøver indsamlet over hele verden, hvilket resulterede i fylogenier for begge patogener, hvor de danske prøver blev placeret i klader af globalt udbredte subtyper.

Afhandlingen afsluttes i et kapitel, der sammenfatter resultaterne af projekterne og diskuterer aspekter af arbejdet, der kan forbedres i fremtiden for at fortsætte udviklingen.

Summary

The emergence and adaptation of new technologies in DNA sequencing in the last decade led to the broader application of sequencing. Whole-genome sequencing of pathogens for food-safety, human and animal health has gained traction, and authorities of various countries have adopted the technology and made it part of their routine procedures. Genomic epidemiology, the use of genomic data in outbreak investigations and disease surveillance, has increased the number of solved outbreaks and reduced the cluster sizes, improving public health at a national level. Furthermore, the deluge of WGS data published in public repositories could be utilized as well. However, new software solutions were needed to provide an overview of this data, and going beyond that, to compare and analyse the samples of the same subtypes to detect potential outbreak clusters. My PhD is comprised of projects with the aim of developing scalable and continuous, i.e. "evergreen", computational methods for genomic epidemiology.

This thesis introduces first the basics of subtyping in microbiology, DNA sequencing technologies and phylogenomics, to serve as background for the projects in genomic epidemiology.

The first two included manuscripts centered around a pipeline for automated phylogenomic analysis of bacterial isolates (PAPABAC). As whole-genome sequencing data of bacterial isolates from clinical and environmental samples collected worldwide are being released by the thousands, comparison of these isolates with the aim of outbreak detection has become possible, by applying a scalable and autonomous method. PAPABAC was shown in **Paper I** to accurately cluster outbreak isolates of *Escherichia coli*, *Campylobacter jejuni* and *Listeria monocytogenes* benchmarking datasets, even in the company of a thousand non-outbreak related samples. Therefore, it was utilized in Evergreen Online, a platform for surveillance of foodborne outbreaks. Its performance on clustering was compared to the NCBI Pathogen detection pipeline, and found to be similar. In **Paper II**, we analysed WGS data from 5,655 methicillin-resistant *Staphylococcus aureus* (MRSA), and 2,572 *Enterococcus faecium*

patient isolates collected in Denmark over a 5-year period. PAPABAC underwent upgrades to handle unknown and missing bases more flexible. The SNP-based trees were in concordance with those achieved with cgMLST. For MRSA, epidemiological data were also available, and the PAPABAC clusters contained known nosocomial outbreaks and other epidemiologically linked isolates.

The concept behind PAPABAC could be applied to other sequence data as well, not just WGS data derived from bacteria. **Paper III** features Krummholz, a modified version of PAPABAC, suited to subtype and cluster consensus sequences obtained from viruses. To demonstrate the feasibility of the method, Krummholz trees were generated for norovirus VP1 sequences downloaded from Genbank. At 85% similarity threshold, the pipeline separated the VP1 sequences along their genogroups, creating an overview of the genetic diversity of the deposited sequences in Genbank. Displayed in Microreact, the krummholz trees of the genogroups GII.2, GII.3, GII.4 and GII.17 exhibit the temporal and spatial spread of samples collected during human gastroenteritis outbreaks. Thus, Krummholz was shown to aid and simplify the surveillance of viral pathogens, which is of high importance for vaccine- and diagnostics development.

The final project, described in **Paper IV**, analysed whole-genome sequenced fungal patient isolates collected for a point prevalence survey from Danish Clinical Microbiology Laboratories in 2018. The 51 samples were typed, and thereafter thirty *Candida albicans* and eight *Candida glabrata* samples were subjected to phylogenetic analysis. No evident correlation was detected between phylogenetic placement and sampling institute or source of the Danish samples. The samples were also tested for antifungal susceptibility, and ultimately, all *C. albicans* samples were phenotypically susceptible to all tested antifungal agents, and only two *C. glabrata* samples were phenotypically resistant to azoles. The phylogenetic analysis was performed with a pipeline written for a popular workflow manager, ensuring repeatability and easing the analysis of additional samples. Thus, we also included samples collected worldwide, which resulted in phylogenies for both pathogens, where the Danish samples were placed in clades of globally prevalent subtypes.

The thesis concludes in a chapter recapitulating the results of the projects and discussing aspects of the work that could be improved upon in the future, to progress further.

Preface

The PhD studies took place at the Technical University of Denmark, in the departments Bioinformatics (2017.08.01 - 2018.02.28) and National Food Institute (2018.03.01 - 2020.10.31), with Professor Ole Lund as main supervisor, and Professor Frank Møller Aarestrup as co-supervisor. An external stay was conducted at Erasmus Medical Center, Rotterdam, The Netherlands, under the supervision of Professor Marion Koopmans and Miranda de Graaf.

The PhD studies have been funded as part of the COMPARE project, that has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 643476., and as part of the VEO project, that has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No. 874735.

Kongens Lyngby, 30$^{\text{th}}$ October 2020
Judit Szarvas

Acknowledgements

I would like to express my heartfelt thanks to everyone who contributed in any manner to the completion of my PhD.

First, I would like to thank my main supervisor, Ole Lund, for entrusting me with this incredible project and supporting me throughout it. I would like to also extend my thanks to my co-supervisor, Frank Aarestrup, who keeps me motivated to do great research.

I would also like to say thank you to Miranda de Graaf and Marion Koopmans, for hosting me for my external stay at Erasmus Medical Center, and introducing me to viral genomics. To Mette Damkjær Bartels, Henrik Westh, and Ana Rita Rebelo for the fruitful collaborations. To Johanne Ahrenfeldt, for infecting me with her love for phylogenetic trees and letting me work with her during my master thesis. To Anna Maria Tetzschner, Derya Aytan, Philip Clausen and Alfred Florensa for all the bioinformatics discussions, coffee and beers we had.

And I'm truly grateful to have been part of great communities both at the former DTU Bioinformatics and at the Research Group for Genomic Epidemiology.

Last, but not least, I'm glad for having the support of my family and friends throughout my PhD.

List of publications

Research articles included in the thesis

Paper I J. Szarvas, J. Ahrenfeldt, J. L. B. Cisneros, M. C. F. Thomsen, F. M. Aarestrup, and O. Lund, "Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform," Commun. Biol., vol. 3, no. 1, p. 137, 2020.

Paper II J. Szarvas, M. D. Bartels, H. Westh, and O. Lund, "Rapid open-source SNP-based clustering offers alternative to core-genome MLST for outbreak tracing in the hospital setting" (Manuscript in preparation)

Paper III J. Szarvas, O. Lund and M. de Graaf, "Visualization of norovirus diversity and evolution trough ever-growing trees" (Manuscript in preparation)

Paper IV J. Szarvas, A. R. Rebelo, One Day in Denmark Consortium, and F. M. Aarestrup, "Danish whole-genome sequenced *Candida albicans* and *Candida glabrata* samples fit into global clusters" (Manuscript in preparation)

Research articles not included in the thesis

- J. Ahrenfeldt et al., "Metaphylogenetic analysis of global sewage reveals that bacterial strains associated with human disease show less degree of geographic clustering," Sci. Rep., vol. 10, no. 1, pp. 1–11, 2020.

Problem statement

Whole-genome sequencing of microbial organisms has become widely used outside research institutes, and transformed the daily routine at public health and food-safety institutes. This development has meant a massive increase in the amount of sequence data generated, calling for new methods to, amongst other things, predict the relatedness of samples with the intent of detecting outbreaks.

In this PhD thesis, we set out to develop computational methods for large-scale and update-able phylogenies, that enable continuous processing of sequencing data.

Abbreviations

bp	Base-pairs
cgMLST	Core-genome multi-locus sequence typing
ddNTP	Dideoxynucleotide-triphosphate
dNTP	Nucleoside-triphosphate
DNA	Deoxyribonucleic acid
FDA	Food and Drug Administration
GISAID	Global Initiative on Sharing All Influenza Data
GTR	General time reversible
HCV	Hepatitis C Virus
HIV	Iuman Immunodeficiency Virus
HKY	Hasegawa-Kishino-Yano substitution model
IRIDA	Integrated Rapid Infectious Disease Analysis
kbp	Kilo base-pairs
MALDI-TOF	Matrix-assisted laser desorption/ionization– time of flight
MLST	Multi-locus sequence typing
MSA	Multiple sequence alignment
MST	Minimum spanning tree
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing

NNI	Nearest neighbour interchange
PCR	Polymerase chain reaction
PFGE	Pulsec-field gel electrophoresis
SMS	Single-molecule sequencing
SRA	Short Read Archive
ST	Sequence type
UPGMA	unweighted pair group method with arithmetic mean
wgMLST	Whole-genome multi-locus sequence typing
WGS	Whole-genome sequencing

Contents

Resumé i

Summary iii

Preface v

Acknowledgements vii

List of publications ix
Research articles included in the thesis ix
Research articles not included in the thesis ix

Problem statement xi

Abbreviations xiii

Contents xv

1 Introduction 1
1.1 DNA sequencing in microbiology 1
1.1.1 First-generation sequencing 1
1.1.2 Next-generation sequencing 2
1.1.3 Single-molecule sequencing 3
1.2 Typing for microbial diagnostics and epidemiology 4
1.2.1 Molecular subtyping of bacteria 4
1.2.2 Multi-locus sequence typing and its variations 5
1.3 Reference based variant discovery 8
1.4 Phylogenomics 10
1.4.1 Distance based clustering methods 10

1.4.2 Maximum likelihood methods 12
1.4.3 Quantifying uncertainty in phylogenies 14
1.5 Genomic epidemiology 15
1.5.1 Bacterial pathogens 15
1.5.2 Viral pathogens 17
1.5.3 Fungal pathogens 18

2 Manuscripts **19**
2.1 Paper I 20
2.2 Paper II 30
2.3 Paper III 44
2.4 Paper IV 63

3 Conclusion and future perspectives **79**

Bibliography **83**

CHAPTER 1

Introduction

1.1 DNA sequencing in microbiology

1.1.1 First-generation sequencing

DNA sequencing methods were developed starting from the 1960s, after protein and RNA sequencing had been accomplished. The first method to sequence double-stranded DNA fragments up to a 100 bp in reasonable time was by Maxam and Gilbert in 1977 [1]. It relied on partial cleavage of the DNA strand at specific bases, and thereafter sorting the fragments by size via gel-electrophoresis. Soon afterwards, Sanger *et al.* [2] published their dideoxynucleotide chain-termination method. Similarly to the Maxam-Gilbert method, it produced mixtures of DNA fragments ending on a different base for each type of nucleotide, which were then size-separated by gel-electrophoresis to determine the sequence. However, it achieved this by synthesizing a complement strand of the single-stranded DNA of interest, randomly incorporating 2',3'-dideoxynucleotide analogues that terminate the chain elongation. [3]

The two methods, with incremental improvements, were both widely used in sequencing projects. The whole-genome sequence of the G4 bacteriophage [4], cauliflower mosaic virus [5], variola smallpox [6], *Haemophilius influenzae* [7], *Escherichia coli* K-12 [8], *Escherichia coli* O157:H7 [9], *Salmonella* Typhi [10], *Salmonella* Typhimurium [11], and many more organisms were determined in the next two decades [12].

The modern version of Sanger's chain-termination method, now simply called **Sanger sequencing**, uses a mixture of fluorescent dye labelled ddNTP molecules [13], and dNTPs in one reaction to synthesize the complement strands to DNA fragments up to 900 bp in length [14]. Size separation and detection of labels happen by capillary electrophoresis [15], that greatly increased the sequencing throughput. Sanger sequencing has been long considered the golden standard in targeted amplicon sequencing, for example for genotyping house-keeping genes of bacteria and fungi, or viral resistance markers, however NGS and SMS methods are now gaining traction

in these areas as well [16, 17].

The template DNA for the sequencing was initially amplified *in vivo*, in plasmid vectors. In 1981, a *in vivo* **shotgun sequencing** method was published that simplified single-strand library preparation by using M13 bacteriophage as expression system [18]. In the 1980s, the PCR technique was developed [19, 20], which gave rise to new *in vitro* methods for preparing sequencing libraries [21]. For WGS, the target genome was fragmented randomly into smaller, overlapping fragments either by mechanical or enzymatic means, in a process that is referred to as shotgun approach, which necessitated computational methods for piecing the sequences together [22]. *Haemophilius influenzae* was the first bacterium to be whole-genome sequenced, in order to demonstrate the possibility of it [7].

1.1.2 Next-generation sequencing

The Human Genome Project [6], started in 1990, hastened the innovation of high-throughput sequencing technologies. Several different **next-generation sequencing** solutions were commercialized, starting with 454's pyrosequencing platform in 2005, and the competition reduced the cost of sequencing, democratising the technology for use outside research [21]. NGS has several applications in bacterial genomics research, for example functional genomics, comparative genomics, and phylogenetics [12]. In addition, WGS has also become a beneficial tool [23] in diagnostic microbiology for species determination [24,25], subtyping [26], antimicrobial resistance gene prediction [27, 28], virulence gene prediction [29], and phylogenomics [30].

Illumina machines (Illumina Inc., USA), utilising a **sequencing-by-synthesis** method developed by Solexa, are now the most widespread in use. The templates for the sequencing are generated *in vitro*, during a process called bridge amplification: after the sequencing library is loaded onto a flow cell, the DNA fragments are hybridising with the primers (adaptors) on the surface of the flow cell, and the PCR amplification takes place in situ, producing clusters of identical single stranded fragments attached to the surface. Sequencing happens with reversible terminators, that are labelled with fluorescent dye. In each cycle, the appropriate nucleotide is added to each growing complement strand, the non-incorporated labelled-dNTPs are washed out of the flow cell, the optical sensor captures the emitted light and the terminator and dye are cleaved off, allowing the synthesis to continue in the next cycle. The method allows for sequencing the same templates from both directions, and the index sequences ligated to the fragments make it possible to link them in order to obtain

paired-end reads. As the sequencing progresses sequencing errors become more likely due to polymerase errors, therefore the number of sequenced nucleotides is limited, resulting in 150 - 250 bp long sequencing reads. [31]

The probability of base-calling error is calculated for each base and expressed in a **PHRED score**,

$$q = -10 \times log_{10}(p) \tag{1.1}$$

where p is the estimated probability of the base being erroneously called [32]. Illumina reads typically have PHRED scores that are higher than 20, i.e. error probabilities lower than 1%.

Another sequencing-by-synthesis platform is **Ion Torrent** (Life Technologies, USA). It relies on emulsion PCR for template amplification, and the detection of an incorporated base happens via change in the pH. The dNTPs are added separately, but the reaction is not terminated after a dNTP is added, homopolymers thus pose a challenge to the technology. The read length is longer than for Illumina, 200 - 400 bp. However, the platform can only produce single-end reads. [33]

1.1.3 Single-molecule sequencing

The **third-generation sequencing** methods were developed with the aim of bypassing template amplification, which could introduce errors and mask nucleotide modifications. Yet, the accuracy of these methods are, despite continuous advances, still lower than for NGS. However, the longer read lengths (>1 kbp) make it possible to resolve repeating elements, detect long structural variants, and phase alleles in non-haploid organisms. For example, hybrid assemblies made from long and short reads are ideal for sequencing novel bacterial genomes. [12, 34]

Pacific Biosciences platforms use single-molecule real-time sequencing technology, which is a sequencing-by-synthesis method, that uses a special optical setup to detect individual labelled-dNTPs as they are incorporated into the complement strand. The molecules in the sequencing library are circular, and one fragment can be sequenced multiple times, resulting in circular consensus sequences with higher accuracy. [34]

Nanopore sequencing (Oxford Nanopore Technologies Inc., United Kingdom) abandons synthesis altogether, in favour of direct measurement of electric current generated by the molecules of single-stranded DNA (or RNA) passing through a protein nanopore. [35] The electric signal is translated into the nucleotide bases during base-calling. However, that is a complex task, thus nanopore base-calling has a lower

accuracy, than short read technologies, with error rates of 5 - 15% [36, 37]. Higher sequencing depth increases the consensus accuracy, but polishing with short reads could still be necessary. [35]

1.2 Typing for microbial diagnostics and epidemiology

The identification of the infectious agent, i. e. typing, is an crucial step in microbial diagnostics. Microbiological samples suspected to be harbouring pathogenic bacteria or fungi are first traditionally cultured on media to isolate the pathogen. Thereafter, the species is determined by Gram staining (for bacteria), biochemical tests and morphology. Recently, MALDI-TOF mass spectrometry based typing is also available in microbiology laboratories. [38] In contrast, culturing viral pathogens is expensive and time-consuming. Instead, immunological assays for the detection of the viral antigens and host antibodies, and various methods for the amplification of viral RNA or DNA are in use to type the viral agent directly from a clinical sample [39].

1.2.1 Molecular subtyping of bacteria

For surveillance and infection control, determining just the species of a bacterial isolate is insufficient and subtyping is necessary, preferably with high resolution in order to better differentiate between strains of the same species. Established *in vitro* methods include immunological reactions for serotyping, and nucleic-acid based tests for fingerprinting.

In vitro **serotyping** involves slide agglutination procedures with specific antibodies (antisera) against selected bacterial surface antigens, such as polysaccharides on the cell wall of Gram-negative bacteria (O-antigen), flagellar components (H-antigen) or capsular polysaccharides (K-antigen). For example, *Escherichia coli* serovars are defined by O and H antigenic combinations, though K-antigen could also be detected in specific serotypes [40], just like in *Salmonella enterica* that has pathogenic serovars with Vi capsular antigens [41]. Both species display considerable antigenic variation, and performing the full serotyping analysis is time-consuming and resource intensive. However, the serotype of a pathogen is easily shared between public health and other governmental agencies and often informative with regards to the disease outcome. Therefore alternative PCR, microarray, and later WGS based *in silico* methods [26, 42] have been developed to detect the genes encoding the H-antigens and the proteins of the biosynthesis pathways for the antigenic polysaccharides. [43]

Although serotyping can delineate outbreaks to a certain degree, it does not have discriminatory power to gauge the relatedness of isolates of the same subtype like molecular fingerprinting could. Restriction endonuclease analysis with **pulsed-field gel electrophoresis** has been considered the gold standard for that in the past decade. The method is generalizable by changing the restriction endonuclease(s) to suit the given species. Laboratories in the PulseNet network [44], that carry out surveillance of foodborne pathogens, use standardized PFGE procedure for increased inter-laboratory repeatability. The chromosomal DNA extracted from bacterial isolates is enzymatically cleaved at specific motifs with restriction endonuclease(s). Thereafter the large DNA fragments are size-separated and the pattern created by PFGE is characteristic of the isolate. Outbreak strains are clonally related and expected to produce identical patterns. However, if the endonuclease recognition sites undergo point mutation that prevents the enzyme from recognizing the cleavage site or if mobile genetic elements are present, closely related strains could have markedly different patterns. Conversely, populations separated by micro-evolutionary events may produce identical patterns. [45, 46]

1.2.2 Multi-locus sequence typing and its variations

While WGS was infeasible in the past for routine application in microbial diagnostics, genotyping shorter DNA fragments has been possible and available for decades. Moreover, sequencing based methods have greater reproducibility, than other molecular methods, permitting inter-laboratory comparisons. **Conventional (or legacy) MLST** [47] of bacterial and fungal pathogens is performed by determining the sequence of 6-10 housekeeping genes or gene fragments. The sequence variation of genes under neutral selection in a population characterise strains below the species level. For each species, curators maintain an MLST scheme with numbers assigned to the known allelic variants of each loci. The isolate alleles are compared to this scheme, and the sequence type of an isolate is defined by the combination of alleles at each loci. Ideally, a central profile database exists containing the STs with their unique allelic combinations. This setup eases outbreak information sharing and communication between public health authorities. [48, 49] Currently, MLST schemes and profiles are available through EnteroBase (`https://enterobase.warwick.ac.uk`), Institut Pasteur MLST (`https://bigsdb.pasteur.fr`) and PubMLST (`https://pubmlst.org/`) websites.

Closely related samples from the same lineage of a clonal population would have the same subtype with conventional MLST. As this is the case in outbreaks, samples

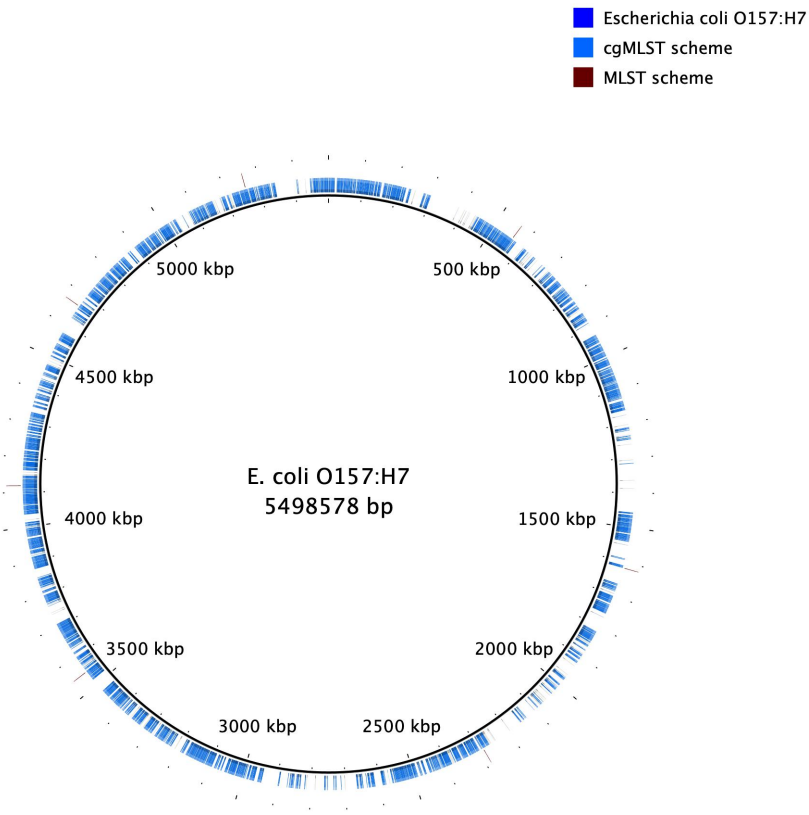

Figure 1.1. Loci in *Escherichia coli* cgMLST and conventional MLST schemes, displayed on the *Escherichia coli* O157:H7 complete reference genome; illustrating the coverage and resolution these typing schemes allow. cgMSLT is superior to the conventional MLST, yet the large number of accessory genes in *E. coli* serovars limit the discriminatory power, in comparison to whole-genome SNP analysis.

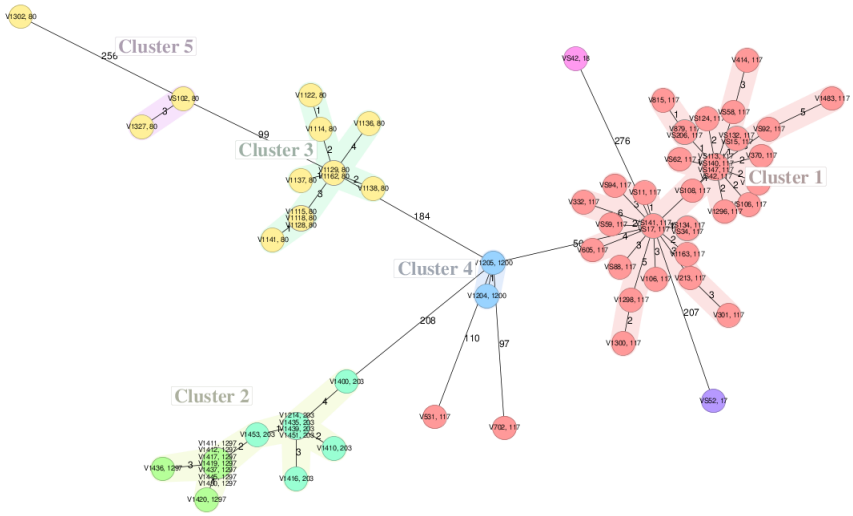

Figure 1.2. Minimum spanning tree representation of a bacterial population based on cgMLST, generated by Ridom SeqSphere+. The clustering threshold is set to 20 allelic differences. Identically coloured nodes placed together are part of the same cluster. Samples sharing the same node have identical allelic variants.

from the same outbreak would most often share ST, unless it is caused by a mixed population. However, conventional MLST studies have shown, that pathogenic populations are frequently made up of a few populous STs. [47, 50] Furthermore, there are species and subtypes that exhibit limited genetic diversity [51]. Therefore, samples with identical legacy STs could be incidental, and not necessarily epidemiologically linked. One prosaic solution to the lack of discriminatory power is to increase the number of genotyped loci. **Core-genome MLST** is such an extension of MLST, where the examined loci are part of the species core genome. Whole-genome MLST takes the concept one step further by including also the accessory genome in the analysis to enhance the resolution. These subtyping methods provide higher resolution than PFGE, and they were expected to delineate outbreaks more precisely.

The wider application of WGS made it possible to asses whole microbial genomes with this gene-by-gene approach, and software solutions for defining the schemes and analysing samples, for example BIGSdb [52], EnteroBase [53], chewBBACA

[54], Pathogenwatch [55], BioNumerics (Applied Maths NV, Belgium) or Ridom SeqSphere+ (Ridom Gmbh, Germany), accelerated the user engagement. cgMLST schemes have been defined and are being maintained for many of the important human pathogens, like *Enterococcus faecium* [56], *Listeria monocytogenes* [57], *Salmonella enterica* [53,58], *Staphylococcus aureus* [59], etc. It has been demonstrated for several species, that cgMLST or wgMLST clusters are highly concordant with SNP-based clusters [60, 61]. However, for *Klebsiella pneumoniae*, cgMLST was less accurate tool to infer the phylogeny of a clonal group of often carbapenese-resistant hospital-associated K. pneumoniae strains [62]. [63]

MLST is not just a practical and generalizable method for characterisation of isolates, but could also yield information about the phylogeny of samples, by applying clustering or graph methods to distances calculated on the allelic profiles. Dendrograms and minimum spanning trees are two of the resulting outputs. MSTs (Figure 1.2) represent connected edges with the lowest sum in the graph of the samples, where the edges are the allelic differences between a given pair of samples. Missing allelic information is normal for cgMLST, loci could be deleted from a given strain, or be absent due to low genome assembly quality. Applications typically have a threshold for the maximum allowed missing loci, but novel clustering methods could also handle higher rates of missing loci [64].

Majority of the pipelines utilizing the gene-by-gene approach rely on assembling the raw reads into draft genomes, and thereafter local alignment is performed between the contigs and the alleles in the applicable MLST scheme via BLAST [65]. The allelic variants are selected using a software specific ranking method [66–68]. Alternatively, raw reads could be mapped to the MLST scheme [69, 70]. Different sequencing platforms and bioinformatic pipelines could produce differing allele calls for the same samples.

1.3 Reference based variant discovery

NGS technologies typically produce sequence fragments that are shorter than the target gene or genome, therefore the reads need to be pieced together. The reference-free method to perform this step is called *de novo* assembly. However, for phylogenetic analysis with the aim of outbreak detection (see Figure 1.3 for an example), using a reference is preferable. Mapping identifies the possible origin(s) of the sequence fragment amongst the reference sequences, and alignment determines the exact placement the fragment in relation to the reference. Usually, the programs follow the seed-chain-

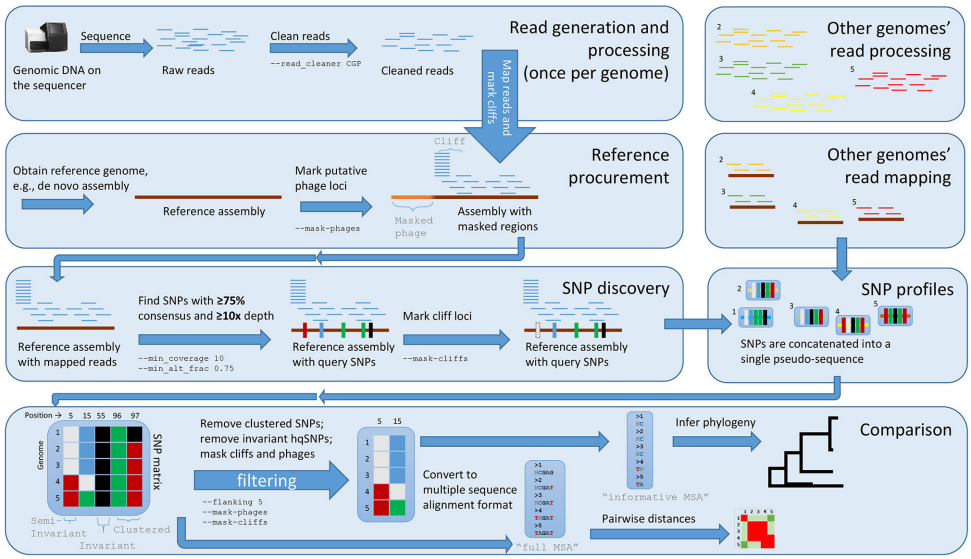

Figure 1.3. The Lyve-SET hqSNP-pipeline is a typical SNP-based outbreak investigation pipeline. Reproduced from [71]

align paradigm. Seeding is the identification of matches between short slice of the query and the reference, via indexing or hashing. Thereafter, the seeds are linked together during chaining, and then the rest of the query is aligned as well. Alignment is the basis for determining the similarity between two sequences, and read alignment algorithms could perform local or global alignment between the read and the reference. Gaps are allowed both in the reference and the query to account for insertions and deletions, respectively. Popular short read alignment programs are BWA [72,73], bowtie2 [74] and smalt[1]. In addition, the more recently developed minimap2 [75] and KMA [76] could also handle third-generation long reads.

Typically, the output of read alignment software is a sequence alignment file, which needs to be further processed to identify the differences (single nucleotide polymorphisms, insertions, deletions) from the reference or to gain a consensus sequence. Therefore, the majority of the **SNP-pipelines** first select all single-nucleotide variants and thereafter filter out the low quality SNPs. GATK [77] and VarScan [78] are utilities for gathering and filtering SNPs. The pipelines, e. g. CSI-phylogeny [79], CFSAN SNP pipeline [80], Lyve-SET pipeline [71], SNVPhyl [81], SnapperDB [82] differ in their filtering strategies and stringency. Generally, a SNP-matrix containing

[1] `https://www.sanger.ac.uk/tool/smalt-0/`

only the variant sites of the analysed sequences is one of their outputs. The shorter sequence matrix allows faster phylogenetic reconstruction, but analysis of additional samples require repetition of the matrix-construction step. In contrast, KMA and the read-aligner program in NDTree [83] can create consensus sequences by evaluating the aligned bases in each position of the reference, and taking the one supported by the majority of the reads, if criteria for the number of supporting reads and significance are fulfilled. These consensus sequences are the same length as the reference sequence, thus an additional sample does not necessitate the re-creation of the previous consensus sequences to form a multiple alignment.

1.4 Phylogenomics

A phylogeny traditionally describes the evolutionary relationship between taxa. Phylogenomics refers to inferring phylogenies from genome-scale data, a method to reveal more recent evolution between samples of interest [84], excluding homologous recombination or horizontal gene-transfer, which require alternative approaches. [85, 86].

Figure 1.4 shows a phylogenetic tree with eight taxa, labelled A-H. The root of the tree is placed at its midpoint, because it does not contain an outgroup that would define the root of the tree and direction of the evolutionary process. The horizontal edges are the branches of the tree, signifying the evolutional distances (changes) between the nodes, most commonly given in SNPs or substitutions per site. Taxa that have common ancestors belong to the same clade, for example F, G and H.

Phylogenies could be inferred from different data types (binary data, amino acid sequences, DNA sequences, etc.) and using various methods (parsimony, distance matrix based, maximum likelihood based and Bayesian methods) [87]. The phylogenies in the included projects were inferred by distance or maximum likelihood based methods, and the underlying data were DNA sequences.

1.4.1 Distance based clustering methods

The direct input for these methods are matrices containing the distance between the taxa. In case of DNA sequences, this distance (Hamming distance) is calculated on multiple alignments or SNP-matrices. Multiple alignments contain both variant and invariant sites, while SNP-matrices only retain the variant sites from the alignment, thus reducing the length of the comparable sequences. There are two ways it could be done: (1) taking each pair of samples, the observed genetic distance between them

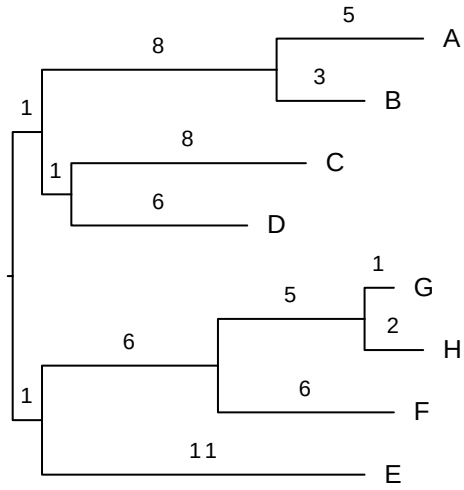

Figure 1.4. Distance matrix based phylogenetic tree with eight tips or taxa (A-H), rooted on the midpoint of the tree. The lengths of the branches are given in SNPs. Taxa A and B, and G and H both have most recent common ancestors. However, G and H are more closely related (3 differences), than A and B (8 differences).

is the number of positions with differing nucleotides, discounting positions with gaps and unknown bases; (2) after the removal of all positions with gaps and unknown bases from the multiple alignments or SNP-matrices, the observed genetic distance between a pair of samples is the number of positions with differing nucleotides. The observed distances should be corrected to account for any unobserved change to get the correct tree, but methods that assume molecular clocks are tolerant of uncorrected values. The resulting tree should have the lowest sum of branch lengths out of all possible trees that fit the input data. [87]

The **UPGMA algorithm** [88] is a clustering method (see Algorithm 1), that assumes a molecular clock, and therefore works best with closely related taxa or samples collected at the same time. Meanwhile, the **neighbour-joining algorithm** [89, 90], described in Algorithm 2, does not assume a molecular clock, and could be applied to data sampled at different time. Both algorithms are available in the PHYLIP software package[2].

[2] `http://evolution.genetics.washington.edu/phylip.html`

Algorithm 1: UPGMA algorithm [87, 88]

Data: m taxa/groups g, list of group sizes n , $m \times m$ distance matrix D
Result: UPGMA tree
Initialize tree, $n_m = 1$
repeat
  Find smallest D_{ij}
  Create $g_{(ij)}$ with size $n_{(ij)} = n_i + n_j$
  Connect nodes i and j to a new node (ij) on the tree with two branches of length $D_{ij}/2$
  Calculate distances from $g_{(ij)}$ to all others with
  $D_{(ij),k} = (\frac{n_i}{n_i + n_k})D_{ik} + \frac{n_j}{n_j + n_k})D_{jk}$
  Delete columns and rows for i and j from matrix
  Add (ij) to matrix
  $m \longleftarrow m - 1$
until $m == 1$

1.4.2 Maximum likelihood methods

Maximum likelihood is a parameter estimation method. In brief, the method provides model parameter estimates that maximize the probability (likelihood) of the observed data, given the model. For phylogenetic inference, the observed data is a discrete-character dataset, e. g. DNA multiple alignment. The evolution of the characters over the sites is described by probabilistic models of nucleotide substitutions (transitions and transversions). Several **evolutionary models** of different complexity have been described. One of the more simplistic model is the Hasegawa-Kishino-Yano model [91], that assumes that the base frequencies at equilibrium (π) are variable, and has one transition (α) and one transversion rate (β) (see Table 1.1). The general time-reversible model [92] is a more complex model, which has a symmetric substitution-rate matrix with six rates, and variable base frequencies. The appropriate model is chosen for a given dataset ideally by testing each model, for example comparing log-likelihood ratios or calculating information criteria. However, there is evidence, that a GTR model with heterogeneous rates of evolution across sites (+G) and a portion of invariant sites (+I) gives as good tree topology as a test-selected model [93]. Given these parameters, one can calculate the probability of transitioning from one given state on a branch to another one. Both HKY and GTR are reversible models, which means that the probability of transitioning from a state at one end of the branch to at one at the other end and vice versa is the same. Therefore phylogenetic trees

Algorithm 2: Neighbour-joining algorithm [87, 89, 90]

Data: n taxa/groups g, $n \times n$ distance matrix D
Result: NJ tree
Initialize tree with star topology
repeat
  foreach g_m **do**
    Calculate $u_i = \sum_{j:j\neq i}^{n} D_{ij}/(n-2)$
  end
  Find smallest $D_{ij} - u_i - u_j$
  Create $g_{(ij)}$
  Connect nodes i and j to a new ancestral node (ij) on the tree with branch lengths

$$v_i = \frac{1}{2} \times D_{ij} + \frac{1}{2} \times (u_i - u_j)$$

$$v_j = \frac{1}{2} \times D_{ij} + \frac{1}{2} \times (u_j - u_i).$$

  Calculate distances from $g_{(ij)}$ to all others with
  $D_{vk} = (D_{ik} + D_{jk} - D_{ij})/2$
  Delete columns and rows for i and j from matrix
  Add (ij) to matrix
  $n \longleftarrow n - 1$
until $n == 2$
Connect nodes i and j with branch length D_{ij}

computed with maximum likelihood methods are unrooted by default. [87, 94]

There are two important assumptions to calculate the likelihoods: the evolution is independent (1) at different sites, (2) on different branches of the phylogeny. This is necessary for calculating the joint probabilities of the transitions between the nodes across the phylogeny with the selected model of substitution. [87, 94]

RAxML, PhyML, FastTree and IQ-tree [95–98] are software solutions for inferring maximum likelihood trees. They take a discrete character dataset, a model of evolution and a seed tree. Generally, they alternate between two stages of the optimisation process. One to estimate the branch lengths and evolution model parameters for a fixed tree topology, and a second stage to find a tree topology that fits the observed data for the given model parameters. For large datasets, the amount of possible bifurcating tree topologies necessitate limiting the search space and the use of heuristic

Table 1.1. Substitution rate matrix for HKY model [91]. The columns contain the initial nucleotide and the row the final nucleotide.

	T	C	A	G
T	$-(\alpha\pi_C + \beta\pi_A + \beta\pi_G)$	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
C	$\alpha\pi_T$	$-(\alpha\pi_T + \beta\pi_A + \beta\pi_G)$	$\beta\pi_A$	$\beta\pi_G$
A	$\beta\pi_T$	$\beta\pi_C$	$-(\alpha\pi_G + \beta\pi_T + \beta\pi_C)$	$\alpha\pi_G$
G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	$-(\alpha\pi_A + \beta\pi_T + \beta\pi_C)$

algorithms, thus the final phylogenetic tree might not be the one with global maximum likelihood, but rather one that is at a local optimum of the search-space. [94]

1.4.3 Quantifying uncertainty in phylogenies

Phylogenetic inference comes with uncertainties that need to be taken into account when interpreting the inferred phylogenetic tree. **Nonparametric bootstrap** [99, 100] is often used to estimate the confidence of the splits in a phylogenetic tree. It resamples the original multiple sequence alignment with replacement n times, and the tree inference method is applied to each of these bootstrap replicates independently. The support for each split in the consensus tree is calculated from the n bootstrap trees. For maximum likelihood based tree inference, standard bootstrapping could be computationally expensive, therefore alternative methods for quantifying uncertainty have been developed. The **ultrafast bootstrap** (UFBoot2) [101, 102] in IQ-tree collects candidate bootstrap trees for the bootstrap MSAs by resampling estimated log-likelihoods for a given tree and the bootstrap MSAs during searching the tree-space of the original MSA. For each bootstrap MSA, the bootstrap tree is chosen randomly from a set of candidate trees for that MSA, that exceeded a minimum log-likelihood and are within ε_{boot} log-likelihood of the best candidate tree in the set. Bootstrap values could be overestimated for splits with decreased sampling variance, and tend to increase with the number of sites included in the MSA. [103] Another statistic, the **nonparametric approximate likelihood-ratio test** (SH-aLRT) [104] computes the support for a given branch by comparing the log-likelihood of the tree with the log-likelihood of the best alternative tree produced by a nearest neighbour interchange move around that branch. Simulations show, that SH-aLRT overestimate

the branch support in cases, when an alternate tree topology exists with better log-likelihood, than the ones produced by NNI. [101]

1.5 Genomic epidemiology

Epidemiological investigation compiles information from various sources to uncover cases (instances of illness), that are connected in time and space, and form an outbreak. Genomic epidemiology includes genomic information in support of traditional outbreak investigation methods.

1.5.1 Bacterial pathogens

Foodborne bacterial infections are a high-profile public health concern. PFGE has long been an essential tool in detecting outbreaks, and the PulseNet initiative paved the way for sharing and comparison of outbreak data beyond the institution where it was generated. Highly discriminatory molecular subtyping methods are capable of indicating possible sources of the infection as well. Adaptation of sequence based methods further increased the reproducibility of subtyping, and thereafter NGS technology brought the promise of highly discriminatory, sequence-based subtyping. In 2010, Harris *et al.* published a study, where NGS and reference based whole-genome wide variant discovery provided SNP-matrices, that made it possible to infer the phylogeny of closely related healthcare–associated, methicillin-resistant *S. aureus* samples [30]. WGS was quickly proved to be useful in genotyping other monomorphic organisms, supplementing classic epidemiological investigation and shedding light on transmission routes, that conventional methods could not disambiguate [105–109]. One early example of international collaboration and real-time outbreak investigation was the German Shiga-toxin producing *E. coli* O104:H4 outbreak in 2011, where WGS data, generated on different sequencing platforms, were publicly released and analysed as the outbreak unfurled, leading to swift characterisation of the novel, hyper-virulent strains [110, 111].

National initiatives were launched to utilize WGS for routine surveillance of infectious disease. **Foodborne pathogens**, such as *E. coli*, *S. enterica*, and *Campylobacter jejuni* are major concern for public health and food-safety authorities. WGS generates data that can be used to predict phenotypes - antimicrobial resistance, virulence - with shorter turn-around-time, replacing a range of *in vitro* tests for characterising an isolate. Moreover, methods for genomic epidemiology have the potential

for global standardisation, and the Global Microbial Identifier consortium was set up for this purpose, doing so by defining minimum metadata standards and data sharing policies. [38, 49, 112, 113]

The GenomeTrakr project was started by the U.S. Food and Drug Administration in 2012, initially with the aim of building a historical database of *Salmonella* isolates collected by the agency. WGS data from the project were uploaded to the Short Read Archive of National Center for Biotechnology Information and made publicly available. The next year, a joint inter-agency project between the U.S. Center for disease Control, U.S. Department of Agriculture's Food Safety and Inspection Service, U.S. FDA and other partners was set up to sequence all *Listeria monocytogenes* samples collected in the United States. The data have been processed both by a commercial wgMLST pipeline (BioNumerics) and open-source high-quality SNP-based pipelines (CFSAN, Lyve-SET). WGS has demonstrably augmented listeriosis outbreak investigations in the United States, producing more refined clusters and linking more cases to food sources. [46] Other foodborne bacteria, e.g. *E. coli*, *Shigella*, *C. jejuni*, were subsequently added to the project. Isolates, that meet the quality control criteria, are uploaded to NCBI, where they are analysed by the Pathogen Detection pipeline[3]. It assembles the raw reads, calculates k-mer based distances between the isolates, and creates single-linkage based clusters with at maximum 50 SNPs distance. However, the clusters created by the NCBI Pathogen Detection pipeline are not used directly in regulatory decisions, for example the FDA utilizes the CFSAN pipeline to infer phylogenies for isolates of interest. [114]

France ran a pilot project for evaluating the use of WGS in *L. monocytogenes* surveillance in 2015-2016. They found, that cgMLST had significantly better discriminatory power, than PFGE, and detected outbreaks earlier. Starting January 2017, France discontinued PFGE analysis and relied fully on WGS for monitoring listeriosis cases. [115]

Denmark made the same switch in September 2013, and the SNP based analysis helped to elucidate a nationwide listeria outbreak already in 2014 [116], then afterwards a prolonged multi-country outbreak 2015 - 2017, demonstrating the practicality of WGS for sharing data between national authorities [117].

In the United Kingdom, England and Wales switched to routine WGS typing of *Salmonella* isolates in April 2015 [118], and *E. coli* in October 2015 [119], using both SnapperDB and the wgMLST module of BioNumerics. Performance analysis of the new system revealed, that besides detecting small, geographically constrained

[3] `https://www.ncbi.nlm.nih.gov/pathogens/about/`

outbreaks, WGS is also capable of identifying long-term, national outbreaks, made up of cases previously taken for sporadic cases [120].

The Canadian Public Health Agency transitioned to prospective sequencing of all human *S. enterica* isolates in May 2017, with food and environmental samples following soon thereafter. The IRIDA platform[4] is the Canadian developed, open-source platform for outbreak investigation, containing the SNVPhyl pipeline amongst other tools. Clusters are defined via BioNumerics wgMLST, and the new technology led to increased number of investigated *Salmonella* clusters. [121]

On a smaller scale, WGS is convenient for surveillance of **nosocomial infections**. Routine sequencing of selected drug-resistant pathogens, for example methicillin-resistant *S. aureus* [30, 122] and vancomycin-resistant Enterococci [123–125], coupled with integrated data analysis of genomic and patient data proved to be useful for infection control in healthcare settings. [126, 127]

The gene-by-gene approach is popular in public health and food-safety laboratories, because it is easier to validate and standardise this method via the underlying schemes, and the commercial platforms have graphical user interfaces that are more intuitive to use, than command line applications. On the other hand, gene-by-gene methods are reliant on the cgMLST or wgMLST schemes existing for a given organism. Frequently, SNP-based pipelines are also applied to outbreak data, for the higher discriminatory power. At the initiation of this PhD project, no large-scale SNP-based solution existed, that was capable of comparing bacterial WGS data collected worldwide.

1.5.2 Viral pathogens

WGS of viral pathogens, such as influenza viruses, is less adapted into the clinical practice, than of bacterial pathogens. Viral genomes usually require target enrichment and amplification before library preparation, therefore amplicon sequencing of regions of interest is more prevalent. Targeted regions include genes implicated in drug resistance, encoding antigenic epitopes, or sequences used in subtyping the pathogen. [17] Molecular surveillance programmes sequence pathogens to monitor them for changes in these regions for development of diagnostic tools and vaccines [128, 129]. However, tiled amplicon sequencing protocols do exist for whole viral genomes [130–132], and WGS has been crucial for epidemiological investigations during the recent Ebola [133], Zika [134–136], and SARS-CoV-2 outbreaks [137, 138].

[4] T. Matthews et al., "The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform," bioRxiv, p. 381830, Jul. 2018. Preprint

Frequently, only the consensus sequences are published, that contain just the major variants. NCBI Genbank (and its mirrors) host a wide variety of viral sequence data, that is often utilized for meta-analysis [139, 140]. In addition, there are several initiatives for publishing curated or pathogen-specific data. For example, there are the Los Alamos HCV [141] database, the HIV database[5], and the Global Initiative on Sharing All Influenza Data Foundation's database [128]. These databases are important resources for surveillance, for example, the seasonal influenza genomic data from GISAID is used in phylodynamic analysis displayed on Nextflu [142]. This analysis pipeline, called Augur, has also been utilized for other viral pathogen datasets presented on Nextstrain [143].

1.5.3 Fungal pathogens

Pathogenic fungi could cause disseminated infections in immuno-compromised individuals. *Candida* species are frequently the culprits behind invasive candidiasis. *C. albicans*, *C. glabrata* and *C. parapsilosis* are the most prevalent infectious agents. Molecular epidemiology of these pathogens has been mainly investigated by MLST [144–153]. WGS based population genetics studies [154–158] and limited studies of serial clinical isolates [159, 160] have been made just recently.

[5] `https://www.niaid.nih.gov/research/hiv-database-analysis-unit`

CHAPTER 2

Manuscripts

COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-020-0869-5 OPEN

Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform

Judit Szarvas [1,2]✉, Johanne Ahrenfeldt [1,2], Jose Luis Bellod Cisneros[1], Martin Christen Frølund Thomsen[1], Frank M. Aarestrup [1] & Ole Lund [1]

Public health authorities whole-genome sequence thousands of isolates each month for microbial diagnostics and surveillance of pathogenic bacteria. The computational methods have not kept up with the deluge of data and the need for real-time results. We have therefore created a bioinformatics pipeline for rapid subtyping and continuous phylogenomic analysis of bacterial samples, suited for large-scale surveillance. The data is divided into sets by mapping to reference genomes, then consensus sequences are generated. Nucleotide based genetic distance is calculated between the sequences in each set, and isolates are clustered together at 10 single-nucleotide polymorphisms. Phylogenetic trees are inferred from the non-redundant sequences and the clustered isolates are added back. The method is accurate at grouping outbreak strains together, while discriminating them from non-outbreak strains. The pipeline is applied in Evergreen Online, which processes publicly available sequencing data from foodborne bacterial pathogens on a daily basis, updating phylogenetic trees as needed.

[1] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark. [2] These authors contributed equally: Judit Szarvas, Johanne Ahrenfeldt. ✉email: jusz@dtu.dk

Epidemiological typing of bacteria is used by hospitals and public health authorities, as well as animal health authorities, to detect outbreaks of infectious diseases and determine trends over time. Traditionally, this includes culturing and isolating the pathogen, followed by species identification and subtyping using various conventional microbiological and molecular methodologies.

For outbreak investigation, however, it is necessary to place the infectious agent into a more discriminatory category than species, to establish links between cases and sources. Multi-locus sequence typing (MLST) has been a frequently used molecular subtyping method, where sequence types are assigned to the isolates based on the combinations of alleles for 6–10 housekeeping genes[1].

Whole-genome sequencing (WGS) has opened a new chapter in microbial diagnostics and epidemiological typing. WGS data can be used to determine, amongst other characteristics, both MLST types and serotype of several bacterial species[2,3]. Furthermore, several studies for multiple bacterial species have shown the value of WGS for elucidating the bacterial evolution and phylogeny, and identifying outbreaks[4–6].

The use of WGS has enabled the unbiased comparison of samples processed in different laboratories, boosting surveillance and outbreak detection, but the methods for sharing and comparing a large number of samples have not been established yet[7,8]. Therefore, a number of national, regional, and international initiatives have been launched with the aim of facilitating the sharing, analyses and comparison of WGS data[9–11].

Since 2012, the US Food and Drug Administration (FDA) is leading a network of public health and university laboratories, called GenomeTrakr. These laboratories sequence bacterial isolates from food and environmental samples, and upload the data to the National Center for Biotechnology Information (NCBI). GenomeTrakr is restricted to foodborne pathogens and currently includes data from seven such bacterial species[12]. All raw WGS data are publicly shared through NCBI, facilitating the collaboration between laboratories. Furthermore, the raw data are picked up by the NCBI Pathogen Detection pipeline[13], that assembles the samples into draft genomes to predict the nearest neighbors and construct phylogenetic trees for each within-50-SNPs cluster using an exact maximum compatibility algorithm[14]. This approach requires access to all of the raw data or assembled genomes, and very extensive computational resources for larger databases, like *Salmonella enterica*. In addition, no sub-species taxonomical classification has so far been implemented in the pipeline.

Focusing on the same bacterial species as GenomeTrakr, PulseNet USA has also established procedures for use of WGS data for outbreak detection. In their vision, an extension of the highly successful MLST approach into a core-genome (cgMLST) or whole-genome (wgMLST) scheme, with in the order of a thousand genes, would allow for sharing information under a common nomenclature[11]. MLST schemes are offered from several databases[15–17], and a number of, at times conflicting, cg- and wgMLST schemes have recently been proposed for a limited number of bacterial species[16,18–24]. Moreover, few of the proposed schemes provide a definitive nomenclature of sequence types to go with the allele profiles. The existing schemes do not cover all of the potential allelic variation: a recent study showed, that for *Campylobacter jejuni*, that has maintained MLST schemes, only approximately 53% of the strains of animal origin could be assigned to an existing unique allelic profile[25]. Continuous curation of the hundreds of relevant bacterial species, that are known human, animal and plant pathogens, would require great effort. A centralized database for the distribution of the allele profiles and sequences would be also necessary. Furthermore, for comparable results, and surveillance, the same analysis pipeline or software should be used for the prediction of the allelic profiles. For example, single-linkage cgMLST clusters can be generated of public and private uploaded data on EnteroBase[15], and up to 1000 sequences on Pathogenwatch[26], by manual selection of strains to be included in the analysis.

The results generated by gene based approaches often lack the necessary resolution, and in most cases, selected WGS data are further analyzed using single nucleotide profiling. Here, genomic variants (single nucleotide polymorphisms (SNPs), insertions and deletions) are derived by aligning WGS reads to a reference genome. For each bacterial species, custom single nucleotide profiling (SNP validation, cluster threshold determination, etc.) is necessary in order to achieve results that are biologically relevant and informative. Here, the choice of the reference genome is crucial to maximize the number of SNPs that could be detected, as SNPs outside the regions covered by the reference are overlooked. The analyzed samples and the reference genome are chosen on a case-by-case basis, usually based on the subtyping results. Various offline SNP analysis pipelines are used by laboratories and research groups for inferring phylogenetic trees for isolates of interest[27–32]. For example, Public Health England developed and uses SnapperDB for outbreak detection without initial cluster analysis by cg- or wgMLST. SnapperDB consists of tools to create a database of SNPs compared to a given reference sequence, and assign each isolate a SNP address based on single linkage clustering[33]. As for web based solutions, in addition to cgMLST clustering, EnteroBase also offers SNP analysis of user selected strains based on the predicted genotypes[15]. Real-time tracking of bacterial pathogens with nucleotide resolution is performed by NCBI Pathogen Detection platform, that processes WGS data from selected bioprojects[13]. For viral pathogens, real-time tracking and visualization of evolution has been realized for seasonal influenza with Nextflu[34], and generalized in Nextstrain[35].

We present here a whole-genome, SNP-based method for subtyping and preliminary phylogenomic analysis of bacterial isolates, that circumvent the known limitations of current gene- and SNP-based approaches. PAPABAC carries out rapid and automated subtyping of bacterial whole-genome sequenced isolates and generates continuously updated phylogenetic trees based on nucleotide differences. We demonstrate two applications, a standalone version for local monitoring of bacterial isolates, and Evergreen Online, for global surveillance of foodborne bacterial pathogens. We also suggest a stable naming scheme for each isolate, making the results from the pipeline easier to communicate to others. To the best of our knowledge, no such tool exists at the moment.

Results

Automated phylogenomic analysis of bacterial WGS data. We developed PAPABAC (Fig. 1), a pipeline for automated phylogenomic analysis of bacterial isolates, that needs no additional input besides WGS data (FASTQ files) and generates clusters of closely related isolates. PAPABAC first matches the isolates to complete bacterial chromosomal genome reference sequences with greater than 99.0% sequence identity and a minimum average depth of 11. These reference sequences serve as templates for the alignment of the raw reads. The aligned bases at each position are statistically evaluated to determine the consensus sequence, as previously described for a nucleotide difference method[36]. Positions that do not fulfill the significance criteria remain ambiguous, get assigned "N", and are disregarded during the pairwise genetic distance calculation. These steps ensure that there is high confidence in the consensus sequence that is the basis of the genetic distance estimation.

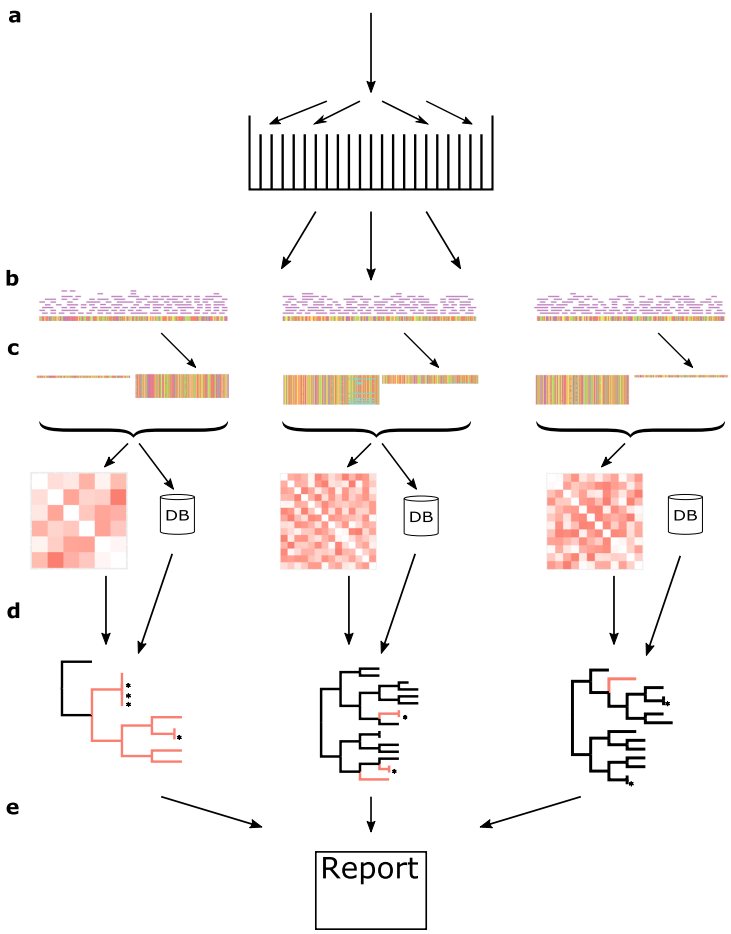

Fig. 1 Overview of PAPABAC. a The input raw read files are classified into sets based on k-mer similarity to NCBI RefSeq complete prokaryotic chromosomal genomes. **b** The raw reads are mapped to the reference genome and a consensus sequence is generated via strict statistical evaluation ($p < 0.05$) of the mapped bases in each position. **c** The resulting consensus sequences are of equal length in each template set. The new isolates in each set are clustered to the non-redundant isolates already in the set if the pairwise nucleotide difference based genetic distance is less than 10. The remaining new isolates undergo the same clustering process. **d** Pairwise genetic distance between all non-redundant isolate in the set is used as input for neighbor-joining algorithm. If there are less than 600 non-redundant isolates in a set, an approximately maximum likelihood phylogenetic tree is also inferred based on the consensus sequences (red: new isolates). The clustered isolates are placed back onto the trees with 0 distance to the cluster representative (marked with an asterisk). **e** The information about the acquired isolates, the sets, the clusters, and the phylogenetic trees is stored in SQLite databases, which are queried once all sets with new isolates are processed to output the results to the users.

The pipeline retains analysis results in such a manner that input is added to the previously processed data. The phylogenomic analysis is carried out on the current input and the previously found non-redundant isolates (singletons and cluster representatives). The genetic distance is estimated in a pairwise manner, comparing the given two sequences for all non-ambiguous positions, i.e., positions where none of the two sequences have an "N" assigned. The distances between the previously processed runs are stored on disk, saving computational time, and only the distances to the new isolates are computed in a given run.

A clustering step during the genetic distance calculation forms clusters of closely related isolates and reduces the number of similar sequences in each set, and thereby also reduces the computation time. After identifying a non-redundant isolate and a closely related isolate to it, the one previously deemed non-redundant will be the cluster representative and kept, while the clustered one is omitted from the subsequent runs of the pipeline. However, the information about the clustering is added to a database and the clustered isolate will be placed on the inferred phylogenetic tree. The cluster representatives remain constant through the subsequent runs of the pipeline, and the clusters only

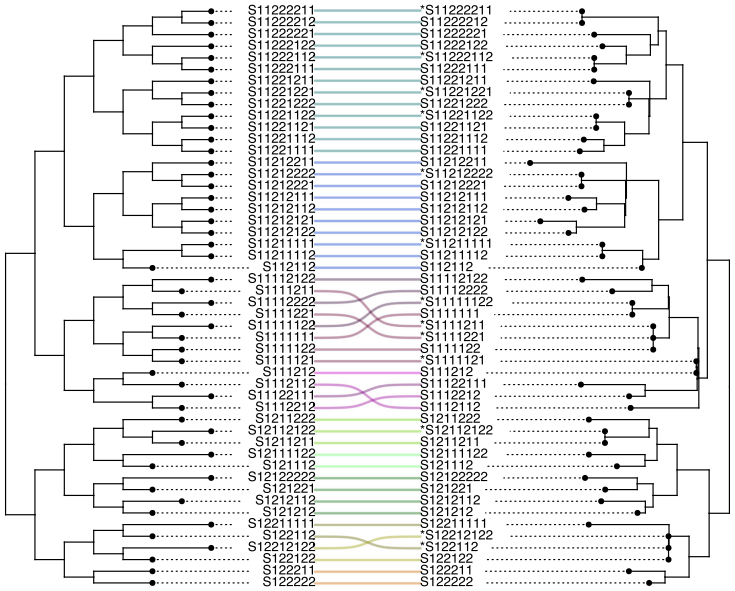

Fig. 2 Benchmarking of PAPABAC. Comparison of the ideal tree (left) to the PAPABAC maximum likelihood tree made of the in vitro experiment dataset[36]. Taxa with an asterisk were clustered together with the taxa in the same clade.

increase in size if new isolates are clustered with the representative. Therefore, each cluster is stable in the sense that an isolate will never change which representative it is associated with and each cluster can be reliably identified by the template name and the identifier of its cluster representative.

The pipeline can be run on a computer with 8 Gb RAM and Unix system. The computational time is reduced compared to re-running the whole analysis each time new samples are added, even without parallelization (Supplementary Fig. 1).

PAPABAC was benchmarked against three SNP pipeline benchmarking datasets. An *Escherichia coli* in vitro evolution experiment dataset[37] provided 50 closely related samples on a short temporal scale with less than 100 nucleotide differences across the dataset. The algorithm clustered together seven out of ten samples with the same ancestor that were taken on the same day and presumably had less than ten nucleotide differences between them. The PAPABAC maximum likelihood (Fig. 2) and neighbor-joining (Supplementary Fig. 2) trees with the clustered isolates pruned to resolve the polytomies were comparable to the ideal phylogeny of the in vitro experiment dataset: the normalized Robinson-Foulds distances were 0.18 and 0.12, respectfully. Benchmarking against the *Campylobacter jejuni* (Supplementary Fig. 3a, b) and the *Listeria monocytogenes* (Supplementary Fig. 3c, d) datasets from Timme et al.[38], PAPABAC correctly clustered the related outbreak strains (colored) and the outgroups, where the genetic distance was below the clustering threshold. The topologies of the maximum likelihood phylogenetic trees closely resembled the tree topologies given.

Online surveillance of foodborne bacterial pathogens. Evergreen Online was built on PAPABAC. Raw WGS data files of five major foodborne pathogens (*C. jejuni*, *E. coli*, *L. monocytogenes*, *Salmonella enterica*, and *Shigella* spp.) are downloaded daily from public repositories with the aim of global surveillance of potential outbreaks worldwide. The inferred phylogenetic trees and information about all of the isolates in the system are available and searchable on the website (http://cge.cbs.dtu.dk/services/Evergreen). The full phylogenetic trees can be viewed in the platform, or in Microreact[39], where the temporospatial information of the samples are also presented visually. Recognizing that phylogenies with more than a few hundred nodes are difficult to browse, the hyperlink of the cluster representative leads to subtrees, limiting the taxa shown to those surrounding the given cluster. Refined phylogenetic trees, based on SNP differences between all isolates around the cluster representative, are also published in Evergreen Online. The platform has been available since October 1st 2017, with logs reliably saved since October 28th 2017. The number of raw read files downloaded fluctuates with the work week of the public health laboratories. On busier days, more than 800 isolates are downloaded. The average number of isolates downloaded per day is 418. Downloading and mapping to the reference genomes take 130 min on average, with the majority of the time spent on downloading. Alignment of the raw reads and the generation of the consensus sequences takes on average 9 min per isolate. The computing time for the template sets is dependent on the number of non-redundant and new sequences in each set, but in most cases even the slowest is finalized within 5 h (Fig. 3 and Supplementary Data 1).

As of June 26th 2018, the pipeline downloaded 82,043 isolates. Out of these, 63,276 isolates have been mapped to references with at least 99.0% identity and average depth of 11 (Supplementary Fig. 4a). The majority of the isolates were typed as *Salmonella enterica* (59.1%), followed by *Escherichia coli* (19.4%) (Supplementary Fig. 4b). The two largest template sets are *S.* Dublin and *S.* Typhimurium serovars, with both close to 9500 isolates in total. After the homology reduction there were 3216 and 5093 non-redundant sequences in these sets, respectively. As of July 13th

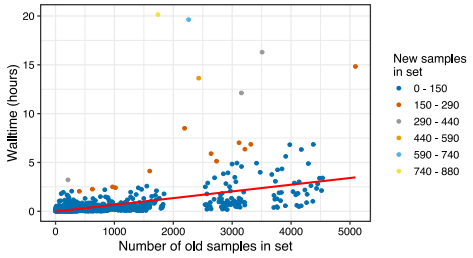

Fig. 3 Walltime of the phylogenetic analysis. Time requirement of the phylogenomic analysis for given number of non-redundant and new strains, on 20 CPUs.

2018, on average, 67% of the sequences are non-redundant in the template sets, while the E. coli template sets are the most diverse and the *Listeria monocytogenes* ones are the least diverse (Supplementary Fig. 4c). There were 122 isolates predicted to have a type not specified by the query (Supplementary Table 1). Of these, 14 isolates were mixed samples, composed of both the queried and the non-queried organisms.

The *L. monocytogenes* SNP pipeline benchmarking dataset[38] was added to the template set (Listeria_monocytogenes_07PF0776_NC_017728_1) of the corresponding reference genome in Evergreen Online, to test the sensitivity and accuracy of the clustering in large datasets. This template set at that moment contained more than 2400 isolates, of which 1398 were non-redundant. The isolates were placed onto a clade of a clonal lineage. The outbreak and outgroup isolates were separated in concordance with the ideal phylogeny (Fig. 4). The smaller clade of outbreak samples clustered to a sample (SRR538386) of an environmental swab in 2014, from California, USA.

Isolates that were presumed to be from an *E. coli* O157:H7 outbreak were selected for the comparison of Evergreen Online and the NCBI Pathogen Detection platform (NCBI-PD). They were located on the Escherichia_coli_O157_H7_str_Sakai_chromosome_NC_002695_1 neighbor-joining tree from Evergreen Online and the PDS000000952.271 SNP cluster tree from NCBI-PD. The labeled isolates appeared in three clusters on the neighbor-joining tree. There were 19.9 nucleotide differences between the yellow and the red cluster representatives and 12.6 nucleotide differences between the yellow and the blue cluster representative. On the NCBI-PD tree, the isolates marked with red circles were on the same clade, while the ones marked with blue and yellow were intermixing on clades that were, at most, 15 compatible characters apart (Fig. 5). On the refined subtree encompassing the labeled isolates, the yellow and blue labeled isolates intermix, similarly to the NCBI-PD tree (Supplementary Fig. 5).

Discussion

Whole-genome sequencing, performed alongside the traditional methods in routine microbiology, yields hundreds to thousands of WGS isolates yearly in hospital, public health and food safety laboratories. This amount of data is overwhelming for many, and there is a lack of methods to generate a quick overview and help prioritize resources. The timely analysis of the sequencing data would allow the detection of more bacterial outbreaks and aid the prevention of further spread. However, lack of human and computational resources for this demanding task often hampers the prompt analysis of the data. Automating the initial subtyping phase would facilitate the start of an outbreak investigation. PAPABAC offers rapid subtyping for a wide range of prokaryotic organisms: the supplied database covers all bacterial subtypes with complete genomes present in NCBI RefSeq. Further reference genomes could be added to increase the covered sequence space, but the active curation of the reference database is not required for routine use. The selection of the reference sequence for the phylogenomic analysis is fast and robust. It is independent of pre-assumptions about the isolates: misclassification during previous analysis does not introduce errors into the downstream analysis. Contamination from another species is discarded during the consensus sequence generation. The subtyping step via k-mer based mapping to a close reference also serves as a sequencing quality control measure, because low-quality sequencing runs will typically result in isolates with low identity to any reference and/or low depth. These isolates do not progress further to the phylogenomic analysis, as they would not yield reliable results.

The phylogenomic analysis performed on the template sets has higher discriminatory power than cg- or wgMLST. The underlying nucleotide difference method was validated in five different studies[6,36,37,40,41]. By using all positions in the consensus sequences for estimating the genetic distance, instead of considering only selected loci, we ensure a high level of sensitivity, as we also include mutations that occur between genes.

The clustering step during the genetic distance calculation was introduced in order to reduce the homology in the template sets and thus reduce the computational burden as the template sets increase in size. However, the clustering threshold of 10 nucleotide differences also constructs informative clusters of highly similar isolates. Benchmarking with the *E. coli* in vitro evolution experiment dataset (Fig. 2) showed that the algorithm was capable of correctly clustering isolates that were derived from the same ancestor, while distinguishing them from other closely related strains. The same sensitivity was demonstrated on empirical outbreak datasets (Supplementary Fig. 3), where the pipeline clustered the outbreak-related strains and separated them from the outgroup strains. Both the maximum likelihood inferred and the neighbor-joining trees placed the outbreak strains correctly in the phylogeny. These results show, that PAPABAC provides quick and reliable information about the close relatives of an outbreak strain to provide candidates to perform a more thorough analysis on.

The design of PAPABAC means that once an isolate passed the homology reduction step, it will be present in the subsequent runs of the pipeline. When an incoming isolate is highly similar to a non-redundant one, the more recent will be the one that is clustered, added to the database and removed from further runs. Hence, the cluster representatives and clusters are robust to the addition of new data to the analysis. Therefore, PAPABAC yields a stable and communicable name for the clusters, comprised of the template name and the cluster representative. This is an advantage over cg- and wgMLST, where allelic profiles don't necessarily have communicable names, and the clusters could merge.

Evergreen Online has been steadily processing WGS data of foodborne bacterial pathogen isolates collected worldwide in real time (Supplementary Fig. 4a). It has been able to keep pace with the flow of the generated data that mainly came from public health and food safety laboratories. Excluding the download time and the optional maximum likelihood based phylogenetic inference, the whole analysis is done in less than a day, even for template sets with thousands of isolates (Fig. 3). This turnover time facilitates quick response in a potential outbreak scenario.

The isolates are not distributed equally across the templates in the system (Supplementary Fig. 4b). Out of the five queried species, *S. enterica* isolates are disproportionally represented. Sequences in the *S.* Dublin and the *S.* Typhimurium LT2 template sets comprise in total approximately half of the *S. enterica*

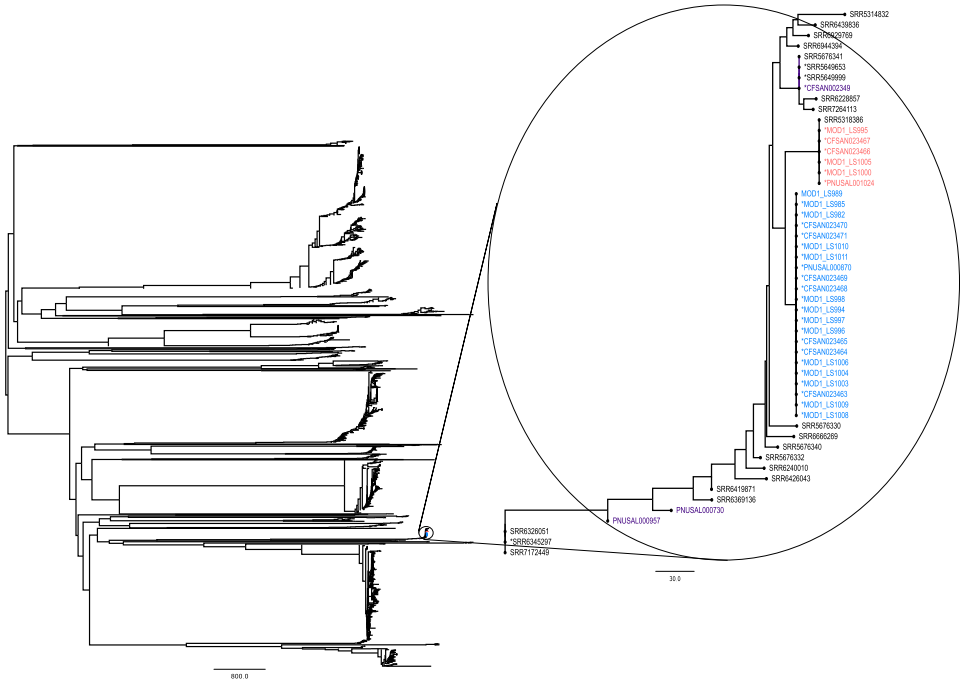

Fig. 4 Benchmarking of Evergreen Online on benchmark dataset for phylogenomic pipelines. Neighbor-joining tree for the Listeria_monocytogenes_07PF0776_NC_017728_1 set after the samples of the *L. monocytogenes* dataset were added. The red and blue isolates belong to the two original outbreak clades, and the purple isolates are outgroups in the benchmark dataset.

isolates. The sequence diversity in the template sets is varied, but the homology reduction on the template sets reduces the number of sequences approximately by a third, considerably decreasing the computational time. The *L. monocytogenes* template sets were the least diverse, which could be due to sampling bias: bacteria that are present in the environment are routinely sampled from food production sites multiple times, producing highly similar sequences, that are then removed from the ongoing analysis. We also tested how a large number of sequences already present in a template set would affect the ability of the pipeline to discriminate between samples (Fig. 4). The template set that corresponded to the stone fruit *L. monocytogenes* outbreak dataset reference had more than 1000 non-redundant isolates, which was ideal for the test analysis. The isolates that were part of the same outbreak clustered together and formed the two expected outbreak clusters, despite the confounding presence of the sequences already in the template set. The smaller clade, however, had a different cluster representative when using all data for the template set, compared with analysis of the outbreak data alone: an environmental sample, that could be related to the outbreak, as it was sampled from the same US state and year (California, 2014) as the samples in the outbreak dataset. These findings indicate that the pipeline is capable of identifying closely related samples, however it is necessary to conduct epidemiological analysis and apply other knowledge when interpreting the results.

Evergreen Online allows for automated selection of closely related isolates out of thousands, which is also the objective of NCBI-PD. *E. coli* isolates, situated on three clusters in Evergreen Online and supposedly from an outbreak, were located in NCBI-PD and their placement in the SNP cluster tree was compared to the Evergreen Online tree (Fig. 5). One cluster (red) was in agreement between the two platforms, and samples from the other two (yellow and blue) clusters were intermixing on a clade on the NCBI-PD tree. The nucleotide difference counts between these samples are low and the differences between the phylogenomic methods could lead to differences in the finer details of the inferred phylogenies. On the refined subtree (Supplementary Fig. 5), the labeled samples form similar topology to the NCBI-PD tree. The homology reducing clustering in Evergreen Online means that any sample in the cluster is less than ten nucleotides differences from the cluster representative, however, the differences between the samples could amount to 18 nucleotides. The compatible character distances on the NCBI-PD tree between the mixed samples are less than that. Taking this into account, the observed distribution of the labeled samples in the two platforms are concordant.

In summary, we developed PAPABAC with the aim of rapid subtyping and continuous phylogenomic analysis on a growing number of bacterial samples. PAPABAC overcomes limitations of cg- and wgMLST approaches by tolerating genomic variation during subtyping, but providing greater sensitivity during the phylogenomic analysis. It was benchmarked on datasets created for testing SNP-based pipelines, and was proved to be accurate in discriminating between outbreak related and non-related samples. The software is open source and fulfills expectations put to WGS-based surveillance pipelines (Table 1). Evergreen Online, an application made for the global surveillance of foodborne bacterial pathogens, demonstrates the accuracy, speed, stability and

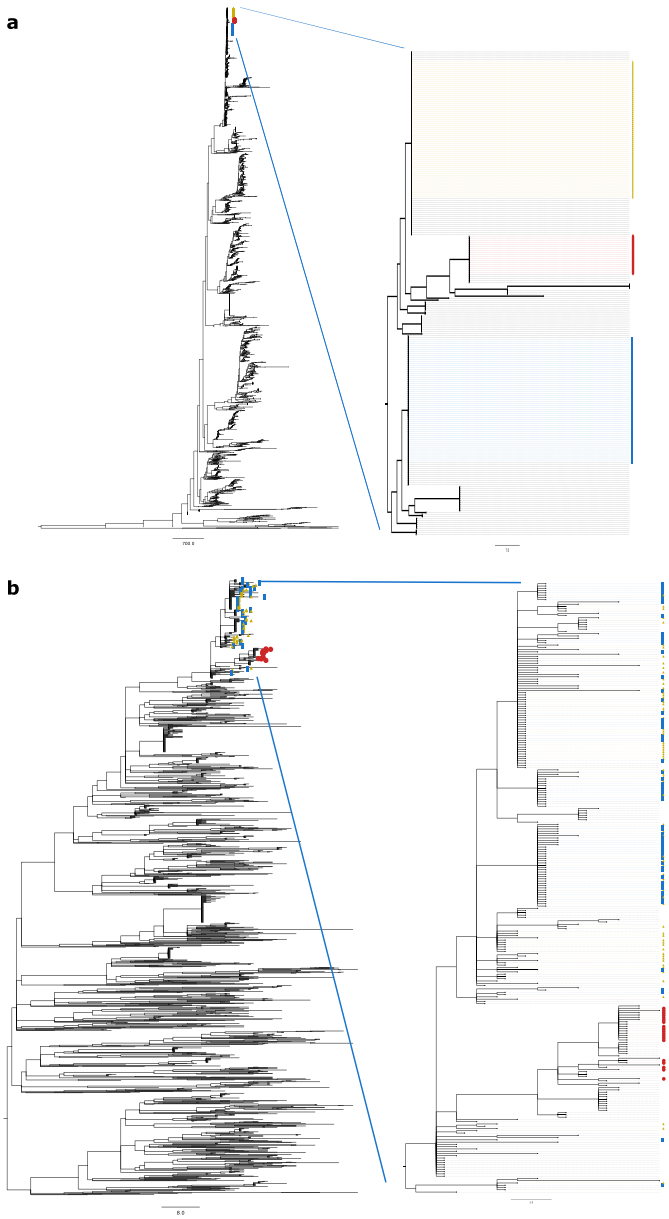

Fig. 5 Comparison of Evergreen Online and the NCBI Pathogen Detection platform. Selected isolates in the **a** Escherichia_coli_O157_H7_str_Sakai_chromosome_NC_002695_1 NJ tree and on the **b** PDS000000952.271 SNP cluster maximum compatibility tree. The three largest clusters of the selected samples on the NJ tree are labeled with yellow, red, and blue dots. These isolates were marked with the same labels on the NCBI-PD tree. The red labeled ones are on a single clade on the PD tree, while the blue and yellow isolates are mixing on two other clades.

Table 1 Comparison of pipelines for large-scale surveillance for pathogenic bacteria.

	SnapperDB	NCBI-PD	PAPABAC
For a wide range of bacterial species	x	x	x
Only sequencing data is required input	—	x	x
Whole-genome based	x	x	x
Assembly-free	x	—	x
Quality control steps	x	x	x
Automated phylogenomic analysis	—	x	x
Stable clustering of samples across runs	—	—	x
Communicable nomenclature for subtype and cluster	x	—	x
Open source	x	—	x

practicality of PAPABAC on thousands of samples via an ongoing analysis, where the results are published online.

Methods

PAPABAC pipeline. The pipeline takes raw whole-sequencing reads (FASTQ files) as input. Matching reference sequences (templates) in our reference database, that have greater than 99.0% identity and a minimum average depth of 11, are identified for the isolates using 16-mers via KMA[42] in sparse mode. Multiple templates are accepted, if they meet the criteria, allowing for the procession of mixed samples. Information about the runs and their templates are inserted into the main SQLite database. The isolates are grouped into sets according to the matched templates. The next steps are performed in these sets in parallel. The isolate reads are mapped to the template using the mapping algorithm of NDtree[36], yielding equal-length consensus sequences. The *Z*-score threshold for accepting a base is set to 1.96, and the majority base have to be present in 90% of the mapped reads.

Genetic distance based on nucleotide difference is calculated pairwise between the previous, non-redundant isolates, and the new isolates. Positions with ambiguous bases are discarded. The new isolates are clustered to the non-redundant ones with a threshold of 10, in order to reduce the homology in each set and form informative clusters. In this step, the non-redundant isolate is prioritized over the new isolate and becomes the cluster representative. After the clustering, the remaining new isolates are clustered together with the Hobohm 1 algorithm[43]. In this case, the cluster representative is the one that has already passed the redundancy threshold. The information about new or extended clusters is saved to the main SQLite database. A distance matrix is constructed for all non-redundant isolates and saved to disk for use in the next run. A distance-based phylogenetic tree is inferred by neighbor-joining[44,45]. If there are less than 600 non-redundant isolates in the set, then a whole-genome based approximate maximum likelihood phylogenetic tree is also inferred using IQ-tree[46], where the neighbor-joining tree is the starting tree and the GTR nucleotide substitution model is used. The clustered isolates are placed back onto the clades with zero distances to the cluster representative. Their tip labels start with an asterisk. The information about the trees is saved to the main SQLite database.

When all the phylogenetic trees with new isolates have been inferred, then the main SQLite database is queried for the list of all isolates, their templates, cluster representatives (if there is any) and the latest phylogenetic tree they are on. This information is printed to a tab-separated file.

Evergreen Online platform. A query is made to the National Center for Biotechnology Information (NCBI) Sequencing Read Archive (SRA) for the newly published Illumina paired-end sequenced isolates of *Campylobacter jejuni*, *Escherichia coli*, *Listeria monocytogenes*, *Salmonella enterica*, and *Shigella spp*. on a daily basis. Fastq files of raw sequencing reads and the corresponding metadata (collection date, location, institute, source, etc.) are acquired either from SRA or from the European Nucleotide Archive (ENA). The sample inclusion criteria is known metadata for collection date and location, and in addition, samples are included from the following institutions: Unites States Center for Disease Control, United States Food and Drug Administration, Food Safety and Inspection Service, Public Health England, University of Aberdeen, University Hospital Galway, Statens Serum Institut, Norwegian Institute of Public Health. The downloaded isolates are the input to PAPABAC. The metadata are saved in the main SQLite database, and added to the tip labels on the phylogenetic trees.

Individual subtrees are inferred from isolates with less than 20 SNPs distance from each cluster-representative, considering only the positions in the sequences where there is no missing data. No tree is inferred, if no genetic difference is found. The subtrees are inserted into an SQLite database.

Once all instances of the second wrapper script have finished, then the SQLite databases are queried for the list of available phylogenetic trees (the maximum likelihood trees preferred over neighbor-joining ones), changes in the clusters and the list of all isolates in the system, which is then used to update the website.

The phylogenetic trees are interactively visualized on the website (https://cge.cbs.dtu.dk/services/Evergreen/) using the Phylocanvas API (http://phylocanvas.org).

For visualization in external programs, such as Microreact[39], the phylogenetic trees can be downloaded as newick files and the corresponding metadata as tab separated files. The isolates and clusters can be searched by SRA run ID, which allows the quick localization of the clusters that increased in size via their cluster representative.

Architecture. The pipeline is written in Python 2.7 and Bash in Unix environment. In addition to the standard Anaconda Python 2.7 packages, it also requires ETE Toolkit v3.0[47] and Joblib v0.11 (https://pythonhosted.org/joblib) packages to be installed. Neighbor program from the PHYLIP package v3.697 (http://evolution.genetics.washington.edu/phylip.html) and IQ-tree v1.6.4[46] are used for the phylogenetic tree inference. The SQL database management is performed with SQLite v3.20.1 (https://www.sqlite.org).

The two main parts of the pipeline have their own wrapper scripts. PAPABAC can be run on a personal computer with as few as four cores. Evergreen Online is running on a high-performance computing cluster, utilizing the Torque (Adaptive Computing Inc., USA) job scheduler. The first wrapper is run in one instance on 20 cores, meanwhile the second wrapper is run once on 20 cores for each template that has at least one new run, in a parallel fashion. When all of these instances are finished running, a Bash script is launched to collect the information from the SQL database, the website is updated and the job for the next day is scheduled.

Reference database. The reference sequences are complete prokaryotic chromosomal genomes from the NCBI RefSeq database. Homology reduction was performed at a 99.0% sequence identity threshold with the Hobohm 1 algorithm. The curated NCBI prokaryotic reference genomes were given priority in the process. The reference sequences could be downloaded via ftp (ftp://ftp.cbs.dtu.dk/public/CGE/databases/Evergreen/).

Computational time comparison. One hundred and one samples from the Escherichia coli in vitro evolution experiment dataset by Ahrenfeldt et al. were batched according to their sampling time. The parallelization in PAPABAC was disabled. The traditional method meant that the analysis was carried out on all the samples up to the given batch, starting anew each time, but using the same scripts as PAPABAC.

Benchmarking with dataset by Ahrenfeldt et al. The last samples in each lineage of the Escherichia coli in vitro evolution experiment dataset[37] were selected for the benchmarking. Therefore, the benchmarking dataset constituted 50 tips on the ideal phylogeny. These samples were batched according to their sampling time (6th, 7th, and 8th day). The batches were processed by PAPABAC chronologically. The pipeline was run with the default parameters. Both maximum likelihood and neighbor-joining trees were inferred.

The phylogenetic trees inferred on all 50 isolates were trimmed for the reference sequence and compared with the ideal phylogeny using the phytools R package (v0.6–60)[48]. The normalized Robinson-Foulds distance was calculated between the ideal and the maximum likelihood, and the ideal and the neighbor-joining trees, after the clustered isolates are removed from each pair of trees. The RF.dist function was utilized from the phangorn R package (v2.4.0)[49].

Benchmarking with datasets from Timme et al. The *Campylobacter jejuni* 0810PADBR-1 and the *Listeria monocytogenes* 1408MLGX6-3WGS dataset[38] was downloaded with the provided script into a distinct directory. The pipeline was run individually on the datasets with default parameters. If the isolates were mapped to more than one template, the phylogenetic trees of the template set with the highest number of isolates were evaluated. The maximum likelihood trees were visually compared to the ideal phylogenies and checked for the distribution of the isolates amongst the clades.

Using the default options on the pipeline, the *L. monocytogenes* SNP dataset was added to a copy of the Listeria_monocytogenes_07PF0776_NC_017728_1 template set of Evergreen Online on 2018-06-15.

Comparison with the NCBI pathogen detection platform. *Escherichia coli* isolates were queried from the SQL database of Evergreen Online for the period of 2018-03-15 and 2018-06-01, corresponding to a multistate outbreak of *E.coli* O157:H7 in the USA[50]. These samples were subtyped using traditional MLST[2], as it was assumed, that the sequence type with the most isolates would also include the outbreak samples. Sequence type 11, which commonly corresponds to the O157:H7 serotype, was selected for further analysis. The corresponding samples and their SNP clusters were found in the NCBI-PD platform. The phylogenetic tree for the SNP cluster with the most samples (PDS000000952.271) was downloaded. The common samples (Supplementary Data 2) were noted on both the NCBI-PD and the Evergreen Online phylogenetic tree (Escherichia_coli_O157_H7_str_Sakai_chromosome_NC_002695_1, downloaded on 2018-08-07). The refined subtree around SRR6766978 was downloaded on 2019-10-30, and pruned to contain only the isolates that were on the Evergreen Online tree. The common samples on the three biggest clusters on the Evergreen Online tree were labeled, and their placement on the NCBI-PD tree and the refined tree was visually inspected.

Statistics and reproducibility. The code for running PAPABAC is provided, and running it on the same input data would yield reproducible results for the consensus sequences, clusters and neighbor-joining trees, for they are produced with deterministic algorithms.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

No novel datasets were generated during the current study. All analyzed data are available in this published article, or on the website: https://cge.cbs.dtu.dk/services/Evergreen/.

Code availability

Scripts and installation instructions for the pipeline are publicly available on Bitbucket: https://bitbucket.org/genomicepidemiology/evergreen.

Received: 4 February 2019; Accepted: 3 March 2020;
Published online: 20 March 2020

References

1. Maiden, M. C. J. Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* **60**, 561–588 (2006).
2. Larsen, M. V. et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **50**, 1355–1361 (2012).
3. Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 2410–2426 (2015).
4. Köser, C. U. et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *N. Engl. J. Med.* **366**, 2267–2275 (2012).
5. Mellmann, A. et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751 (2011).
6. Joensen, K. G. et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *J. Clin. Microbiol.* **52**, 1501–1510 (2014).
7. WHO. *Whole Genome Sequencing for Foodborne Disease Surveillance: Landscape Paper* (World Health Organization, 2018).
8. Deng, X., den Bakker, H. C. & Hendriksen, R. S. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.* **7**, 1–22 (2016).
9. Whole Genome Sequencing (WGS) Program | FDA. https://www.fda.gov/food/science-research-food/whole-genome-sequencing-wgs-program. Accessed 12 June 2019.
10. COMPARE Europe. http://www.compare-europe.eu.
11. Nadon, C. et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill.* **22**, 30544 (2017).
12. Timme, R. E., Sanchez Leon, M. & Allard, M. W. Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback. in *Foodborne Bacterial Pathogens. Methods in Molecular Biology* **1918**, 201–212 (2019).
13. Pathogen Detection—NCBI. https://www.ncbi.nlm.nih.gov/pathogens/. Accessed 27 June 2018.
14. Cherry, J. L. A practical exact maximum compatibility algorithm for reconstruction of recent evolutionary history. *BMC Bioinform.* **18**, 127 (2017).
15. Alikhan, N.-F., Zhou, Z., Sergeant, M. J. & Achtman, M. A genomic overview of the population structure of Salmonella. *PLOS Genet.* **14**, e1007261 (2018).
16. Cody, A. J., Bray, J. E., Jolley, K. A., McCarthy, N. D. & Maiden, M. C. J. Core genome multilocus sequence typing scheme for stable, comparative analyses of Campylobacter jejuni and C. coli human disease isolates. *J. Clin. Microbiol.* **55**, 2086–2097 (2017).
17. Institut Pasteur MLST databases and software. https://bigsdb.pasteur.fr/. Accessed 28 May 2019.
18. Ghanem, M. & El-Gazzar, M. Development of Mycoplasma synoviae (MS) core genome multilocus sequence typing (cgMLST) scheme. *Vet. Microbiol.* **218**, 84–89 (2018).
19. Higgins, P. G., Prior, K., Harmsen, D. & Seifert, H. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of Acinetobacter baumannii. *PLoS ONE* **12**, e0179228 (2017).
20. Ghanem, M. et al. Core genome multilocus sequence typing: a standardized approach for molecular typing of *Mycoplasma gallisepticum*. *J. Clin. Microbiol.* **56**, e01145 (2017).
21. Bletz, S., Janezic, S., Harmsen, D., Rupnik, M. & Mellmann, A. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. *J. Clin. Microbiol.* **56**, e01987-17 (2018).
22. Zhou, H., Liu, W., Qin, T., Liu, C. & Ren, H. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Klebsiella pneumoniae. *Front. Microbiol.* **8**, 371 (2017).
23. Kohl, T. A. et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. *J. Clin. Microbiol.* **52**, 2479–2486 (2014).
24. Moran-Gilad, J. et al. Design and application of a core genome multilocus sequence typing scheme for investigation of Legionnaires' disease incidents. *Eurosurveillance* **20**, 21186 (2015).
25. Leekitcharoenphon, P. et al. Comparative genomics of quinolone-resistant and susceptible Campylobacter jejuni of poultry origin from major poultry producing European countries (GENCAMP). *EFSA Support. Publ.* **15**, 1398E (2018).
26. Pathogenwatch | A Global Platform for Genomic Surveillance. https://pathogen.watch/. Accessed 28 May 2019.
27. Kvistholm Jensen, A. et al. Whole-genome sequencing used to investigate a nationwide outbreak of listeriosis caused by ready-to-eat delicatessen meat, Denmark, 2014. *Clin. Infect. Dis.* **63**, 64–70 (2016).
28. Schjørring, S. et al. Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017. *Eurosurveillance* **22**, 17-00762 (2017).
29. Ford, L. et al. Incorporating whole-genome sequencing into public health surveillance: lessons from prospective sequencing of Salmonella Typhimurium in Australia. *Foodborne Pathog. Dis.* **15**, 161–167 (2018).
30. Holmes, A., Dallman, T. J., Shabaan, S., Hanson, M. & Allison, L. Validation of whole-genome sequencing for identification and characterization of Shiga toxin-producing Escherichia coli to produce standardized data to enable data sharing. *J. Clin. Microbiol.* **56**, e01388–17 (2018).
31. Woksepp, H., Ryberg, A., Berglind, L., Schön, T. & Söderman, J. Epidemiological characterization of a nosocomial outbreak of extended spectrum β-lactamase Escherichia coli ST-131 confirms the clinical value of core genome multilocus sequence typing. *APMIS* **125**, 1117–1124 (2017).
32. Davis, S. et al. CFSAN SNP pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput. Sci.* **1**, e20 (2015).
33. Dallman, T. et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* **81**, 3946–3952 (2018).
34. Neher, R. A. & Bedford, T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* **31**, 3546–3548 (2015).
35. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
36. Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O. & Aarestrup, F. M. Evaluation of whole genome sequencing for outbreak detection of Salmonella enterica. *PLoS ONE* **9**, e87991 (2014).
37. Ahrenfeldt, J. et al. Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods. *BMC Genomics* **18**, 19 (2017).
38. Timme, R. E. et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **5**, e3893 (2017).
39. Argimón, S. et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* **2** (2016).
40. Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M. & Lund, O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE* **9**, e104984 (2014).
41. Joensen, K. G. et al. Evaluating next-generation sequencing for direct clinical diagnostics in diarrhoeal disease. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 1325–1338 (2017).

42. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinform.* **19**, 307 (2018).
43. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. Selection of representative protein data sets. *Protein Sci.* **1**, 409–417 (1992).
44. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
45. Studier, J. & Keppler, K. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* **5**, 729–731 (1988).
46. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
47. Huerta-Cepas, J. et al. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
48. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
49. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
50. CDC. Multistate Outbreak of E. coli O157:H7 Infections Linked to Romaine Lettuce (Final Update) | Investigation Notice: Multistate Outbreak of E. coli O157:H7 Infections April 2018 | E. coli | CDC. https://www.cdc.gov/ecoli/2018/o157h7-04-18/index.html. Accessed 7 August 2018.

Acknowledgements

The authors would like to thank every public health institute, that releases their sequencing data to the public in a prompt manner. Global surveillance would not be possible without the sharing of data. This project was part of"Collaborative management platform for detection and analyses of (re-) emerging and foodborne outbreaks in Europe" (COMPARE), that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 643476.

Author contributions

O. L. developed the overall idea for the project. J.Sz. and J.A. contributed equally: both contributed to the design of the pipeline, wrote the code and wrote the manuscript. J.Sz. carried out the analysis. J.L.B.C. and M.C.F.T. created software that was used in the work. O.L. and F.M.A. advised the study and helped to write the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42003-020-0869-5.

Correspondence and requests for materials should be addressed to J.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Rapid open-source SNP-based clustering offers alternative to core-genome MLST for outbreak tracing in a hospital setting

Judit Szarvas[1], Mette Damkjær Bartels[2,3], Henrik Westh[3, 4] and Ole Lund[1]

1 Research Group of Genomic Epidemiology, National Institute of Food, Technical University of Denmark, Denmark
2 MRSA Knowledge Center, Copenhagen University Hospital Hvidovre, Denmark
3 Department of Clinical Microbiology, Hvidovre University Hospital, Denmark
4 Institute of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark

Abstract

Traditional genotyping methods for infection control of antimicrobial resistant bacteria in health-care settings have been supplemented by whole-genome sequencing, often relying on a gene-based approach, e..g. core-genome multilocus sequence typing (cgMLST), to cluster related samples. In this study, we compared clusters of methicillin resistant *Staphylococcus aureus* (MRSA) and *Enterococcus faecium* analyzed with the commercial cgMLST software Ridom SeqSphere+ and with an open-source single-nucleotide polymorphism (SNP) based phylogenetic analysis pipeline (PAPABAC). 5,655 MRSA and 2,572 *E. faecium* patient isolates, collected between 2013 and 2018, were processed. Clusters of 1,844 MRSA and 1,355 *E. faecium* isolates were compared to cgMLST results, and epidemiological data was included when available. The phylogenies inferred by the two different technologies were highly concordant and the MRSA SNP-tree re-captured known hospital-related outbreaks and epidemiologically linked samples.

Introduction

The spread of antimicrobial-resistant bacteria such as methicillin resistant *Staphylococcus aureus* (MRSA), vancomycin resistant *Enterococcus faecium* (VRE) and carbapenemase producing organisms (CPO) in hospitals and health care facilities is a growing problem worldwide. Resistant bacteria are estimated to cause the death of 33,000 people in Europe [1] and 35,000 people in the US [2] each year. In recent years, whole genome sequencing (WGS) has become increasingly prevalent in routine clinical microbiology and infection control and has proven to be an excellent tool for outbreak investigations both in health care settings [3], [4] and in the foodborne transmission of pathogens [5]. The combination of epidemiological data and WGS makes it possible to track the transmission between patients in order to reveal an outbreak and avoid further transmission. However, the huge amount of data generated by WGS calls for a fast and reliable tool for sequence analysis of the genome data. Generally, either a single nucleotide polymorphism (SNP) based [6]–[9] or an allele based approach such as core-genome multilocus sequence typing (cgMLST)[10], [11] or whole-genome multilocus sequence typing (wgMLST) is used for the analysis of the genome data. The former relies on a suitable closed reference genome and a method to call SNPs with a high confidence, while the latter requires an available scheme for the species in question, a constantly updated database of alleles for each locus, correctly assembled genomic data and predicted genes.
The purpose of this study was to compare the SNP based open-source PAPABAC (pipeline for automated phylogenomic analysis of bacterial isolates) with cgMLST based on the commercial software Ridom SeqSphere+. The study included 5,655 methicillin-resistant *Staphylococcus aureus* (MRSA) patient isolates and 2,572 *Enterococcus faecium* patient isolates from Denmark from 2013 to 2018 sequenced on Illumina MiSeq. Available epidemiological data was included for some MRSA isolates to determine the most likely clustering. The clustering of isolates by the two methods was compared to evaluate the performance of the SNP based analysis.

Methods

Whole-genome sequences

5,655 methicillin-resistant *Staphylococcus aureus* (MRSA) and 2,572 *Enterococcus faecium* patient isolates that were whole-genome sequenced (WGS) at the Department of Clinical Microbiology at Hvidovre Hospital between January 2013 and January 2018, were included for further analysis. Sequencing runs with unknown multilocus sequence types (ST), or with differing STs in Ridom SeqSphere+ (Ridom GmbH, Germany) and in PAPABAC were excluded from the analysis.

As described before[12], [13], sequencing libraries were made with Nextera XT DNA sample preparation kit (Illumina, United States (US)) after DNA concentrations were normalized with Qubit (Invitrogen, United Kingdom (UK)). Paired-end sequencing was done on MiSeq (Illumina, US) yielding 150 basepair reads.

Core-genome MLST and phylogenetic analysis

Raw reads were assembled using SPAdes [14]. Multi locus sequence typing (MLST) and core-genome MLST (cgMLST) was done with SeqSphere+ v2.3-6.0 (Ridom GmbH, Germany). Samples with sub-par assembly quality were not uploaded to SeqSphere+, and in case of re-sequencing, the best quality assembly was added to SeqSphere+. The cgMLST schemes for *S. aureus* and *E. faecium* include 1,861 and 1,423 alleles, respectively. For *S. aureus*, SeqSphere+ creates clusters with a threshold of 24 or fewer allelic differences and for *E. faecium* with a threshold of 20 or fewer allelic differences.

Phylogenetic analysis was performed with PAPABAC v2.0 [8]. In short, raw reads were matched and aligned to reference sequences of chromosomal genomes already in the PAPABAC database with k-mer identities greater than 85.15%, creating equal length consensus sequences for each sample. These were gathered for each genome, and SNP based distances were calculated between them. Sequencing runs that had more than 12% of unknown bases in their consensus sequences were excluded from the analysis. To filter low quality regions from the alignment, all columns were removed from the sequence alignment that had one or more unknown base up until 30% of total missing data over the alignment, and thereafter all columns were removed that had more than 5% unknown bases. Sequences were homology reduced with a threshold of 10 SNPs, creating technical clusters. The distance-based clustering method was unweighted pair group method with arithmetic mean. Sequencing runs where batched according to the date of sequencing, and processed chronologically.

MRSA samples of ST97 and single locus variants (SLV) of ST97 were subjected to additional analysis with PAPABAC, using a MRSA reference genome of ST97 from a well-described hospital outbreak in Copenhagen [15]. This genome was closed with a combination of Nanopore sequencing and Illumina MiSeq sequencing and yielded a genome of 2,781,777 basepairs (bps) and a plasmid of 20,403 bps. A maximum likelihood phylogeny was also inferred with IQ-TREE v1.6.12 [16] with GTR+I+G nucleotide substitution model.

Comparison of cgMLST and PAPABAC results

For MRSA isolates, the clusters generated by cgMLST were compared with the clustering by PAPABAC. Some of the MRSA isolates were known to be from well-described hospital outbreaks and for many of the other isolates it was known whether the isolates were from persons from the same households or if they were epidemiologically connected in other ways. We did the comparison on the template set for the MRSA

ST80 reference sequence *Staphylococcus aureus subsp. aureus 11819-97* (accession: NC_017351.1), that matched the highest number of MRSA samples out of the 19 available *S. aureus* templates. Clusters of diverse STs (ST1, ST97, ST852 and ST1835) were located in the dendrogram and the minimum spanning trees from SeqSphere+, and the clustering of samples were compared, together with the available epidemiological information.

For the *E. faecium* samples, there were 4 templates in the database, and we used the VRE ST203 reference sequence *Enterococcus faecium Aus0085* (accession: NC_021994.1) for the comparison. Epidemiological data was not included for VRE samples.

Due to the assembly or consensus sequence quality requirements for inclusion, samples might not be present in both systems.

Because cgMLST creates clusters for *S. aureus* at 24 or fewer allelic differences and for *E. faecium* at 20 or fewer allelic differences[17], we considered sequences with the same SNP thresholds to be related.

Results

MRSA outbreaks and other epidemiologically connected samples

Between January 2013 and January 2018, 5,655 sequencing runs were generated from DNA extracted from MRSA isolates. Out of these, 1,930 genomes were matched to the reference sequence *Staphylococcus aureus subsp. aureus 11819-97*, and after discarding low-coverage sequences, 1848 genomes were included in the last phylogenetic tree. 1,804 genomes were *in silico* predicted to contain the *mec*A gene. The genomes in the phylogenetic tree belonged to 39 sequence types (STs). The most abundant sequence type was ST6, followed by ST1 and ST8.

630 genomes did not cluster to another at the 10 SNPs technical threshold, while 1,218 genomes clustered in 352 technical clusters. Technical clusters have a median size of 2 and a mean size of 3.46 runs. Looking at the distribution of STs in these 352 clusters, there were seven (2,0%) that also contained at least one genome that was a single-locus variant (SLV) of the ST of the cluster.

Some examples of MRSA SNP clusters are shown in Figure 1. One ST1 cluster (Figure 1A) contained seven genomes in the SNP based trees, with a maximum SNP distance of 18, while the maximum allelic distance was 13 in SeqSphere+ (Figure S1A). The technical clustering connected sequencing runs of the same sample, temporally displaced samples from the same patient, and samples from epidemiologically related patients. Another cluster with eight samples of a maximum allelic difference of seven in SeqSphere+ (Figure S1B) presented in our tree with a maximum distance of 17 SNPs (Figure 1B), encompassing one technical cluster of five isolates from two patients from the same household, another one with two isolates from the

same patient from this household, and a singleton sample from the other patient. The last cluster of ST1 we inspected had 51 samples in the SNP tree, divided over two clades and three additional singletons, with a maximum distance of 18 SNPs. In SeqSphere+, the maximum allelic difference was 15. Samples were from members from four connected families and two unrelated individuals (Figure 1C).

Eleven ST1835 samples formed a cluster of a previously described outbreak [18] with both clustering technologies (Figure S2). For ST852, six samples from a household clustered together in SeqSphere+ with a maximum allelic difference of 11, while our tree placed them within 20 SNPs, but into two clades and one sample as a singleton (Figure 1D).

A known hospital outbreak of ST97 MRSA [15] spanning five years had only 22 isolates in the SNP tree, as the remaining nine isolates related to the outbreak did not have the required identity to the reference sequence *Staphylococcus aureus subsp. aureus 11819-97*, and they were assigned to a different reference thereby splitting the outbreak isolates. The 22 outbreak isolates in the SNP tree were situated across four technical clusters with a distance of 42 SNPs. We repeated the analysis of all ST97 samples from the database (N=79) with a closed reference genome from this particular outbreak. All 31 outbreak samples (Figure 2) were now assigned to the outbreak reference genome, but were divided into four technical clusters and four singletons, within 40 SNPs (Figure 3)

E. faecium clusters

2,572 sequencing runs were produced between March 2013 and January 2018, that were typed to be *Enterococcus faecium*. The reference sequence *Enterococcus faecium Aus0085* collected 1,755 samples, of which 1,355 genomes had the necessary coverage to be included in the phylogenetic tree. These genomes were distributed over 17 sequence types, with ST203, ST117 and ST80 being the most abundant STs. 1,247 genomes were clustered into 89 technical clusters with the 10 SNPs distance threshold. The median size of these were three genomes and the mean size was 13.27 runs. 83 clusters contained just one ST, six clusters (6.7%) contained one or more genomes being SLVs of the dominant ST of the cluster.

We compared SNP-based clusters that were larger than 2 sequences to SeqSphere+ cgMLST clusters. All samples that were clustered with PAPABAC also clustered together via cgMLST (Figure S3).

Discussion

Surveillance and infection control in hospitals, that embrace WGS, require a system for processing and comparing thousands of sequences. Ideally, SNP-based solutions would be preferable, however that would require significant computational power or human resources. cgMLST-based solutions have become popular as these reduce the computational complexity by only comparing a few thousand gene alleles

instead of millions of base-pairs per sample and allow for all-against-all comparisons with a common nomenclature. We previously demonstrated a method for large-scale phylogenetic analysis to be accurate from few dozen to several hundred samples of food-borne Gram-negative bacteria. The stand-alone pipeline, PAPABAC, is available as open-source software. It can run automatically on new input raw sequencing data, quickly generating stable clusters of closely related isolates, making it ideal for use in continuous genomic surveillance. The clusters can be referred to by the sample identifier of the cluster representative, i. e. the first isolate in thecluster, and the template name. Moreover, it can accommodate the growing number of samples without significant increase in CPU time. [8]

Incorporating enhanced filtering for uncertain bases, we applied PAPABAC to Gram-positive bacteria, that are relevant in nosocomial infections, namely *S. aureus* and *E. faecium.* In this study, the resulting clusters, produced by the 10 SNPs technical threshold, encompassed sequences with the same STs, or in seven cases for MRSA and six cases for *E. faecium* only differed in one SNP in one locus, indicating a good specificity for related sequences.

Final phylogenetic trees, after the last batch had been included for analysis, were compared to cgMLST clusters generated with SeqSphere+, and for the MRSA samples, the available epidemiological data was included to determine the most likely clustering. Although we could not compare all clusters, due to their numbers, we expect similar level of concordance in the rest of the clusters. We found, that there was general agreement between the two methods as to what sequences could have common origins or be part of an outbreak, if we designate 24 allelic or SNP differences or fewer for MRSA, and 20 allelic or SNP differences or fewer for *E. faecium* to indicate this. We identified known MRSA outbreaks (Figure 1C, Figure 3, Figure S2), and patients who were close contacts (Figure 1A-B, D). However, epidemiological data could not support the links between all samples in some clusters, like in Figure 1A. Yet it remained true, that epidemiologically connected samples were closely related genetically.

Choosing the correct reference genome is important when using a SNP-based analysis. This is highlighted by the MRSA ST97 outbreak presented here, where some sequences were more similar to a different reference sequence in the default database, splitting the outbreak between two phylogenetic trees. One solution to avoid this, is to decrease the identity threshold for matching sequencing runs to reference sequences, however, low identity sequences would decrease the alignment length, leading to apparent higher similarities between sequences. A better solution is to use in-house reference genomes, that match closely with the pathogens present in a given population and location. We found, that aligning to the in-house ST97 reference genome, all of our outbreak-related samples could be placed onto the same phylogenetic tree, with 40 SNPs between them at most. This is larger, than what is usually considered to be closely related, but can be explained by the fact that the outbreak isolates were spanning over a five years

period. Using one of the outbreak isolates as reference genome for the SNP-based method gave a marginally higher resolution between the genomes in the outbreak, than using a more distant reference genome. CgMLST schemes are only available for a number of bacteria, and profiles available in SeqSphere+ are maintained and updated through the SeqSphere+ software. New alleles in the query sequence(s) are not included in the distance calculation unless submitted to cgMLST.org. Local cgMLST schemes could be defined, however, it's a laborious task compared to selecting a suitable reference genome.

In short, the two methods yield concordant phylogenies, and PAPABAC has the advantage over Ridom SeqSphere+ to generate stable, referable clusters without the need for sequence assembly, and it is a free of charge, open-source alternative to the commercial software.

Figures

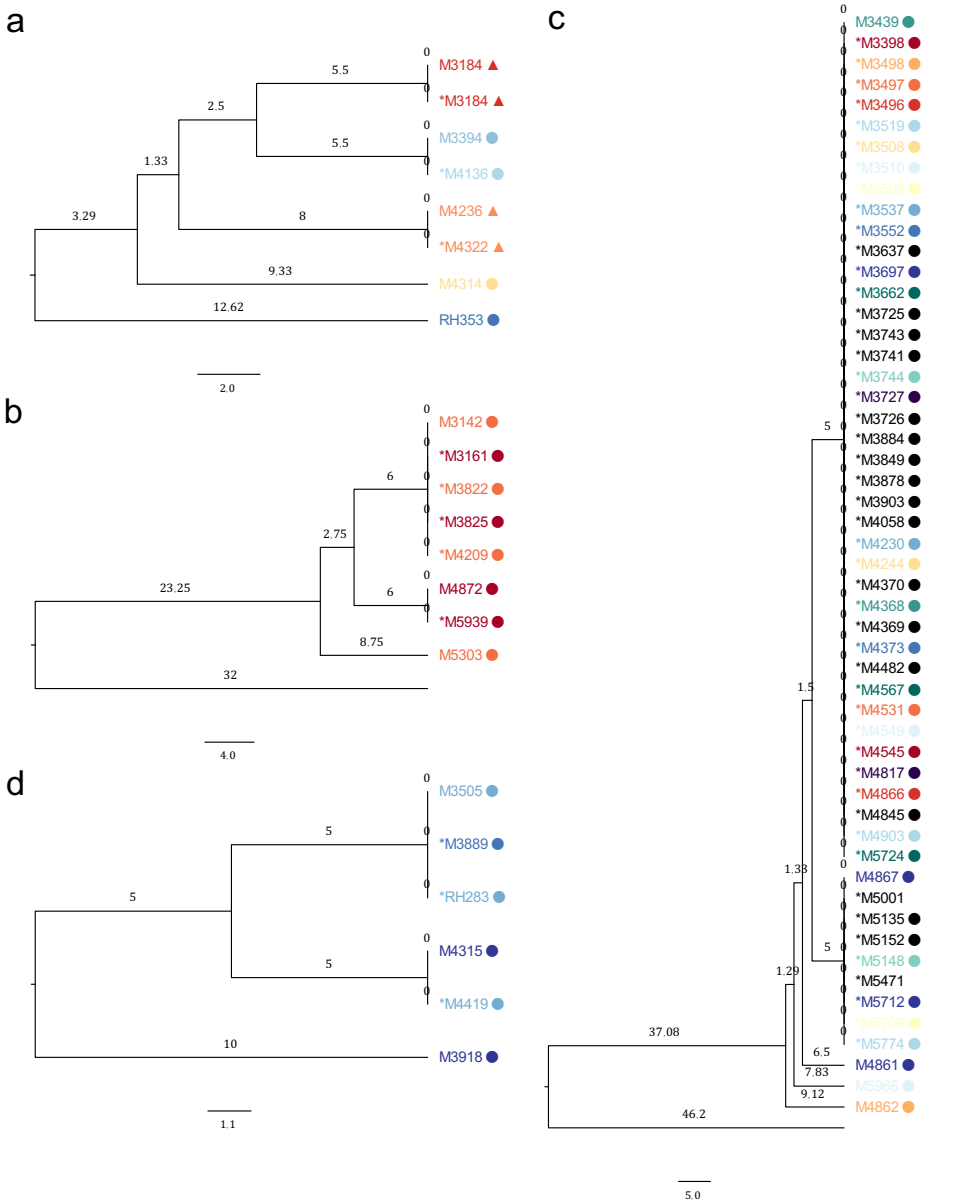

Figure 1: Subsections of the UPGMA tree made on the reference sequence Staphylococcus aureus subsp. aureus 11819-97, showing SeqSphere+ clusters, with samples from the same patient colored the same, epidemiologically related samples display the same symbol. The asterisk denotes samples in the technical clusters. Branch lengths are single nucleotide differences. (A) ST1 clusters with seven samples, sequences clustered were from the same biological sample, or from patients sharing a household, or from the

same patient. (B) ST1 cluster with eight sequences, from the same household. (C) ST1 cluster with 51 sequences, from members of connected families and unrelated individuals. (D) ST852 cluster with six sequences, from the same household.

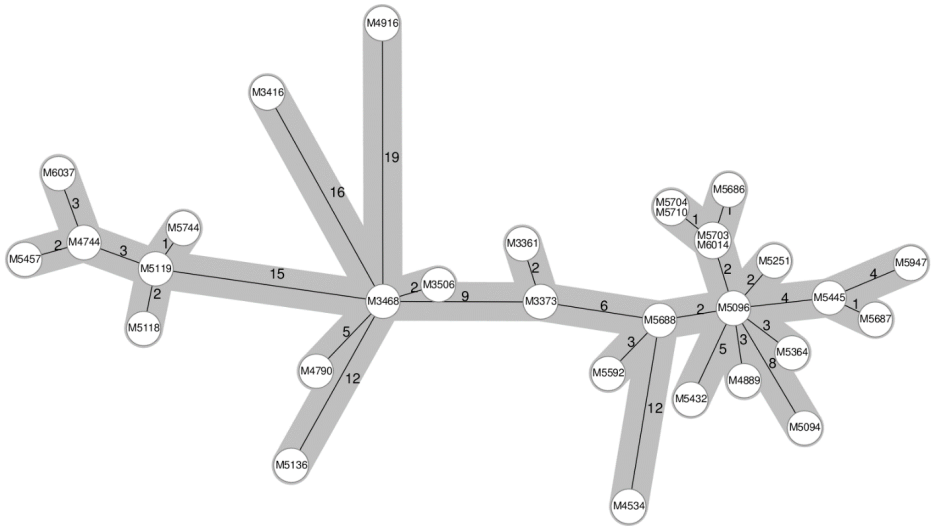

Figure 2: Ridom SeqSphere+ cgMLST minimum spanning tree of a five-year long hospital-related outbreak. The numbers on the edges between the nodes denote the allelic distances between samples.

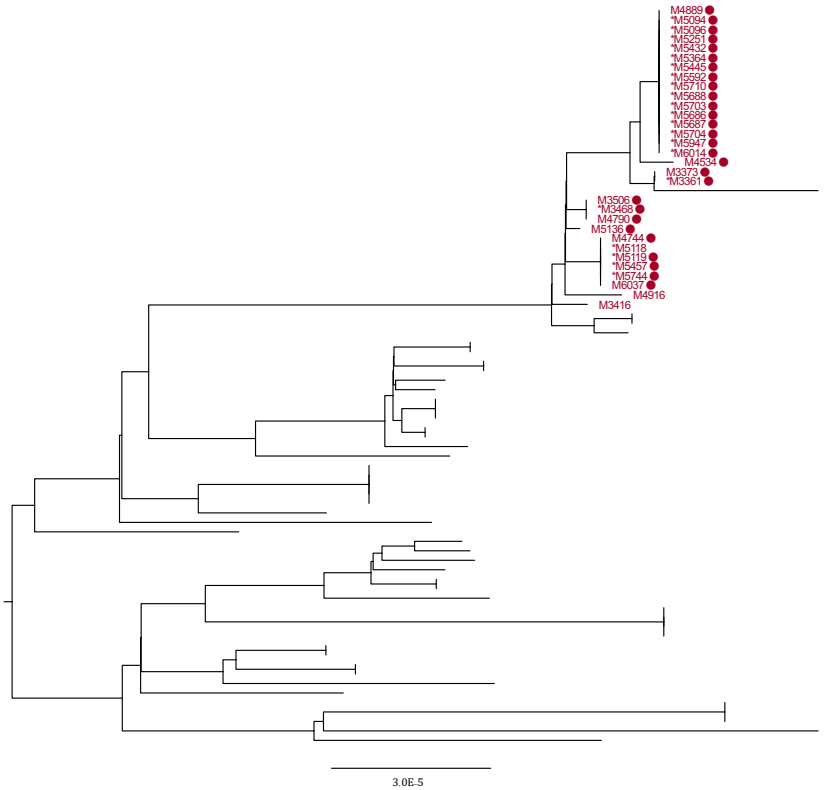

Figure 3: Maximum likelihood phylogenetic tree made with reference Staphylococcus aureus M3261, the samples in red are part of a known hospital outbreak, with supporting epidemiological data denoted with red dots. The asterisk denotes samples in the technical clusters. The scale bar is the equivalent of 0.00003 substitutions per site.

References

[1] A. Cassini *et al.*, "Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis," *Lancet Infect. Dis.*, vol. 19, no. 1, pp. 56–66, Jan. 2019.

[2] US CDC, "Antibiotic resistance threats in the United States, 2019," Atlanta, Georgia, Nov. 2019.

[3] H. Humphreys and D. C. Coleman, "Contribution of whole-genome sequencing to understanding of the epidemiology and control of meticillin-resistant Staphylococcus aureus," *Journal of Hospital Infection*, vol. 102, no. 2. W.B. Saunders Ltd, pp. 189–199, 01-Jun-2019.

[4] C. L. Correa-Martinez, H. Tönnies, N. J. Froböse, A. Mellmann, and S. Kampmeier, "Transmission of Vancomycin-Resistant Enterococci in the Hospital Setting: Uncovering the Patient–Environment Interplay," *Microorganisms*, vol. 8, no. 2, p. 203, Jan. 2020.

[5] M. W. Allard *et al.*, "Genomics of foodborne pathogens for microbial food safety," *Current Opinion in Biotechnology*, vol. 49. Elsevier Ltd, pp. 224–229, 01-Feb-2018.

[6] S. Davis *et al.*, "CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data," *PeerJ Comput. Sci.*, vol. 1, p. e20, Aug. 2015.

[7] T. Dallman *et al.*, "SnapperDB: a database solution for routine sequencing analysis of bacterial

isolates," *Bioinformatics*, vol. 81, no. 12, pp. 3946–3952, Apr. 2018.
[8] J. Szarvas, J. Ahrenfeldt, J. L. B. Cisneros, M. C. F. Thomsen, F. M. Aarestrup, and O. Lund, "Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform," *Commun. Biol.*, vol. 3, no. 1, p. 137, 2020.
[9] "Pathogen Detection - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/pathogens/. [Accessed: 27-Jun-2018].
[10] N.-F. Alikhan, Z. Zhou, M. J. Sergeant, and M. Achtman, "A genomic overview of the population structure of Salmonella," *PLOS Genet.*, vol. 14, no. 4, p. e1007261, Apr. 2018.
[11] "Pathogenwatch | A Global Platform for Genomic Surveillance." [Online]. Available: https://pathogen.watch/. [Accessed: 28-May-2019].
[12] M. Bartels *et al.*, "Monitoring meticillin resistant Staphylococcus aureus and its spread in Copenhagen, Denmark, 2013, through routine whole genome sequencing," *Eurosurveillance*, vol. 20, no. 17, p. 21112, Apr. 2015.
[13] M. Pinholt *et al.*, "Multiple hospital outbreaks of vanA Enterococcus faecium in Denmark, 2012-13, investigated by WGS, MLST and PFGE.," *J. Antimicrob. Chemother.*, vol. 70, no. 9, pp. 2474–82, Sep. 2015.
[14] A. Bankevich *et al.*, "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–77, May 2012.
[15] I. M. Rubin *et al.*, "A Sporadic Four-Year Hospital Outbreak of a ST97-IVa MRSA With Half of the Patients First Identified in the Community," *Front. Microbiol.*, vol. 9, no. JUL, p. 1494, Jul. 2018.
[16] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–74, Jan. 2015.
[17] S. R. Leopold, R. V. Goering, A. Witten, D. Harmsen, and A. Mellmann, "Bacterial whole-genome sequencing revisited: Portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes," *J. Clin. Microbiol.*, vol. 52, no. 7, pp. 2365–2370, 2014.
[18] B. G. U. Ramsing *et al.*, "First Outbreak with MRSA in a Danish Neonatal Intensive Care Unit: Risk Factors and Control Procedures," *PLoS One*, vol. 8, no. 6, Jun. 2013.

Supplementary material

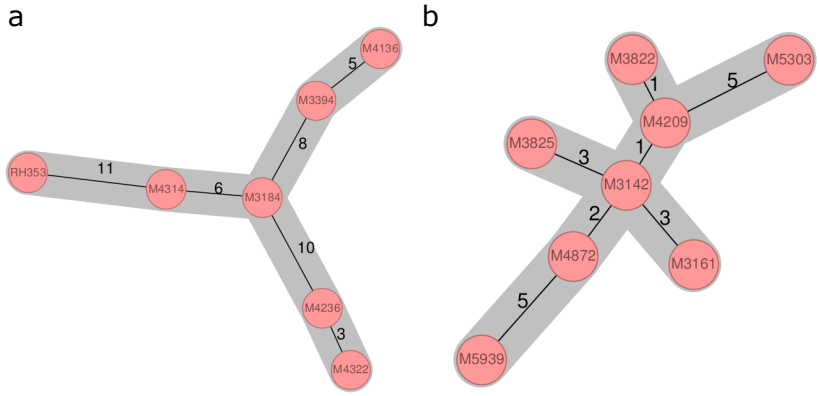

Figure S1: Ridom SeqSphere+ cgMLST minimum spanning trees, the numbers on the edges between the nodes denote the allelic distances between samples. (A) MRSA ST1 cluster shown in *Figure 1*A. (B) MRSA ST1 cluster shown in *Figure 1*B.

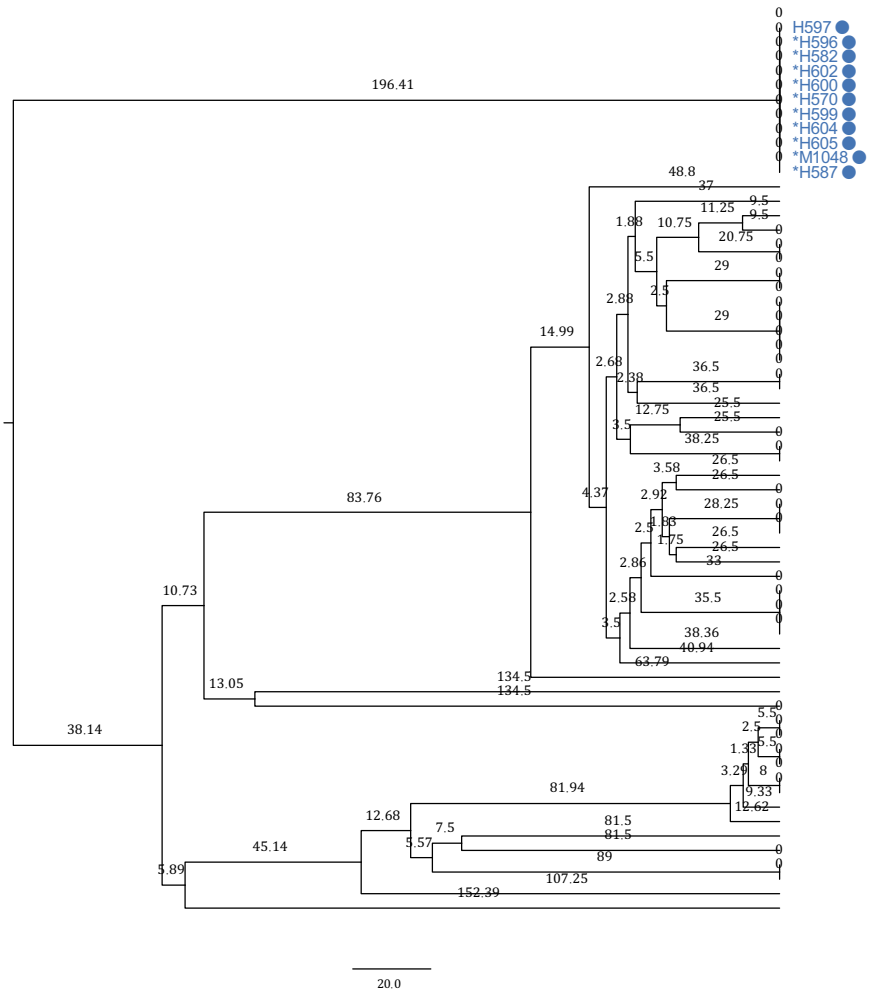

Figure S2: Subtree of the UPGMA tree made on the reference sequence Staphylococcus aureus subsp. aureus 11819-97, highlighting the isolates from the ST1835 outbreak. Scale bar is in single nucleotide differences.

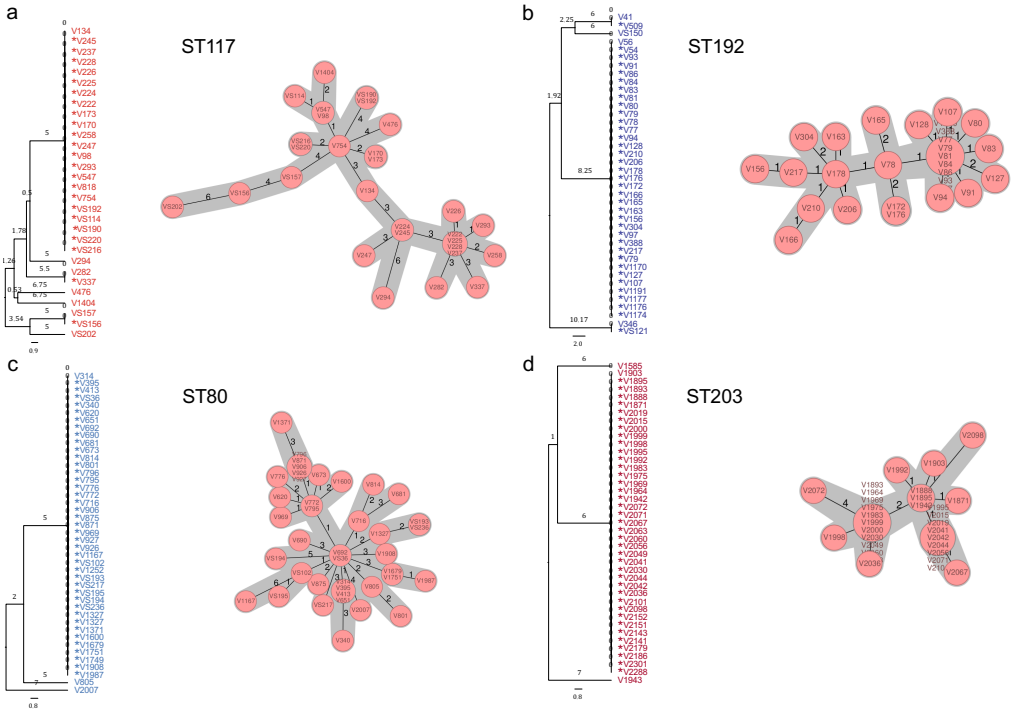

Figure S3: Subsections of the *Enterococcus faecium Aus0085* UPGMA tree and the corresponding samples that are available in the Ridom SeqSphere+ cgMLST minimum spanning trees. The asterisk denotes samples in the PAPABAC technical clusters. Scale bars are given in single nucleotide differences in the SNP tree, while the numbers on the edges between the nodes of the cgMLST cluster denote the allelic distances between samples. (A) *E. faecium* ST117 cluster. (B) *E. faecium* ST192 cluster. (C) *E. faecium* ST80 cluster. (D) *E. faecium* ST203 cluster.

Visualization of norovirus diversity and evolution trough ever-growing trees

Judit Szarvas[1], Frank M. Aarestrup[1], Marion Koopmans[2], Miranda de Graaf[2]*, Ole Lund[1]*

*contributed equally to this manuscript.

[1]Research group for Genomic Epidemiology, Division for Global Surveillance, National Food Institute, Technical University of Denmark

[2]Department of Viroscience, Erasmus MC, The Netherlands

1. Abstract

Molecular surveillance of norovirus caused human gastroenteritis produces a large and increasing number of sequences each year. The overview of the different genogroups and variants, and monitoring of emerging lineages is a non-trivial task. In this study, we developed Krummholz, an open-source bioinformatic pipeline, that gathers sequences from NCBI Genbank as default, and visualizes the evolution of viral sequences. When applied to norovirus VP1 sequences in Genbank, the method created a phylogenetic tree for each genogroup. The spread of the pandemic variants of GII.4, and the expansion of the GII.P16 polymerase positive GII.2 lineage could be observed. New lineages emerge and propagate, which is, in some cases such as the latter case, due to a recombination event between ORF1 and ORF2. For the GII.3 genogroup, a maximum likelihood phylogeny is also inferred, which confirms the polymerase subtype-based and temporal clustering of isolates.

2. Introduction

Norovirus is a common cause of viral gastroenteritis in humans. Genetically the genus is highly divers and can be subdivided in ten genogroups (GI-GX). GI and GII can be further subdivided in more than 40 genotypes (Chhabra et al., 2019). Humans are infected by GI, GII and GIV, while the other genogroups infect animal species (Vinje, 2015). The norovirus genome is approximately 7,000bp long and comprises three open reading frames, encoding a polyprotein (ORF1), capsid protein (ORF2) and a minor capsid protein (ORF3) (Thorne & Goodfellow, 2014). Since there is frequent recombination between ORF1 and ORF2 a dual typing system is adopted where both the polymerase and capsid gene are typed (Kroneman et al., 2013). Norovirus genotyping was recently revisited to accommodate the current known diversity of noroviruses (Chhabra et al., 2019).

International surveillance networks such as NoroNet (www.noronet.nl) and Calicinet have been used to capture the diversity of norovirus circulating in Europe and Asia and have led to important insights in the epidemiology of these viruses (Cannon et al., 2017; van Beek et al., 2018).
The high genetic diversity of norovirus made it technically challenging to obtain sequences for the complete genome or ORFs. However, with next generation sequencing becoming more widely available and cheaper there is an is an upward trend in the amount of complete genome sequences that are being submitted to Genbank at the National Center for Biotechnology Information (NCBI) in recent years. Currently > 48,000 norovirus sequences have been uploaded to NCBI, albeit with a large proportion of short sequences from the polymerase or capsid region. Many of the submitted norovirus sequences do not contain (correct) information on the norovirus genotype, making it challenging to download genotype-specific datasets to investigate the worldwide evolution of these viruses.

With the increasing amount of data there is the need to be able to analyze this data in a more systematic manner. In studies of gastroenteritis in cohorts or outbreaks, the majority of human norovirus infections are caused by the GII.4 genotype (Ahmed et al., 2014). For this particular genotype we observe antigenic evolution over time, with new variants emerging that can evade herd immunity (Lindesmith et al., 2012; Lindesmith et al., 2008). The genetic evolution is epochal and these epidemic variants differed by up to 25 amino acid mutations from their predecessors (Tohma, Lepore, Gao, Ford-Siltz, & Parra, 2019). Thus far, for the non-GII.4 genotypes linear evolution, without replacement among variants, is observed with minimal changes at the protein level and a higher ratio of synonymous compared to non-synonymous substitutions (Tohma, Lepore,

Ford-Siltz, & Parra, 2018). To perform such analyses in a real time manner using sequence data available through NCBI and the dedicated norovirus databases, it is currently necessary to download all norovirus sequences, type these viruses, select the genotypes of interest and infer phylogenetic trees of selected sample sets, which is a time-consuming process. To evaluate the emergence of potentially interesting lineages it would be ideal to visualize newly uploaded sequences with metadata such as location and sample date, without the need to re-run trees for all the different norovirus genotypes.

To provide such a solution, we have now developed a bioinformatics pipeline to generate "krummholz" trees similar to the bacterial "evergreen" trees (Szarvas et al., 2020). Through this pipeline we extract all sequences of interest from Genbank, generate a curated reference strain database that represents all downloaded diversity. Next, phylogenetic trees are build based on reference strains, that allow targeted recalculation of phylogenies when new sequences are added to the dataset, and new trees could be started just by adding new reference strains to the database. The trees are continuously published on www.noronet.nl

3. Methods

Data

Consensus viral nucleotide sequences of the Norovirus genus were downloaded from National Center for Biotechnology Information (NCBI) Nucleotide database, using the following query string: "txid142786[Organism] AND ("1400"[SLEN] : "8000"[SLEN])". First 6887 nucleotide sequences were downloaded on 2019/07/16, then 113 additional ones on 2019/09/10, and 565 on 2020/02/16. Metadata regarding the sample collection date and location was extracted from the GenBank entries.

Template database construction

A total of 99 full length reference capsid protein VP1 sequences were selected ((Chhabra et al., 2019; Kroneman et al., 2011; Vinje, 2015), Supplementary Table S1) and supplemented with 6237 length-sorted, full-length or nearly full-length VP1 sequences from GenBank (≥ 1430 nt). These sequences were homology reduced with KMA (Clausen, Aarestrup, & Lund, 2018) according to the Hobohm 1 algorithm (Hobohm, Scharf, Schneider, & Sander, 1992) using 16-mers at the following thresholds: 0.6143 (97%), 0.4401 (95%), 0.1853 (90%), 0.0743 (85%), 0.0281 (80%). The kept sequences were aggregated and genotyped with Norovirus Typing Tool Version 2.0 (Kroneman et al., 2011). The results for each reference set were analyzed with a custom R script. The genotyping

results were manually checked and corrected, where necessary. Template names were constructed by concatenating the GenBank accession numbers and the genotype descriptors. The template names were later updated to reflect the changes in nomenclature (Chhabra et al., 2019).

Krummholz bioinformatics pipeline

The workflow was modified from our bacterial phylogenomic analysis pipeline (Szarvas et al., 2020) to be suitable for assembled sequences in the format of FASTA sequences, rather than raw reads in the format of FASTQ sequences. The input sequences were mapped to the reference database using the sparse option of KMA, without down sampling the data using a prefix. Hits were accepted as templates, if the template K-mer coverage was higher than 7.43% (corresponding to 85% sequence identity), and in case of multiple hits, if the coverage of the hit was higher than 50% of the sum of the threshold and the highest hit. Each sequence was aligned to its accepted template(s) with KMA using the *cge* flag, to gain template length consensus sequences (stripping out insertions relative to the template). If the consensus sequence had more than 10% unknown bases, then it was excluded from further analysis.

The consensus sequences were grouped based on the templates. In each set, where the existing set size was less than 100 sequences, the genetic distances were calculated on all positions with no unknown bases. On larger sets, the calculation was carried out on positions where the fraction of the known bases was larger than the mean fraction for the set minus one standard deviation. Identical sequences were clustered together.

Phylogenetic trees were inferred using the Neighbor program from the PHYLIP package v3.697 (http://evolution.genetics.washington.edu/phylip.html) and IQ-tree v1.6 (Nguyen, Schmidt, von Haeseler, & Minh, 2015). Selected phylogenetic trees were exported into Microreact (Argimon et al., 2016).

The scripts are publicly available at https://bitbucket.org/jszarvas/viral_surveillance/src/master/

Maximum likelihood tree

Using the accession numbers from the GII.3 krummholz tree all sequences were downloaded from NCBI and typed using the noronet typing tool. A maximum-likelihood tree of the VP1 gene was inferred by IQ-tree software under the symmetrical model (SYM)+R4 (Minh et al., 2020) and visualized using Figtree version v1.4.4.

4. Results and Discussion

Selection of reference sequences

An initial overview of the norovirus VP1 sequence diversity was gained during the template database construction. In total, 572 different sequences were considered as templates as we tested the different homology reducing thresholds. Sequence similarity thresholds of 90% or more yielded several templates for the same genotype based on the classification at that date, with the majority of them being GII and GI genotypes. Meanwhile a low sequence similarity threshold of 80% merged some genotypes (Supplementary Table S2). Therefore, we selected the set containing 74 sequences, that were at most 85% identical to each other, as our template database. These template sequences were compared to the different genogroups and genotypes that have been determined previously (Chhabra et al., 2019). A cut-off of 85% gave separate templates for each of the previously determined genotypes and genogroups but not for each of the GII.4 variants.

Generating krummholz trees for the norovirus VP1 gene

The major challenge solved here is the automatic and continuous analyses of novel sequences of interest submitted to Genbank, and in real time comparing them to all other sequences already submitted. All norovirus VP1 sequences from Genbank were processed with Krummholz in total 10.6 minutes on eight 2.20GHz cores on a linux server. 63 templates had at least one sequence mapped to it, besides the template itself. Of these, 50 had more than 3 sequence assigned. The largest sequence set was for the norovirus GII.4 genotype with 3,450 sequences, followed by GII.2 and GII.17, with 911 and 861, respectively. This is expected as GII.4 has been the predominant genotype causing outbreaks in humans for the past decades. Overall, the genotypes associated with human hosts are overrepresented in Genbank, compared to sequences from canine, bovine or other mammalian hosts (Figure *1*).

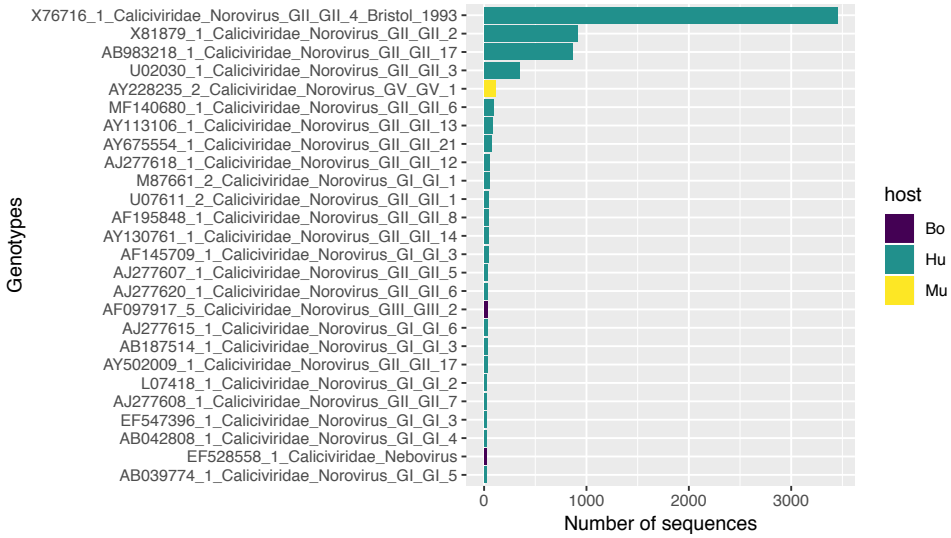

Figure 1: Number of sequences in each template set that had at least 20 sequences, colored by the host associated with each genotype. (Bo: bovine, Hu: human, Mu: rodents)

Visualization of the krummholz trees

To visualize the phylogenetic trees generated by Krummholz, we have used Microreact, an open data visualization and sharing tool for genomic epidemiology, but the data output might also be visualized using other tools. This allows us to look at the worldwide evolution of individual norovirus genotypes over time (Table S3). We have shown example visualizations of four of the most predominant human norovirus genotypes (Figure 2). The software provides an overview of the worldwide distribution of the strains on a map, a timeline of the date of sampling of the strains, and a phylogenetic tree. The nodes representing the samples are colored based on the location or time of isolation. Clicking on a single leaf on the tree will highlight its position on the map and in the timeline. Data points can be selected within the map, tree and timeline, enabling the download of datasets of interest for further analysis. By selecting play on the timeline, the user is able to 'replay' the potential chain of events and explore the relationships between the temporal and spatial spread of data.

For most norovirus genotypes we observed multiple lineages that can co-circulate worldwide. Some lineages will result in evolutionary dead-end clades over time, while others are the predecessor for new variants (Figure 2). In some cases, the separate lineages of the VP1 represent different

polymerase genotype combinations (Tohma et al., 2018). This is also evident when looking at the chain of events for norovirus genotype GII.4, GII.2, GII.17 and GII.3 (Figure 2 and Figure 3).

For GII.4, we can observe the variants that have caused pandemics in the past decades with frequent shifts of variant predominance (Figure 2a)(Siebenga et al., 2007; Tohma et al., 2019). For GII.2, we see a large increase of VP1 sequences in the database after 2016 that form a separate lineage, reflecting the increased number of outbreaks that was observed after it gained an GII.P16 polymerase (Figure 2b) (Barclay et al., 2019; Niendorf et al., 2017; van Beek et al., 2018). For GII.17 (Figure 2c), a replay of the sequence data shows the sporadic detection of GII.17 before 2014 and the emergence of Kawasaki 308–like in 2014 (GII.17 Kawasaki), that became predominant genotype in parts of Asia and circulated worldwide as well as the co-circulating Kawasaki 323–like strains in 2014-2015 (Chan et al., 2017; de Graaf et al., 2015; Fu et al., 2015; Tohma et al., 2018).
For GII.3, we can observe several clusters over that in part co-circulate over time. To perform an in-dept analysis of the GII.3 sequences and compare the maximum likelihood tree with the krummholz tree visualized in Microreact, we downloaded all sequences from NCBI, determined the polymerase genotypes using the NoroNet typing tool and inferred a maximum likelihood tree using IQ-tree (Figure 3). In parallel the krummholz GII.3 tree file was downloaded and colored based on the polymerase genotypes (Figure S1). Both trees revealed clustering based on the polymerase genotype, as well as temporal clustering. The clustering of VP1 sequences with different polymerase types are indicative of recombination between the ORF1 and ORF2.

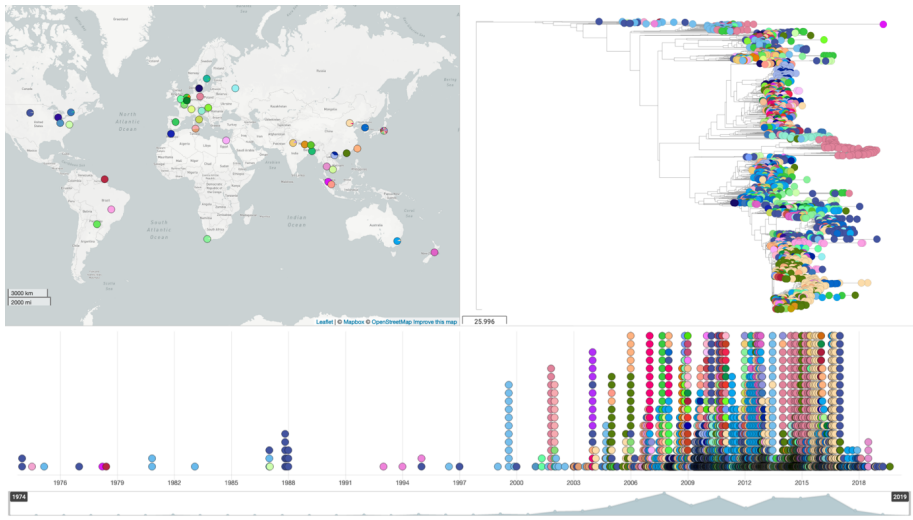

Figure 2a: Krummholz overview of GII.4 VP1 sequences

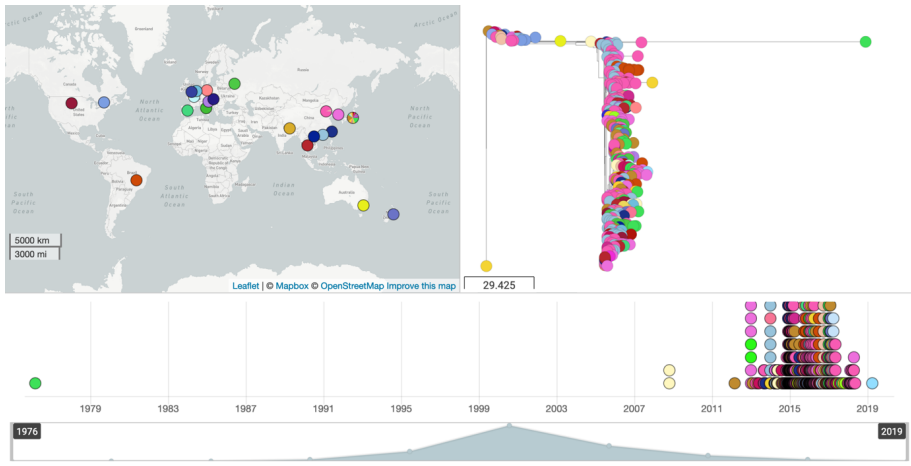

Figure 2b: Krummholz overview of GII.17 VP1 sequences

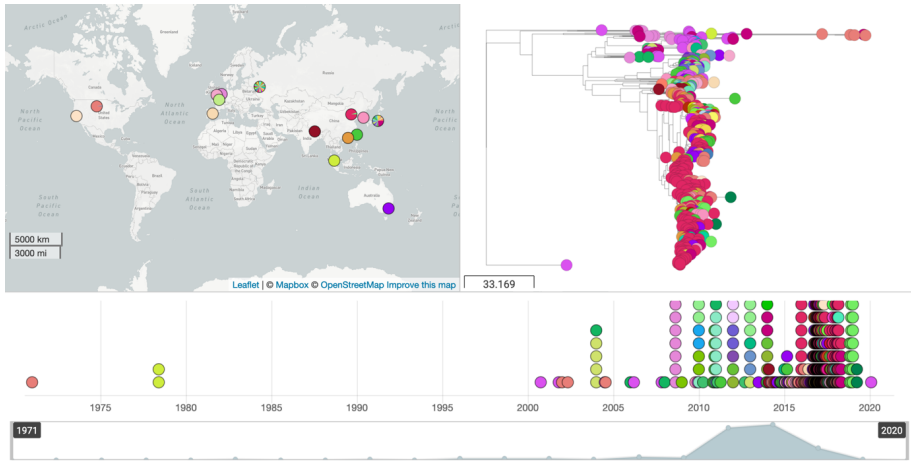

Figure 2c: Krummholz overview of GII.2 VP1 sequences

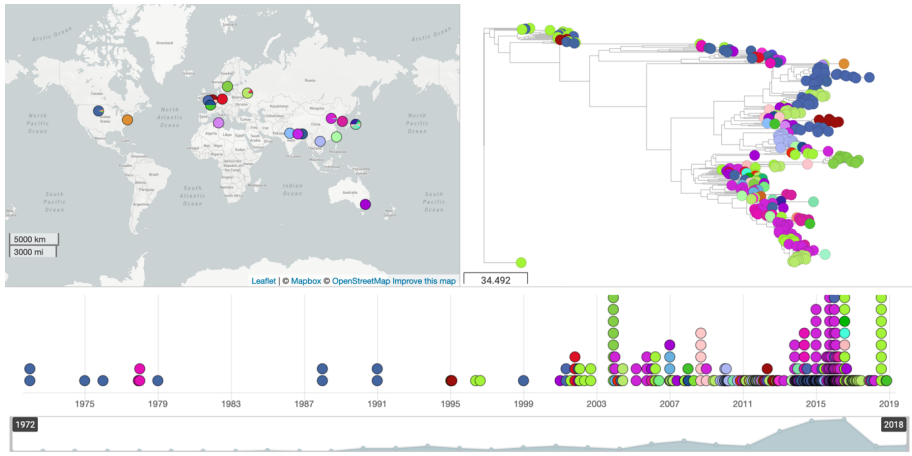

Figure 2d: Krummholz overview of GII.3 VP1 sequences

Figure 3: Maximum likelihood tree inferred for GII.3 VP1 sequences, tip labels colored based on the polymerase subtype, where available. Selected bootstrap values >70 are depicted. Scale bars indicate nucleotide substitutions per site.

Real-time tracking of pathogen evolution through phylodynamics analysis and interactive visualization platform of Nextstrain has led to important insights in viral evolution of emerging viruses (Alm et al., 2020; Dyrdak, Mastafa, Hodcroft, Neher, & Albert, 2019; Hadfield et al., 2018). In addition to the phylodynamics and visualization tools, Krummholz provides a method for collecting and the classification of strains to provide a database for these tools.

In summary, Krummholz has been steadily processing norovirus sequence data collected worldwide in real time. Excluding the download time and the optional maximum likelihood based phylogenetic

inference, the whole analysis is done in a few minutes, even for the larger template sets. The Evergreen method was initially developed for rapid subtyping and continuous phylogenomic analysis of bacterial samples, suited for large-scale surveillance. The present study shows that the method could be adapted to work on viral consensus sequences, and that the generated krummholz trees can also be used for the analysis of rapidly evolving viral sequences. The krummholz trees on Noronet can simplify global surveillance of foodborne viral pathogens with high accuracy, speed, stability and practicality, which will in turn facilitate information needed for vaccine development, outbreak management and the development of diagnostic tools.

5. References

Ahmed, S. M., Hall, A. J., Robinson, A. E., Verhoef, L., Premkumar, P., Parashar, U. D., . . . Lopman, B. A. (2014). Global prevalence of norovirus in cases of gastroenteritis: a systematic review and meta-analysis. *The Lancet. Infectious diseases, 14*(8), 725-730. doi:10.1016/S1473-3099(14)70767-4

Alm, E., Broberg, E. K., Connor, T., Hodcroft, E. B., Komissarov, A. B., Maurer-Stroh, S., . . . group*, G. E. (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Euro Surveill, 25*(32). doi:10.2807/1560-7917.ES.2020.25.32.2001410

Argimon, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., . . . Aanensen, D. M. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom, 2*(11), e000093. doi:10.1099/mgen.0.000093

Barclay, L., Cannon, J. L., Wikswo, M. E., Phillips, A. R., Browne, H., Montmayeur, A. M., . . . Vinje, J. (2019). Emerging Novel GII.P16 Noroviruses Associated with Multiple Capsid Genotypes. *Viruses, 11*(6). doi:10.3390/v11060535

Cannon, J. L., Barclay, L., Collins, N. R., Wikswo, M. E., Castro, C. J., Magana, L. C., . . . Vinje, J. (2017). Genetic and Epidemiologic Trends of Norovirus Outbreaks in the United States from 2013 to 2016 Demonstrated Emergence of Novel GII.4 Recombinant Viruses. *J Clin Microbiol, 55*(7), 2208-2221. doi:10.1128/JCM.00455-17

Chan, M. C. W., Hu, Y., Chen, H., Podkolzin, A. T., Zaytseva, E. V., Komano, J., . . . Chan, P. K. S. (2017). Global Spread of Norovirus GII.17 Kawasaki 308, 2014-2016. *Emerg Infect Dis, 23*(8), 1359-1354. doi:10.3201/eid2308.161138

Chhabra, P., de Graaf, M., Parra, G. I., Chan, M. C., Green, K., Martella, V., . . . Vinje, J. (2019). Updated classification of norovirus genogroups and genotypes. *J Gen Virol, 100*(10), 1393-1406. doi:10.1099/jgv.0.001318

Clausen, P., Aarestrup, F. M., & Lund, O. (2018). Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics, 19*(1), 307. doi:10.1186/s12859-018-2336-6

de Graaf, M., van Beek, J., Vennema, H., Podkolzin, A. T., Hewitt, J., Bucardo, F., . . . Koopmans, M. P. (2015). Emergence of a novel GII.17 norovirus - End of the GII.4 era? *Euro Surveill, 20*(26). doi:10.2807/1560-7917.es2015.20.26.21178

Dyrdak, R., Mastafa, M., Hodcroft, E. B., Neher, R. A., & Albert, J. (2019). Intra- and interpatient evolution of enterovirus D68 analyzed by whole-genome deep sequencing. *Virus Evol, 5*(1), vez007. doi:10.1093/ve/vez007

Fu, J., Ai, J., Jin, M., Jiang, C., Zhang, J., Shi, C., . . . Zhu, Y. (2015). Emergence of a new GII.17 norovirus variant in patients with acute gastroenteritis in Jiangsu, China, September 2014 to March 2015. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 20*(24). Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26111236

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., . . . Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics, 34*(23), 4121-4123. doi:10.1093/bioinformatics/bty407

Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci, 1*(3), 409-417. doi:10.1002/pro.5560010313

Kroneman, A., Vega, E., Vennema, H., Vinje, J., White, P. A., Hansman, G., . . . Koopmans, M. (2013). Proposal for a unified norovirus nomenclature and genotyping. *Arch Virol, 158*(10), 2059-2068. doi:10.1007/s00705-013-1708-5

Kroneman, A., Vennema, H., Deforche, K., v d Avoort, H., Penaranda, S., Oberste, M. S., . . . Koopmans, M. (2011). An automated genotyping tool for enteroviruses and noroviruses.

Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology, 51(2), 121-125. doi:10.1016/j.jcv.2011.03.006

Lindesmith, L. C., Beltramello, M., Donaldson, E. F., Corti, D., Swanstrom, J., Debbink, K., . . . Baric, R. S. (2012). Immunogenetic mechanisms driving norovirus GII.4 antigenic variation. *PLoS pathogens, 8*(5), e1002705. doi:10.1371/journal.ppat.1002705

Lindesmith, L. C., Donaldson, E. F., Lobue, A. D., Cannon, J. L., Zheng, D. P., Vinje, J., & Baric, R. S. (2008). Mechanisms of GII.4 norovirus persistence in human populations. *PLoS medicine, 5*(2), e31. doi:10.1371/journal.pmed.0050031

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol, 37*(5), 1530-1534. doi:10.1093/molbev/msaa015

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol, 32*(1), 268-274. doi:10.1093/molbev/msu300

Niendorf, S., Jacobsen, S., Faber, M., Eis-Hubinger, A. M., Hofmann, J., Zimmermann, O., . . . Bock, C. T. (2017). Steep rise in norovirus cases and emergence of a new recombinant strain GII.P16-GII.2, Germany, winter 2016. *Euro Surveill, 22*(4), 30447. doi:10.2807/1560-7917.ES.2017.22.4.30447

Siebenga, J. J., Vennema, H., Renckens, B., de Bruin, E., van der Veer, B., Siezen, R. J., & Koopmans, M. (2007). Epochal evolution of GGII.4 norovirus capsid proteins from 1995 to 2006. *J Virol, 81*(18), 9932-9941. doi:10.1128/JVI.00674-07

Szarvas, J., Ahrenfeldt, J., Cisneros, J. L. B., Thomsen, M. C. F., Aarestrup, F. M., & Lund, O. (2020). Large scale automated phylogenomic analysis of bacterial isolates and the Evergreen Online platform. *Commun Biol, 3*(1), 137. doi:10.1038/s42003-020-0869-5

Thorne, L. G., & Goodfellow, I. G. (2014). Norovirus gene expression and replication. *The Journal of general virology, 95*(Pt 2), 278-291. doi:10.1099/vir.0.059634-0

Tohma, K., Lepore, C. J., Ford-Siltz, L. A., & Parra, G. I. (2018). Evolutionary dynamics of non-GII genotype 4 (GII.4) noroviruses reveal limited and independent diversification of variants. *J Gen Virol, 99*(8), 1027-1035. doi:10.1099/jgv.0.001088

Tohma, K., Lepore, C. J., Gao, Y., Ford-Siltz, L. A., & Parra, G. I. (2019). Population Genomics of GII.4 Noroviruses Reveal Complex Diversification and New Antigenic Sites Involved in the Emergence of Pandemic Strains. *mBio, 10*(5). doi:10.1128/mBio.02202-19

van Beek, J., de Graaf, M., Al-Hello, H., Allen, D. J., Ambert-Balay, K., Botteldoorn, N., . . . NoroNet. (2018). Molecular surveillance of norovirus, 2005-16: an epidemiological analysis of data collected from the NoroNet network. *Lancet Infect Dis, 18*(5), 545-553. doi:10.1016/S1473-3099(18)30059-8

Vinje, J. (2015). Advances in laboratory methods for detection and typing of norovirus. *Journal of clinical microbiology, 53*(2), 373-381. doi:10.1128/JCM.01535-14

6. Supplemental Material

Table S1: Full length VP1 sequences

Genbank accession number	**Strain**
AY237415	Human calicivirus strain Mc37
AB126320	SW/NV/swine43/JP
AB032758	Chitta virus
AB078334	Norwalk-like virus Kashiwa47
GQ856465	Norovirus Hu/GII.4/Beijing/55028/2007/CHN
GQ856474	Norovirus Hu/GII.4/Beijing/55161/2008/CHN

GQ856476	Norovirus Hu/GII.4/Beijing/55171/2008/CHN
AB983218	Norovirus Hu/GII/JP/2014/GII.P17_GII.17/Kawasaki323
DQ438972	Norovirus Hu/NoV/Katrina-17/2005/US
KC597139	Norovirus Hu/GII.17/C142/1978/GUF
LC037415	Norovirus Hu/GII/JP/2015/GII.P17_GII.17/Kawasaki308
AY823305	Norovirus swine/GII/OH-QW125/03/US
AY823307	Norovirus swine/GII/OH-QW218/03/US
AJ277606	Norovirus HU/NLV/Girlington/93/UK
AB542917	Norovirus Hu/OC07118/2007/JP
AB542915	Norovirus Hu/OC05024/2005/JP
GQ856469	Norovirus Hu/GII.4/Beijing/53931/2007/CHN
AJ844470	Norovirus Hu/NV/Chiba/040502/2004/JP
KJ196291	Norovirus GII/Hu/JP/2011/GII/Yuzawa/Gira2HS
AY134748	Snow Mountain virus
EU187437	Norovirus Hu/5017.34/2003/JPN
AF397156	Human calicivirus NLV/MOH/99
AB039778	Norwalk-like virus Saitama U16
AF414409	Norwalk-like virus NLV/Gwynedd/273/1994/US RNA
AB039780	Norwalk-like virus Saitama U25
DQ379715	Norovirus Hu/GII/Goulburn Valley G5175 C/1983/AUS
EU360814	Norovirus Bo/GIII.1/Aba-Z5/2002/HUN
L23828	Norwalk virus (SRSV-KY-89/89/J)
AJ277610	Norovirus HU/NLV/Whiterose/96/UK
AJ277612	Norovirus HU/NLV/Birmingham/93/UK
AY038598	Human calicivirus NLV/VA98115/1998
GQ856470	Norovirus Hu/GII.4/Beijing/54108/2007/CHN
GQ856471	Norovirus Hu/GII.4/Beijing/54114/2007/CHN
GQ856473	Norovirus Hu/GII.4/Beijing/55042/2007/CHN
AJ277616	Norovirus HU/NLV/Valetta/95/Malta
AJ277621	Norovirus HU/NLV/Thistlehall/90/UK
AJ313030	Norovirus Hu/NLV/Queen's Arms/Leeds/92/UK
AF414406	Norwalk-like virus NLV/Appalachicola Bay/318/1995/US
AM263418	Norovirus Hu/GI/Babbacombe/1996/GBR
AF538678	Human calicivirus NLV/VA497/1999/US
AY675555	Norovirus NLV/IF2036/2003/Iraq
GU299761	Norovirus Hu/GI.8/2008890321/2008/US
GU296356	Norovirus Hu/GI/Lilla Edet/p2/SWE/2008

364
365

366

Table S2: Genotypes genotype according to the genotyping scheme in July 2019, represented as template for 85% and 80% homology reduction thresholds

#	Genotype	85%	80%
1	Caliciviridae Nebovirus unclassified	X	X
2	Caliciviridae Norovirus GI Could not assign	X	X
3	Caliciviridae Norovirus GI GI.1	X	X
4	Caliciviridae Norovirus GI GI.2	X	X
5	Caliciviridae Norovirus GI GI.3	X	X
6	Caliciviridae Norovirus GI GI.4	X	X
7	Caliciviridae Norovirus GI GI.5	X	X
8	Caliciviridae Norovirus GI GI.6	X	X
9	Caliciviridae Norovirus GI GI.7	X	X
10	Caliciviridae Norovirus GI GI.8	X	-
11	Caliciviridae Norovirus GI GI.9	X	X
12	Caliciviridae Norovirus GII Could not assign	X	X
13	Caliciviridae Norovirus GII GII.1	X	X
14	Caliciviridae Norovirus GII GII.10	X	X
15	Caliciviridae Norovirus GII GII.11	X	X
16	Caliciviridae Norovirus GII GII.12	X	-
17	Caliciviridae Norovirus GII GII.13	X	X
18	Caliciviridae Norovirus GII GII.14	X	-
19	Caliciviridae Norovirus GII GII.16	X	-
20	Caliciviridae Norovirus GII GII.17	X	X
21	Caliciviridae Norovirus GII GII.18	X	X
22	Caliciviridae Norovirus GII GII.19	X	X
23	Caliciviridae Norovirus GII GII.2	X	X
24	Caliciviridae Norovirus GII GII.20	X	X
25	Caliciviridae Norovirus GII GII.21	X	X
26	Caliciviridae Norovirus GII GII.22	X	X
27	Caliciviridae Norovirus GII GII.24	X	X
28	Caliciviridae Norovirus GII GII.25	X	-
29	Caliciviridae Norovirus GII GII.3	X	X
30	Caliciviridae Norovirus GII GII.4 Bristol_1993	X	X
31	Caliciviridae Norovirus GII GII.4 Could not assign	X	-
32	Caliciviridae Norovirus GII GII.4 Den_Haag_2006b	-	-
33	Caliciviridae Norovirus GII GII.4 New_Orleans_2009	-	-
34	Caliciviridae Norovirus GII GII.5	X	X
35	Caliciviridae Norovirus GII GII.6	X	X
36	Caliciviridae Norovirus GII GII.7	X	X

37	Caliciviridae Norovirus GII GII.8	X	X
38	Caliciviridae Norovirus GII GII.9	X	X
39	Caliciviridae Norovirus GIII	X	X
40	Caliciviridae Norovirus GIV	X	X
41	Caliciviridae Norovirus GV	X	X
42	Caliciviridae Norovirus GVII	X	X
43	Caliciviridae Norovirus Gx.15	X	X
44	Caliciviridae Norovirus Gy.23	X	X
45	Caliciviridae Sapovirus GI GI.3	X	X

Table S3: Krummholz trees displayed in Microreact

Template	Version	Url
AJ277615_1_Caliciviridae_Norovirus_GI_GI_6	1581812764	https://microreact.org/project/zsN2zzRcq/
AY228235_2_Caliciviridae_Norovirus_GV_GV_1	1581810473	https://microreact.org/project/iCUCzd37Q/
AF195848_1_Caliciviridae_Norovirus_GII_GII_8	1581812751	https://microreact.org/project/XFIJvnTpY/
AB983218_1_Caliciviridae_Norovirus_GII_GII_17	1581812744	https://microreact.org/project/Vndbry_jr/
AJ277618_1_Caliciviridae_Norovirus_GII_GII_12	1581812742	https://microreact.org/project/Ko6aT_NdC/
AF093797_1_Caliciviridae_Norovirus_GI_GI_6	1581812768	https://microreact.org/project/lZ7f7JhDS/
AY549161_1_Caliciviridae_Norovirus_GIII_GIII_2	1581810624	https://microreact.org/project/KXq4G0WBs/
AF195847_1_Caliciviridae_Norovirus_GIV_GIV_1	1581812773	https://microreact.org/project/RNhRkdzIK/
GQ856469_1_Caliciviridae_Norovirus_GII_GII_25	1581812760	https://microreact.org/project/rqT39ZYZv/
AB042808_1_Caliciviridae_Norovirus_GI_GI_4	1581812729	https://microreact.org/project/JaXYJ6-iy/
U07611_2_Caliciviridae_Norovirus_GII_GII_1	1581812758	https://microreact.org/project/bCc8aT8OO/
AY502010_1_Caliciviridae_Norovirus_GII_GII_16	1581812777	https://microreact.org/project/K7ZAg-d-C/
AJ277614_1_Caliciviridae_Norovirus_GI_GI_5	1581812766	https://microreact.org/project/HyUJcOEGF/
AJ277608_1_Caliciviridae_Norovirus_GII_GII_7	1581812736	https://microreact.org/project/Lm7tcalyq/
AF097917_5_Caliciviridae_Norovirus_GIII_GIII_2	1581812775	https://microreact.org/project/jg7gcbBAt/
AB039774_1_Caliciviridae_Norovirus_GI_GI_5	1581812762	https://microreact.org/project/9gsDYsjwB/
EF528558_1_Caliciviridae_Nebovirus	1581810635	https://microreact.org/project/GwAp7l4zm/
L07418_1_Caliciviridae_Norovirus_GI_GI_2	1581812734	https://microreact.org/project/vapMN1E95/
AF427118_1_Caliciviridae_Norovirus_GII_GII_10	1581810606	https://microreact.org/project/Lsp4ihVvy/
AY130761_1_Caliciviridae_Norovirus_GII_GII_14	1581812755	https://microreact.org/project/OBsABzsfJ/
EU224456_1_Caliciviridae_Norovirus_GIV	1581810609	https://microreact.org/project/8Dh-wfVv2/
FJ875027_1_Caliciviridae_Norovirus_GVI_GVI_1	1581810631	https://microreact.org/project/BK-h72-cb/
JN975492_1_Caliciviridae_Norovirus_GV_GV_1	1581810653	https://microreact.org/project/Hf0e_QJcn/
AY502009_1_Caliciviridae_Norovirus_GII_GII_17	1581812749	https://microreact.org/project/xQBqJvlMs/
AJ277607_1_Caliciviridae_Norovirus_GII_GII_5	1581812756	https://microreact.org/project/sgPdoQbJk/
AF145709_1_Caliciviridae_Norovirus_GI_GI_3	1581812724	https://microreact.org/project/nLpjUwi-s/
HQ637267_1_Caliciviridae_Norovirus_GI_GI_9	1581810647	https://microreact.org/project/32gsBjmir/
GQ443611_1_Caliciviridae_Norovirus_GVI_GVI_2	1581810625	https://microreact.org/project/Y77X4I2tD/

AB074893_1_Caliciviridae_Norovirus_GII_GII_11	1581812747	https://microreact.org/project/C80Vyid3x/
AY038599_2_Caliciviridae_Norovirus_GII_GII_9	1581812752	https://microreact.org/project/w91jhVwny/
AJ844469_1_Caliciviridae_Norovirus_GI_GI_7	1581812771	https://microreact.org/project/hUYyoYyWA/
AB187514_1_Caliciviridae_Norovirus_GI_GI_3	1581812763	https://microreact.org/project/AwrDYB4BK/
AJ277609_1_Caliciviridae_Norovirus_GI_GI_7	1581812767	https://microreact.org/project/trIwX-Uy-/
EF450827_1_Caliciviridae_Norovirus_GIV_GIV_2	1581812634	https://microreact.org/project/qS1bH1M59/
AJ277620_1_Caliciviridae_Norovirus_GII_GII_6	1581812730	https://microreact.org/project/kg-RRqRAo/
X76716_1_Caliciviridae_Norovirus_GII_GII_4_Bristol_1993	1581812646	https://microreact.org/project/fx-M4MbcR/
X81879_1_Caliciviridae_Norovirus_GII_GII_2	1581812726	https://microreact.org/project/M_YzyX6QB/
EU373815_1_Caliciviridae_Norovirus_GII_GII_20	1581810601	https://microreact.org/project/KwDK3J_xK/
AY113106_1_Caliciviridae_Norovirus_GII_GII_13	1581812759	https://microreact.org/project/hN972YC3W/
KY225989_1_Caliciviridae_Norovirus_GII_GII_24	1581812776	https://microreact.org/project/e4q0YhWCN/
MK159169_1_Caliciviridae_Norovirus_GIII	1581810646	https://microreact.org/project/t3w9J2Jx8/
AY130762_1_Caliciviridae_Norovirus_GIX_GIX_1	1581812735	https://microreact.org/project/4eNO4IO__/
EF547396_1_Caliciviridae_Norovirus_GI_GI_3	1581812748	https://microreact.org/project/DenVtT1qj/
M87661_2_Caliciviridae_Norovirus_GI_GI_1	1581812772	https://microreact.org/project/XNwjpfILk/
U04469_1_Caliciviridae_Norovirus_GI_GI_3	1581812770	https://microreact.org/project/l4KUju_uz/
MF140680_1_Caliciviridae_Norovirus_GII_GII_6	1581812753	https://microreact.org/project/OXfaAcOcu/
AY675554_1_Caliciviridae_Norovirus_GII_GII_21	1581812738	https://microreact.org/project/myFosoUjK/
AF538679_1_Caliciviridae_Norovirus_GI_GI_8	1581812739	https://microreact.org/project/hutPZWHQR/
MF352142_1_Caliciviridae_Norovirus_GII_GII_26	1581810648	https://microreact.org/project/cn3NH3e-w/
U02030_1_Caliciviridae_Norovirus_GII_GII_3	1581812732	https://microreact.org/project/fVifdLqDq/
KR232647_2_Caliciviridae_Norovirus_GII_GII_23	1581812740	https://microreact.org/project/UkZwOpk3y/

372

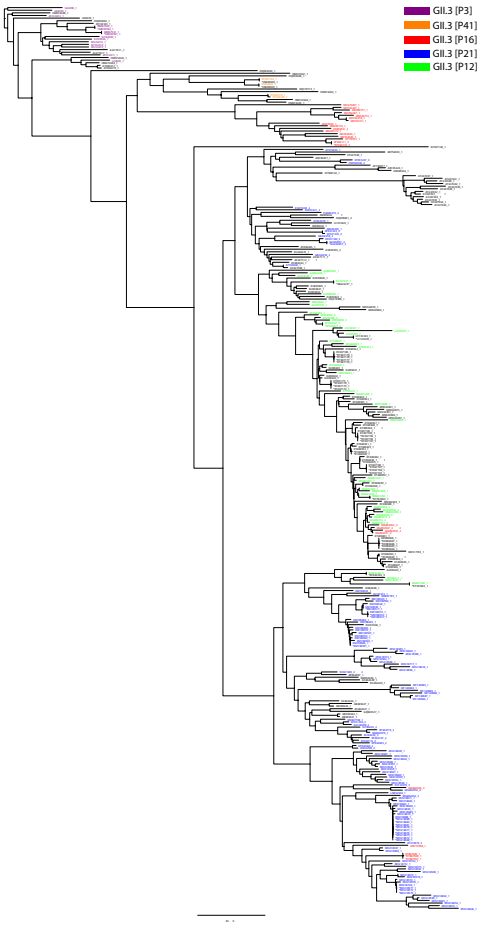

Figure S1: Krummholz tree for GII.3 VP1 sequences, tip labels colored based on the polymerase subtype, where available. Scale bar indicate genetic distance in SNPs.

Danish whole-genome sequenced *Candida albicans* and *Candida glabrata* samples fit into globally prevalent clades

Judit Szarvas[1], Ana Rita Rebelo[1], One Day in Denmark Consortium[2], Frank Møller Aarestrup[1]

1 Research group for Genomic Epidemiology, Division for Global Surveillance, National Food Institute, Technical University of Denmark

2 https://files.dtu.dk/u/nunsJA7p7DiM6ewm/ODID%20Consortium%20August%202020.docx?l

Abstract

Candida albicans and *Candida glabrata* are fungal opportunistic pathogens with increasing incidence in worldwide, and higher than expected prevalence in Denmark. We whole-genome sequenced yeast isolates collected from Danish Clinical Microbiology Laboratories to obtain an overview of the Candida population in the country. The majority of the thirty *C. albicans* isolates were found to belong to three globally prevalent clades, and with one exception, the remaining isolates were also predicted to cluster with samples from other geographical locations. Similarly, most of the eight *C. glabrata* isolates were predicted to be prevalent subtypes. Antifungal susceptibility testing proved all *C. albicans* isolates to be susceptible to both azoles and echinocandins. Two *C. glabrata* isolates presented azole-resistant phenotypes, yet were all susceptible to echinocandins. There is no indication of causality between population structure and resistance phenotypes for either species.

Introduction

Opportunistic fungal infections are an increasing concern in hospital settings, as invasive infections and fungaemia have a high rate of mortality. A majority of these infections are caused by *Candida* species, and the incidence of invasive candidiasis has been rising worldwide in the last decades (Guinea 2014). Denmark is not an exception to this, and although the incidence rate has stabilized at around 8/100,000 inhabitants, it remains higher than in other Nordic countries (Astvad et al. 2017). Moreover, between 2012 and 2015 an increasing trend could be observed in the level of antimicrobial resistance towards azoles and echinocandins, making treatment of candidiasis difficult. ("DANMAP 2018 - Use of Antimicrobial Agents and Occurrence of Antimicrobial Resistance in Bacteria from Food Animals, Food and Humans in Denmark.," n.d.; Astvad et al. 2017).

Candida species normally colonize the skin, intestines or mucosal surfaces, but may cause disseminated infections in immuno-compromised individuals (Verduyn Lunel, Meis, and Voss 1999). The most common species causing candidemia is *Candida albicans*, but the species composition shows a changing pattern with *Candida glabrata* and *Candida parapsilosis* infections becoming more prevalent (Astvad et al. 2017; Guinea 2014; Chapman et al. 2017; Pfaller et al. 2019).

Currently, the main antifungal agents used to treat infections caused by *Candida* species are the azoles (such as ketoconazole, fluconazole and itraconazole, as well as more recent compounds as voriconazole and posaconazole) (Allen et al. 2015), echinocandins (in particular caspofungin, micafungin and anidulafungin) (Mroczyńska and Brillowska-Dąbrowska 2020) and amphotericin B (Lemke, Kiderlen, and Kayser 2005). Several *Candida* species, and notably *C. albicans* and *C. glabrata* are able to acquire resistance to these drugs, especially through mutations in genes coding for the drugs' targets or enzymes involved in their syntheses, or through the over-expression of efflux pumps. Point-mutations in *FKS* genes that encode the β-1,3-glucan synthase, which is the target for echinocandins, decrease susceptibility towards these antifungals (Alexander et al. 2013; Katiyar, Pfaller, and Edlind 2006; Balashov, Park, and Perlin 2006). Several mutations have also been identified in genes involved in lanosterol 14α-demethylase synthesis leading to resistance to azoles (Nishimoto, Sharma, and Rogers 2020; Whaley et al. 2016; Sanguinetti et al. 2005; Yao et al. 2019). Examples of such mutations occur in the *ERG11* gene, leading to reduced binding of the drugs to their target, or over-expression of the gene due to mutations in promoters or transcription factors. Over-expression of efflux pumps is another known common mechanism that leads to azole resistance, mediated by mutations in transcription factors – such as *TAC1* - that lead to the over-expression of ATP-binding cassette transporters such as the one coded by *CDR1* and *CDR2* genes.

C. albicans is a predominantly diploid organism that has a highly clonal population structure. Multilocus sequence typing (MLST)has shown that *C. albicans* samples belong to a limited number of major clades that show some degree of geographical enrichment (Tavanti et al. 2005; Odds et al. 2007; Shin et al. 2011). MLST has the limitation of relying on only seven gene fragments, situated on five chromosomes, which puts constrains on detecting possible recombination events stemming from the parasexual cycle of *C. albicans* (Heitman et al. 2014; Ropars et al. 2018).

C. glabrata, on the other hand, is thought to be obligate haploid and sexual reproduction has not been observed in this species, although genetic analysis indicates the capability for it (Gabaldón and Fairhead 2019; Carreté et al. 2018). Six locus-based MLST studies have demonstrated clonal population and clades that are, similarly to *C. albicans*, enriched in certain geographical areas (Lott et al. 2012; Hou et al. 2017; S et al. 2018).

Recently, whole-genome sequencing has become available for routine studies with fungal pathogens. It has the advantage over MLST to give increased resolution and information about structural variants, that could offer new insights into the mechanisms of virulence. So far, only a limited number of studies have investigated the phylogeny of *Candida* species and have demonstrated that both organisms are capable of recombination and have shown lack of geographical signal in the phylogenies (Biswas et al. 2018; Carreté et al. 2018; Sitterlé et al. 2019; Ropars et al. 2018).

The present study was conducted to obtain a snapshot of the diversity of *Candida* species across all 11 Danish Clinical Microbiology Laboratories (DCM). In addition, the genomes were compared to publicly available genomes from other geographic locations. We show that Danish isolates fit into several global clades both for *C. albicans* and *C. glabrata*.

Materials and Methods

Clinical isolates

On January 10, 2018, bacterial and yeast isolates from routine analyses (n=2,073) were collected from all 11 DCM (Rebelo et al. Submitted), 51 yeast isolates were identified among this collection. Geographical provenience of samples was anonymized by attributing a code (F1 to F11) to each of the 11 DCM. Metadata for the isolates included in the study are in Supplementary Data 1. The project was approved by local Danish Data Protection Agencies in each Region and MTAs were entered between the University and all DCM.

Antifungal susceptibility testing

Antifungal susceptibility testing was performed by broth microdilution using Sensititre™ YeastOne™ YO10 panel (Thermo Fisher Scientific Inc., Waltham, MA, USA) according to manufacturers' instructions. The control strain *Candida parapsilosis* ATCC 22019 was used. Minimum Inhibitory Concentration (MIC) interpretation was performed according to the EUCAST clinical breakpoints (EUCAST Antifungal Clinical Breakpoint Table v. 10.0) or, where clinical breakpoints were not available, EUCAST epidemiological cut-off values (ECOFF).

Whole genome sequencing data and species identification

Genomic DNA was extracted from the 51 yeast isolates using the Easy-DNA™ Kit (Invitrogen, Carlsbad, CA, USA) and DNA concentrations were determined using the Qubit™ dsDNA high-sensitivity (HS) and/or broad-range (BR) assay kits (Invitrogen, Carlsbad, CA, USA). The genomic DNA was prepared for Illumina pair-end sequencing using the Illumina (Illumina, Inc., San Diego, CA, USA) NexteraXT® DNA Library Prep Reference Guide (Document #15031942, v03, February 2018) and NextSeq System Denature and Dilute Libraries Guide (Document #15048776, v03, April 2018). The libraries were sequenced using the Illumina NextSeq 500 platform.

Quality control was performed on the sequencing reads with FastQC v.0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and summarized with MultiQC (Ewels et al. 2016). The raw reads were trimmed with BBDuk2 from BBMaps v36.49 (Bushnell B. n.d.), and assembled with SPAdes v3.11.0 (Bankevich et al. 2012) into draft genomes. *In silico* species identification was performed using KmerFinder v3.1 (Hasman et al. 2014; Larsen et al. 2014; Clausen, Aarestrup, and Lund 2018) on the draft genomes. Only the *C. albicans* and *C. glabrata* isolates were included in this analysis.

Variant calling and phylogenetic analysis

The pipeline for variant calling was constructed using Snakemake (Köster and Rahmann 2012). Trimmed reads were aligned against the reference genome of *Candida albicans* SC5314 or *Candida glabrata* CBS138 with minimap2 v2.6 (Li 2018). BAM files were sorted by coordinates with SAMtools v1.6 (Li et al. 2009), then duplicated reads were removed with sambamba v0.6.5 (Tarasov et al. 2015). Genome coverage was calculated for each sample, and samples with less than 20x over 80% of the total reference length were excluded from the analysis due to low coverage.

Raw variants were called with the Genome Analysis Toolkit (gatk) v.4.1.3.0 (McKenna et al. 2010) HaplotypeCaller with a call confidence threshold of 10.0. Single polymorphic nucleotides (SNPs) and insertions-deletions were separated by gatk SelecVariants for the purpose of hard filtering with VariantFiltration. SNPs were selected for, if they cleared the following filters: QD < 2.0, FS > 60.0, SOR > 3.0, MQ < 40.0, MQRankSum < -10.0, while indels were filtered with the following: QD < 2.0, FS > 200.0, SOR > 10.0, ReadPosRankSum < -20.0. Afterwards, the variants that passed were merged into one VCF per sample. Consensus sequences were generated by a custom Python script, that encoded heterozygous positions as IUPAC ambiguous bases, and deletions and filtered variants as 'N's.

The genomic data were partitioned along chromosomes. Phylogenetic trees were inferred with IQ-TREE v1.6.12 (Nguyen et al. 2015) using its model-test for predicting the best fitting nucleotide substitution models for each partition, and predicting support with 1000 ultrafast-bootstrap (UFBoot) and approximate likelihood-ratio test (SH-aLRT). The phylogenetic trees were visualized in R (R Core Team 2018) using ggtree (Yu et al. 2017), phangorn (Schliep 2011) and ape (Paradis and Schliep 2018) packages.

The genomes of 182 *C. albicans* isolates (PRJNA432884) and 96 *C. glabrata* (SRR2982714-SRR2982719, SRR5239753-SRR5239784, SRR5459192-SRR5459193, SRR7609359-SRR7609360, SRR8068012-SRR8068062, SRR8241569-SRR8241571) isolates were downloaded from the European Nucleotide Archives, and processed in the same manner as described above to create a global phylogeny.

The mixed ST clades on the *C. glabrata* global phylogenetic tree were analyzed with the Mann-Whitney U test for difference between the inter-ST distances and the intra-ST distances.

Multi locus sequence typing

In silico MLST was performed by using the Center for Genomic Epidemiology MLST v2.0.4 (Larsen et al. 2012) on the *C. glabrata* consensus samples as input. The genotypes and ST definitions (Dodgson et al. 2003) were downloaded from the *Candida glabrata* MLST website on pubmlst.org (Jolley, Bray, and Maiden 2018) on 2020-04-30.

Code availability

Scripts are publicly available at https://bitbucket.org/jszarvas/candida_phylogeny_pipeline/

Data availability

Raw sequence data have been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under study accession no. PRJEB40738.

Online Supplementary

Supplementary Data 1: https://files.dtu.dk/u/yXeJhVI3iZvbd4mH/Supplementary_Data_1-Sample_metadata.xlsx?l

Supplementary Data 2: https://files.dtu.dk/u/_8N6dm_Demntw3p0/Supplementary_Data_2-MIC_and_breakpoints_used.xlsx?l

Results

Of the 51 yeast isolates, 30 were found to be *Candida albicans*, eight *Candida glabrata*, seven *Candida krusei*, three *Candida dubliniensis*, two *Candida tropicalis*, and one *Candida guilliermondii*. Only the *C. albicans* and *C. glabrata* genomes were analyzed further.

Antifungal susceptibility testing showed high susceptibility towards azoles among the *C. albicans* isolates: all were clinically susceptible or presented a wild-type phenotype to fluconazole, itraconazole, voriconazole, and posaconazole (Figure 1). The *C. glabrata* isolates had lowered susceptibility to azoles (Figure 2). Two were resistant to fluconazole and simultaneously presented a non-wild type phenotype for posaconazole, with one of those isolates also presenting a non-wild type phenotype for voriconazole. All eight *C. glabrata* isolates had lower MIC values than the ECOFF for itraconazole. Furthermore, all isolates from both species were susceptible to amphotericin B. Seven *C. albicans* isolates showed increased MIC to echinocandins when compared with the respective EUCAST clinical breakpoints: all seven appeared resistant to anidulafungin, and one also resistant to micafungin. All the *C. glabrata* isolates were susceptible to echinocandins. The MIC values and breakpoints used for each antifungal agent can be found in Supplementary Data 2.

Phylogeny of the Danish *C. albicans* isolates identified three clusters with more than five isolates each, and the remaining isolates as clusters of two or singletons. The *C. glabrata* phylogeny was comprised of two clusters with three isolates each, and two singletons. All larger clusters encompassed isolates from different DCM and sources (Figure S1-Figure S2). Site substitution rates on branches did not indicate that any isolate was closely related to another in either phylogeny.

Multi-locus sequence typing of *C. glabrata* isolates (Table 1) predicted five STs, with ST3 for three and ST6 for two isolates.

Table 1: In silico MLST of C. glabrata isolates identified five STs

Sample	**ST**
2018-F1-188-B	148
2018-F5-18	3
2018-F6-160	7
2018-F6-162	6
2018-F6-164	3
2018-F6-165	3
2018-F6-167	6
2018-F9-42	128

In the *C. albicans* global phylogeny (Figure 3), the Danish isolates fit into already established MLST-based clades (Odds et al. 2007; Shin et al. 2011; Ropars et al. 2018): seven isolates were in clade 2, six each in clades 4 and 1, two in clade 12, and one each in clades 13 and 3. One isolate was an out-group to clade 12. Other isolates were clustered in novel clusters suggested by Ropars et al, namely in cluster A, C, E and F. The *C. glabrata* isolates separate along the predicted STs (Figure 4), except for 2018-F9-42, which is placed on a clade with the ST6 samples. There are three instances of clades containing isolates from two STs: ST6 and ST128, ST19 and ST65, and ST2 and ST8. The intra-ST distances do not differ significantly from the inter-ST distances for either of these clades, when tested with the Mann-Whitney U test (p values 0.09073, 0.4228 and 0.4081, respectively).

Discussion

In this study we sequenced clinical *Candida* isolates collected during one day in Denmark, from which the majority of the isolates were typed to be *C. albicans* (n=30), followed by *C. glabrata* (n=8). These results are concordant with the findings of the incidence-based fungemia surveillance from Denmark (Astvad et al. 2017) that reported increased prevalence of *C. glabrata* candidemia in the country, although only a limited number of our isolates were isolated from blood. However, studies have shown (Lott et al. 2012) that invasive candidiasis is frequently caused by commensal strains, therefore less severe opportunistic infections could mirror the prevalence of candidemia.

Antifungal susceptibility testing showed, that all *C. albicans* isolates were susceptible to azoles. Approximately 25% (7/30 isolates) appeared to be resistant to anidulafungin, and one also to be resistant to micafungin, albeit with MIC values only one or two-fold dilutions above the EUCAST breakpoint. This would be a concerning development, as in the 2012-2015 period no anidulafungin-resistant *C. albicans* isolates have been detected in Denmark and considering that echinocandin-resistant *C. albicans* remain rare worldwide (Pfaller et al. 2011; Dannaoui et al. 2012; Mroczyńska and Brillowska-Dąbrowska 2020), with few cases reported and mostly corresponding to a decrease in susceptibility following long-term treatments with these antifungal agents (Fekkar et al. 2014; Shields, Nguyen, and Clancy 2015; Ruggero and Topal 2014; Lortholary et al. 2011; Laverdière et al. 2006; Miller et al. 2006). However, we propose that insufficient standardization of the AST method used is responsible for obtaining the observed higher MIC values, as demonstrated by the classification of these isolates as clinically susceptible to echinocandins when the Clinical and Laboratory Standards Institute (CLSI) breakpoints are applied instead (S≤2 mg/L) (Personal communication). The *C. glabrata* isolates were all susceptible to echinocandins by both interpretations, in contrast to the global observations showing higher prevalence of echinocandin-resistance in *C. glabrata* than *C. albicans* (Grossman, Chiller, and Lockhart 2014; Pristov and Ghannoum 2019). However, the low number of *C. glabrata* isolates tested might not accurately reflect the effective *Candida* epidemiology in the country. Of the *C. glabrata* isolates, 25% (2/8 isolates) presented a non-wild type or resistant phenotype for certain azoles. This is in accordance with the results of the national surveillance reporting increased number of non-wild type *C. glabrata* samples. (Astvad et al. 2017)

The resistant or non-wild type isolates, with the exception of two pairs of isolates, did not cluster phylogenetically (Figure 1-Figure 2), and we could assume that the genetic mechanisms conferring resistance were acquired independently. While isolates 2018-F1-53 and 2018-F1-187 were clustered together and both exhibited increased MIC values to anidulafungin, it has been reported that *C. albicans* acquires resistance-conferring mutations rapidly in the gene coding for the drug target (Spettel et al. 2019). Similarly, 2018-F1-162 and 2018-F1-167 clustered together and presented a non-wild phenotype for azoles, but *C. glabrata* is also known for developing resistance during treatment (Spettel et al. 2019; Garnaud et al. 2015). In both cases, further analysis would be required to investigate the genetic background of the observed phenotypes.

The majority of the *C. albicans* samples (19/30 isolates) were from the most sampled global clades, namely clade 2, 1 and 4 (Figure 3). The within-clade diversity was considerable in these clades, although the inter-clade distance was higher. The Danish isolates were scattered in clade 1, and only 2018-F1-54 and 2018-F1-192 were clustered together, separating from a cluster of French isolates. In clades 2 and 4 the Danish isolates are separated into two clusters each, and the larger clusters contain only Danish isolates. Although the population is under-sampled, geographically constrained clonal expansion could explain the observed phenomenon. The remaining 11 isolates were placed into three MLST-established and four putative clades, and one remained a singleton. One of these isolates was embedded in clade 13, that contained strains

adapted to the genital niche, with a very low degree of polymorphism and heterozygosity, characteristics that 2018-F1-190 also shared.

The *C. glabrata* isolates were largely separated along ST types in the global phylogeny (Figure 4), and the intra-cluster diversity was lower than what could be observed in the case of *C. albicans*. *C. glabrata* is haploid, therefore heterozygosity does not contribute to the genetic diversity of the species. Approximately 63% (n=5) of the samples belonged to the globally prevalent ST3 and ST6. There were three instances of STs intermixing within clades, genetic drift affecting one of the six loci in the MLST scheme resulting in a different ST designation, despite the inter-ST genetic distances being comparable to the intra-ST genetic distances.

In summary, we whole-genome sequenced Danish *Candida* isolates to uncover the population structure. We found that the majority of the isolates belonged to clades that were widespread in the world and despite some evidence of local clonal expansion. This study suggests that both *C. albicans* and *C. glabrata* are truly globally distributed opportunistic pathogens without geographically restrictions on diversity.

References

Alexander, Barbara D., Melissa D. Johnson, Christopher D. Pfeiffer, Cristina Jiménez-Ortigosa, Jelena Catania, Rachel Booker, Mariana Castanheira, Shawn A. Messer, David S. Perlin, and Michael A. Pfaller. 2013. "Increasing Echinocandin Resistance in Candida Glabrata: Clinical Failure Correlates with Presence of FKS Mutations and Elevated Minimum Inhibitory Concentrations." *Clinical Infectious Diseases* 56 (12): 1724–32. https://doi.org/10.1093/cid/cit136.

Allen, David, Dustin Wilson, Richard Drew, and John Perfect. 2015. "Azole Antifungals: 35 Years of Invasive Fungal Infection Management." *Expert Review of Anti-Infective Therapy* 13 (6): 787–98. https://doi.org/10.1586/14787210.2015.1032939.

Astvad, K. M. T., H. K. Johansen, B. L. Røder, F. S. Rosenvinge, J. D. Knudsen, L. Lemming, H. C. Schønheyder, et al. 2017. "Update from a 12-Year Nationwide Fungemia Surveillance: Increasing Intrinsic and Acquired Resistance Causes Concern." Edited by Daniel J. Diekema. *Journal of Clinical Microbiology* 56 (4). https://doi.org/10.1128/JCM.01564-17.

Balashov, Sergey V., Steven Park, and David S. Perlin. 2006. "Assessing Resistance to the Echinocandin Antifungal Drug Caspofungin in Candida Albicans by Profiling Mutations in FKS1." *Antimicrobial Agents and Chemotherapy* 50 (6): 2058–63. https://doi.org/10.1128/AAC.01653-05.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, et al. 2012. "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing." *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Biswas, Chayanika, Vanessa R. Marcelino, Sebastiaan Van Hal, Catriona Halliday, Elena Martinez, Qinning Wang, Sarah Kidd, et al. 2018. "Whole Genome Sequencing of Australian Candida Glabrata Isolates Reveals Genetic Diversity and Novel Sequence Types." *Frontiers in Microbiology* 9 (March). https://doi.org/10.3389/fmicb.2018.02946.

Bushnell B. n.d. "BBMap." Accessed February 16, 2020. https://sourceforge.net/projects/bbmap/.

Carreté, Laia, Ewa Ksiezopolska, Cinta Pegueroles, Emilia Gómez-Molero, Ester Saus, Susana Iraola-Guzmán, Damian Loska, Oliver Bader, Cecile Fairhead, and Toni Gabaldón. 2018. "Patterns of Genomic Variation in the Opportunistic Pathogen Candida Glabrata Suggest the Existence of

Mating and a Secondary Association with Humans." *Current Biology* 28 (1): 15–27.e7. https://doi.org/10.1016/j.cub.2017.11.027.
Chapman, Belinda, Monica Slavin, Debbie Marriott, Catriona Halliday, Sarah Kidd, Ian Arthur, Narin Bak, et al. 2017. "Changing Epidemiology of Candidaemia in Australia." *Journal of Antimicrobial Chemotherapy* 72 (4): 1103–1108. https://doi.org/10.1093/jac/dkw422.
Clausen, Philip T.L.C., Frank M. Aarestrup, and Ole Lund. 2018. "Rapid and Precise Alignment of Raw Reads against Redundant Databases with KMA." *BMC Bioinformatics* 19 (1): 307. https://doi.org/10.1186/s12859-018-2336-6.
"DANMAP 2018 - Use of Antimicrobial Agents and Occurrence of Antimicrobial Resistance in Bacteria from Food Animals, Food and Humans in Denmark." n.d. In .
Dannaoui, Eric, Marie Desnos-Ollivier, Dea Garcia-Hermoso, Frédéric Grenouillet, Sophie Cassaing, Marie Thérèse Baixench, Stéphane Bretagne, Françoise Dromer, and Olivier Lortholary. 2012. "Candida Spp. with Acquired Echinocandin Resistance, France, 2004-2010." *Emerging Infectious Diseases* 18 (1): 86–90. https://doi.org/10.3201/eid1801.110556.
Dodgson, Andrew R., Claude Pujol, David W. Denning, David R. Soll, and Andrew J. Fox. 2003. "Multilocus Sequence Typing of Candida Glabrata Reveals Geographically Enriched Clades." *Journal of Clinical Microbiology* 41 (12): 5709–17. https://doi.org/10.1128/JCM.41.12.5709-5717.2003.
Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. "MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report." *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.
Fekkar, A., E. Dannaoui, I. Meyer, S. Imbert, J. Y. Brossas, M. Uzunov, G. Mellon, et al. 2014. "Emergence of Echinocandin-Resistant Candida Spp. in a Hospital Setting: A Consequence of 10 Years of Increasing Use of Antifungal Therapy?" *European Journal of Clinical Microbiology and Infectious Diseases* 33 (9): 1489–96. https://doi.org/10.1007/s10096-014-2096-9.
Gabaldón, Toni, and Cécile Fairhead. 2019. "Genomes Shed Light on the Secret Life of Candida Glabrata: Not so Asexual, Not so Commensal." *Current Genetics* 65 (1): 93–98. https://doi.org/10.1007/s00294-018-0867-z.
Garnaud, Cécile, Françoise Botterel, Natacha Sertour, Marie-Elisabeth Bougnoux, Eric Dannaoui, Sylvie Larrat, Christophe Hennequin, Jesus Guinea, Muriel Cornet, and Danièle Maubon. 2015. "Next-Generation Sequencing Offers New Insights into the Resistance of Candida Spp. to Echinocandins and Azoles." *Journal of Antimicrobial Chemotherapy* 70 (9): 2556–65. https://doi.org/10.1093/jac/dkv139.
Grossman, Nina T, Tom M Chiller, and Shawn R Lockhart. 2014. "Epidemiology of Echinocandin Resistance in Candida." *Current Fungal Infection Reports* 8 (4): 243–48. https://doi.org/10.1007/s12281-014-0209-7.
Guinea, J. 2014. "Global Trends in the Distribution of Candida Species Causing Candidemia." *Clinical Microbiology and Infection*. Elsevier B.V. https://doi.org/10.1111/1469-0691.12539.
Hasman, Henrik, Dhany Saputra, Thomas Sicheritz-Ponten, Ole Lund, Christina Aaby Svendsen, Niels Frimodt-Moller, and Frank M. Aarestrup. 2014. "Rapid Whole-Genome Sequencing for Detection and Characterization of Microorganisms Directly from Clinical Samples." *Journal of Clinical Microbiology* 52 (1): 139–46. https://doi.org/10.1128/JCM.02452-13.
Heitman, Joseph, Dee A. Carter, Paul S. Dyer, and David R. Soll. 2014. "Sexual Reproduction of Human Fungal Pathogens." *Cold Spring Harbor Perspectives in Medicine* 4 (8). https://doi.org/10.1101/cshperspect.a019281.
Hou, Xin, Meng Xiao, Sharon C.A. Chen, Fanrong Kong, He Wang, Yun Zhuo Chu, Mei Kang, et al. 2017. "Molecular Epidemiology and Antifungal Susceptibility of Candida Glabrata in China (August 2009 to July 2014): A Multi-Center Study." *Frontiers in Microbiology* 8 (MAY).

https://doi.org/10.3389/fmicb.2017.00880.
Jolley, Keith A., James E. Bray, and Martin C.J. Maiden. 2018. “Open-Access Bacterial Population Genomics: BIGSdb Software, the PubMLST.Org Website and Their Applications [Version 1; Referees: 2 Approved].” *Wellcome Open Research* 3. https://doi.org/10.12688/wellcomeopenres.14826.1.
Katiyar, Santosh, Michael Pfaller, and Thomas Edlind. 2006. “Candida Albicans and Candida Glabrata Clinical Isolates Exhibiting Reduced Echinocandin Susceptibility.” *Antimicrobial Agents and Chemotherapy* 50 (8): 2892–94. https://doi.org/10.1128/AAC.00349-06.
Köster, Johannes, and Sven Rahmann. 2012. “Snakemake--a Scalable Bioinformatics Workflow Engine.” *Bioinformatics (Oxford, England)* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.
Larsen, M. V., S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Ponten, F. M. Aarestrup, D. W. Ussery, and O. Lund. 2014. “Benchmarking of Methods for Genomic Taxonomy.” *Journal of Clinical Microbiology* 52 (5): 1529–39. https://doi.org/10.1128/JCM.02981-13.
Larsen, M. V., S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R. L. Marvig, L. Jelsbak, et al. 2012. “Multilocus Sequence Typing of Total-Genome-Sequenced Bacteria.” *Journal of Clinical Microbiology* 50 (4): 1355–61. https://doi.org/10.1128/JCM.06094-11.
Laverdière, Michel, Richard G. Lalonde, Jean Guy Baril, Donald C. Sheppard, Steven Park, and David S. Perlin. 2006. “Progressive Loss of Echinocandin Activity Following Prolonged Use for Treatment of Candida Albicans Oesophagitis.” *Journal of Antimicrobial Chemotherapy* 57 (4): 705–8. https://doi.org/10.1093/jac/dkl022.
Lemke, A., A. F. Kiderlen, and O. Kayser. 2005. “Amphotericin B.” *Applied Microbiology and Biotechnology* 68 (2): 151–62. https://doi.org/10.1007/s00253-005-1955-9.
Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” Edited by Inanc Birol. *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078. https://doi.org/10.1093/BIOINFORMATICS/BTP352.
Lortholary, Olivier, Marie Desnos-Ollivier, Karine Sitbon, Arnaud Fontanet, Stéphane Bretagne, Françoise Dromer, C. Bouges-Michel, et al. 2011. “Recent Exposure to Caspofungin or Fluconazole Influences the Epidemiology of Candidemia: A Prospective Multicenter Study Involving 2,441 Patients.” *Antimicrobial Agents and Chemotherapy* 55 (2): 532–38. https://doi.org/10.1128/AAC.01128-10.
Lott, Timothy J., Joao P. Frade, G. Marshall Lyon, Naureen Iqbal, and Shawn R. Lockhart. 2012. “Bloodstream and Non-Invasive Isolates of Candida Glabrata Have Similar Population Structures and Fluconazole Susceptibilities.” *Medical Mycology* 50 (2): 136–42. https://doi.org/10.3109/13693786.2011.592153.
McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.
Miller, Christopher D., Ben W. Lomaestro, Steven Park, and David S. Perlin. 2006. “Progressive Esophagitis Caused by Candida Albicans with Reduced Susceptibility to Caspofungin.” *Pharmacotherapy* 26 (6 I): 877–80. https://doi.org/10.1592/phco.26.6.877.
Mroczyńska, Martyna, and Anna Brillowska-Dąbrowska. 2020. “Review on Current Status of Echinocandins Use.” *Antibiotics* 9 (5): 227. https://doi.org/10.3390/antibiotics9050227.

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74. https://doi.org/10.1093/molbev/msu300.

Nishimoto, Andrew T., Cheshta Sharma, and P. David Rogers. 2020. “Molecular and Genetic Basis of Azole Antifungal Resistance in the Opportunistic Pathogenic Fungus Candida Albicans.” *Journal of Antimicrobial Chemotherapy* 75 (2): 257–70. https://doi.org/10.1093/jac/dkz400.

Odds, Frank C., Marie-Elisabeth Bougnoux, Duncan J. Shaw, Judith M. Bain, Amanda D. Davidson, Dorothée Diogo, Mette D. Jacobsen, et al. 2007. “Molecular Phylogenetics of Candida Albicans.” *Eukaryotic Cell* 6 (6): 1041–52. https://doi.org/10.1128/EC.00041-07.

Paradis, Emmanuel, and Klaus Schliep. 2018. “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R.” *Bioinformatics* 35 (3): 526–28. https://doi.org/10.1093/bioinformatics/bty633.

Pfaller, Michael A., Shawn A. Messer, Gary J. Moet, Ronald N. Jones, and Mariana Castanheira. 2011. “Candida Bloodstream Infections: Comparison of Species Distribution and Resistance to Echinocandin and Azole Antifungal Agents in Intensive Care Unit (ICU) and Non-ICU Settings in the SENTRY Antimicrobial Surveillance Program (2008-2009).” *International Journal of Antimicrobial Agents* 38 (1): 65–69. https://doi.org/10.1016/j.ijantimicag.2011.02.016.

Pfaller, Michael A, Daniel J Diekema, John D Turnidge, Mariana Castanheira, and Ronald N Jones. 2019. “Twenty Years of the SENTRY Antifungal Surveillance Program: Results for Candida Species From 1997–2016.” *Open Forum Infectious Diseases* 6 (Supplement_1): S79–94. https://doi.org/10.1093/ofid/ofy358.

Pristov, K. E., and M. A. Ghannoum. 2019. “Resistance of Candida to Azoles and Echinocandins Worldwide.” *Clinical Microbiology and Infection*. Elsevier B.V. https://doi.org/10.1016/j.cmi.2019.03.028.

R Core Team. 2018. “No Title.” Vienna, Austria: R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Ropars, Jeanne, Corinne Maufrais, Dorothée Diogo, Marina Marcet-Houben, Aurélie Perin, Natacha Sertour, Kevin Mosca, et al. 2018. “Gene Flow Contributes to Diversification of the Major Fungal Pathogen Candida Albicans.” *Nature Communications* 9 (1): 2253. https://doi.org/10.1038/s41467-018-04787-4.

Ruggero, M.A., and J.E. Topal. 2014. “Development of Echinocandin-Resistant *Candida Albicans* Candidemia Following Brief Prophylactic Exposure to Micafungin Therapy.” *Transplant Infectious Disease* 16 (3): 469–72. https://doi.org/10.1111/tid.12230.

S, Amanloo, Shams-Ghahfarokhi M, Ghahri M, and Razzaghi-Abyaneh M. 2018. “Genotyping of Clinical Isolates of Candida Glabrata From Iran by Multilocus Sequence Typing and Determination of Population Structure and Drug Resistance Profile.” *Medical Mycology* 56 (2). https://doi.org/10.1093/MMY/MYX030.

Sanguinetti, Maurizio, Brunella Posteraro, Barbara Fiori, Stefania Ranno, Riccardo Torelli, and Giovanni Fadda. 2005. “Mechanisms of Azole Resistance in Clinical Isolates of Candida Glabrata Collected during a Hospital Survey of Antifungal Resistance.” *Antimicrobial Agents and Chemotherapy* 49 (2): 668–79. https://doi.org/10.1128/AAC.49.2.668-679.2005.

Schliep, Klaus Peter. 2011. “Phangorn: Phylogenetic Analysis in R.” *Bioinformatics* 27 (4): 592–93. https://doi.org/10.1093/bioinformatics/btq706.

Shields, Ryan K., M. H. Nguyen, and Cornelius J. Clancy. 2015. “Clinical Perspectives on Echinocandin Resistance among Candida Species.” *Current Opinion in Infectious Diseases* 28 (6): 514–22. https://doi.org/10.1097/QCO.0000000000000215.

Shin, Jong Hee, Marie Elisabeth Bougnoux, Christophe D’Enfert, Soo Hyun Kim, Chang Jin Moon, Min Young Joo, Kyungwon Lee, et al. 2011. “Genetic Diversity among Korean Candida

Albicans Bloodstream Isolates: Assessment by Multilocus Sequence Typing and Restriction Endonuclease Analysis of Genomic DNA by Use of BssHII." *Journal of Clinical Microbiology* 49 (7): 2572–77. https://doi.org/10.1128/JCM.02153-10.

Sitterlé, Emilie, Corinne Maufrais, Natacha Sertour, Matthieu Palayret, Christophe d'Enfert, and Marie Elisabeth Bougnoux. 2019. "Within-Host Genomic Diversity of Candida Albicans in Healthy Carriers." *Scientific Reports* 9 (1): 1–12. https://doi.org/10.1038/s41598-019-38768-4.

Spettel, Kathrin, Wolfgang Barousch, Athanasios Makristathis, Iris Zeller, Marion Nehr, Brigitte Selitsch, Michaela Lackner, Peter-Michael Rath, Joerg Steinmann, and Birgit Willinger. 2019. "Analysis of Antifungal Resistance Genes in Candida Albicans and Candida Glabrata Using next Generation Sequencing." Edited by Joy Sturtevant. *PLOS ONE* 14 (1): e0210397. https://doi.org/10.1371/journal.pone.0210397.

Tarasov, Artem, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. 2015. "Sambamba: Fast Processing of NGS Alignment Formats." *Bioinformatics* 31 (12): 2032–34. https://doi.org/10.1093/bioinformatics/btv098.

Tavanti, Arianna, Amanda D. Davidson, Mark J. Fordyce, Neil A.R. Gow, Martin C.J. Maiden, and Frank C. Odds. 2005. "Population Structure and Properties of Candida Albicans, as Determined by Multilocus Sequence Typing." *Journal of Clinical Microbiology* 43 (11): 5601–13. https://doi.org/10.1128/JCM.43.11.5601-5613.2005.

Verduyn Lunel, Frans M., Jacques F.G.M. Meis, and Andreas Voss. 1999. "Nosocomial Fungal Infections: Candidemia." In *Diagnostic Microbiology and Infectious Disease*, 34:213–20. https://doi.org/10.1016/S0732-8893(99)00035-8.

Whaley, Sarah G, Elizabeth L Berkow, Jeffrey M Rybak, Andrew T Nishimoto, Katherine S Barker, and P David Rogers. 2016. "Azole Antifungal Resistance in Candida Albicans and Emerging Non-Albicans Candida Species." *Frontiers in Microbiology* 7 (JAN): 2173. https://doi.org/10.3389/fmicb.2016.02173.

Yao, Dongting, Jia Chen, Weiqin Chen, Zhen Li, and Xiaobo Hu. 2019. "Mechanisms of Azole Resistance in Clinical Isolates of Candida Glabrata from Two Hospitals in China." *Infection and Drug Resistance* 12 (April): 771–81. https://doi.org/10.2147/IDR.S202058.

Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan Yuk Lam. 2017. "Ggtree: An r Package for Visualization and Annotation of Phylogenetic Trees with Their Covariates and Other Associated Data." *Methods in Ecology and Evolution* 8 (1): 28–36. https://doi.org/10.1111/2041-210X.12628.

Figures

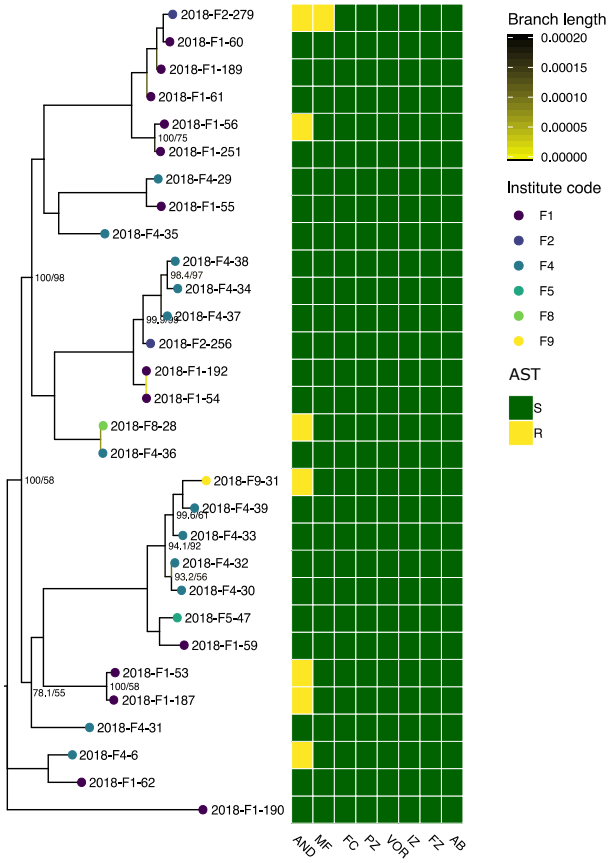

Figure 1: Maximum likelihood phylogenetic tree of *Candida albicans* isolates using the reference sequence *Candida albicans* SC5314, rooted on midpoint and displaying UFBoot/SH-aLRT support for splits with less than 100% support. Tips are colored according to the institution codes. The heatmap displays the antifungal susceptibility testing results against anidalufungin (AND), micafungin (MF), 5-Flucytosine (FC), posaconazole (PZ), voriconazole (VOR), itraconazole (IZ), fluconazole (FZ), amphotericin B (AB), R is MIC above and S is MIC below the EUCAST clinical breakpoints.

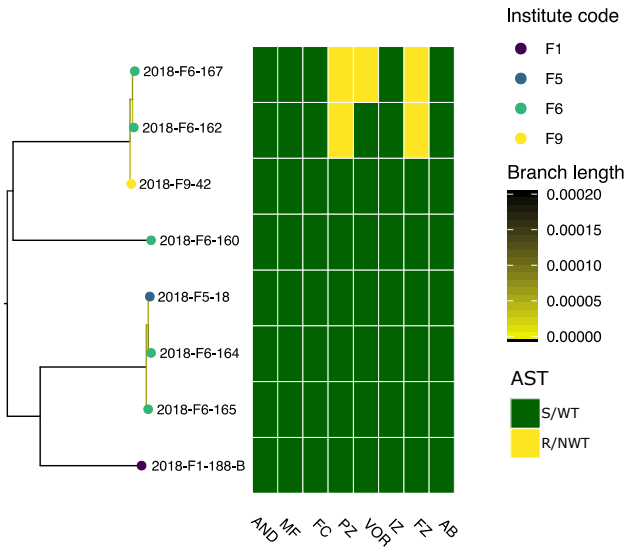

Figure 2: Maximum likelihood phylogenetic tree of *Candida glabrata* isolates using the reference sequence *Candida glabrata* CBS138, rooted on midpoint. Tips are colored according to the institution codes. The heatmap displays the antifungal susceptibility testing results against anidalufungin (AND), micafungin (MF), 5-Flucytosine (FC), posaconazole (PZ), voriconazole (VOR), itraconazole (IZ), fluconazole (FZ), amphotericin B (AB). R is MIC above and S is MIC below the EUCAST clinical breakpoints, WT is wild-type phenotype and NWT means non-wild-type phenotype according to the ECOFF.

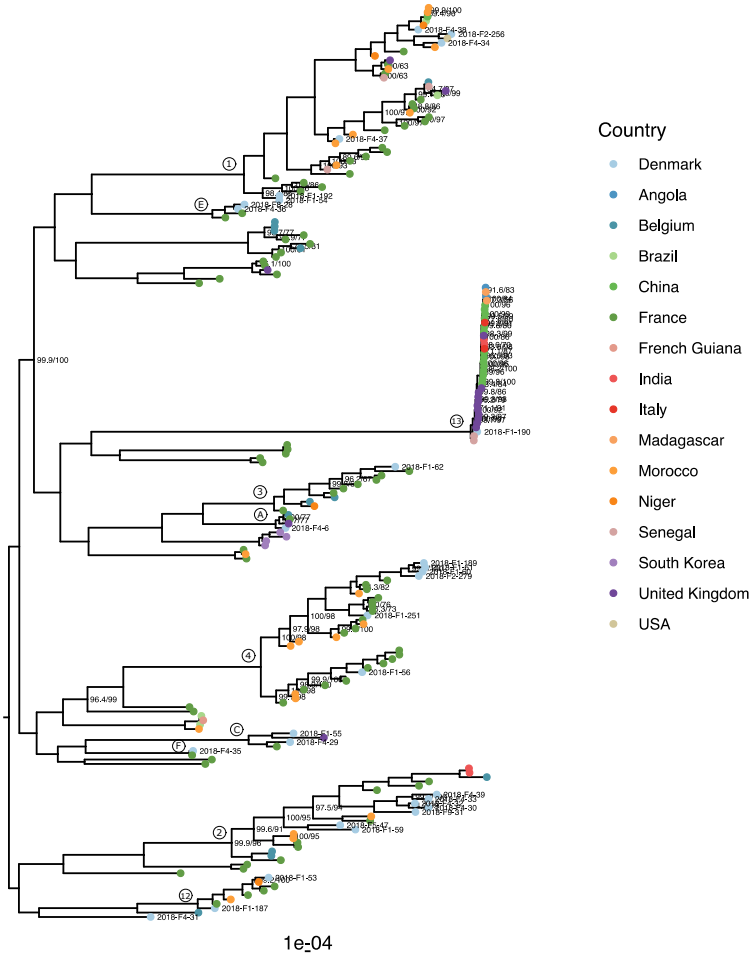

Figure 3: Global maximum likelihood phylogeny of *C. albicans* isolates, with the Danish isolates labelled. Tips are colored by country, the UFBoot/SH-aLRT support is shown where it is lower than 100%, the scale bar is 0.0001 substitutions per site. Clades labeled after Ropars et al.

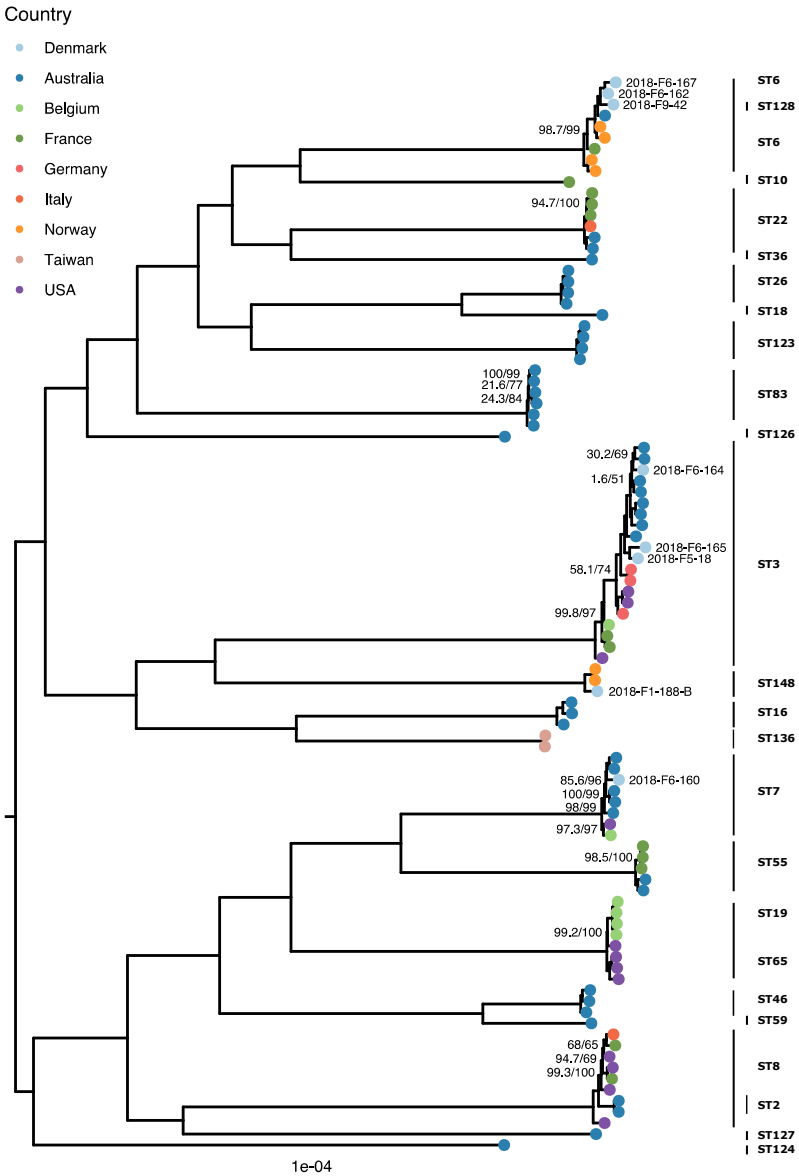

Figure 4: Global maximum likelihood phylogeny of *C. glabrata* isolates, with the Danish isolates labelled. Tips are colored by country, the UFBoot/SH-aLRT support is shown where it is lower than 100%, the scale bar is 0.0001 substitutions per site.

Supplementary material

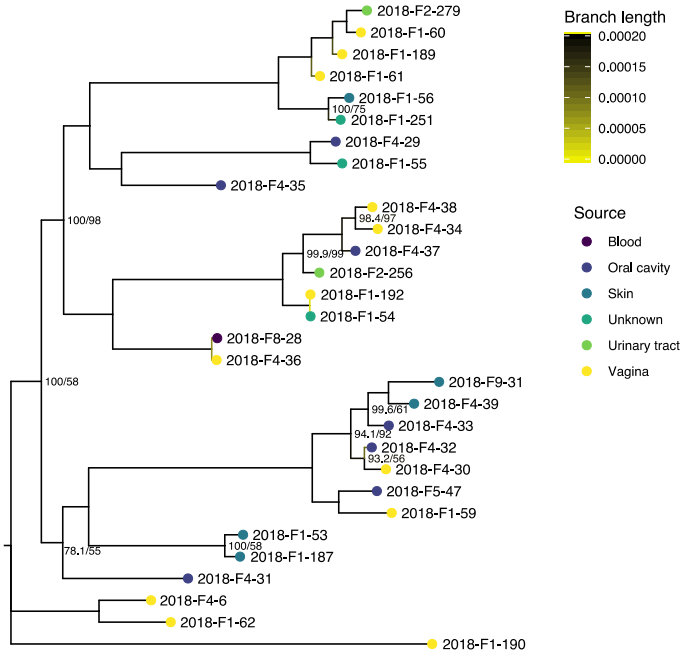

Figure S1: Maximum likelihood phylogenetic tree of *Candida albicans* isolates using the reference sequence *Candida albicans* SC5314, rooted on midpoint and displaying UFBoot/SH-aLRT support for splits with less than 100% support. Tips are colored according to the standardized sample sources.

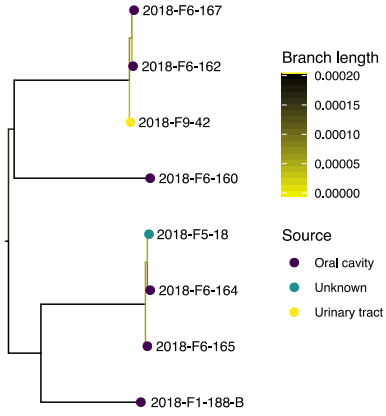

Figure S2: Maximum likelihood phylogenetic tree of *Candida glabrata* isolates using the reference sequence *Candida glabrata* CBS138, rooted on midpoint. Tips are colored according to the standardized sample sources

CHAPTER 3

Conclusion and future perspectives

The maturation of sequencing technologies brought whole-genome sequencing to non-academic settings, and the sequencing of pathogens is no longer limited to retrospective studies of outbreaks. Sequencers are now present in laboratories of food safety and public health authorities, and clinical microbiology departments of hospitals, where WGS is frequently performed as part of the routine workflow. Genomic epidemiology has incontrovertibly improved the effectiveness of outbreak investigations. However, there was a need for new methods for outbreak detection to fully utilize the potential of the new technology.

The project described in **Paper I** aimed to create a software solution for monitoring publicly available isolates of foodborne pathogens for potential outbreaks. The routine sequencing of clinical and environmental bacterial isolates means that the amount of data to be processed accumulates quickly, therefore the method needed to be scalable. The surveillance pipeline (PAPABAC) we developed is capable of accurately comparing the growing amount of bacterial isolates to each other with a computational time that does not increase exponentially with the number of samples. A surveillance platform for *Campylobacter jejuni*, *Escherichia coli*, *Shigella*, *Listeria monocytogenes* and *Salmonella enterica* was launched in October 2017, and it has since processed more than 200,000 isolates. A feature is planned that allows users the direct upload of raw read data from isolates and returns the individual results to them. Adhering to minimum metadata requirements would be a criterion for analysing the user-submitted data, to contribute to the growing collection of isolates.

In **Paper II**, we applied the stand-alone version of the pipeline in a hospital setting, for surveillance of methicillin-resistant *Staphylococcus aureus* and *Enterococcus faecium*. The study was retrospective, comparing output from the SNP-based

pipeline with cgMLST generated minimum-spanning trees, and the clusters detected by the two systems were highly concordant. The surveillance pipeline was updated to utilize more sophisticated filtering of unknown positions from the alignments, to have greater flexibility with regards to the genome content of the analysed organisms, while retaining high discriminatory power. This is crucial for the method to be generalizable to a wide range of bacterial species, and scalable to unconventionally large datasets. Additionally, the memory usage of the distance calculation was also optimized, in order to be able to run the program on processors with limited amount of RAM.

One of the limitations of the surveillance pipeline is the need for closely matching reference genomes. If a new subtype arises, that does not have a suitable reference in the reference database, it will not be processed. The database could, naturally, be updated manually, and the isolates in question could be re-analyzed, however, it would be more prudent to automate this step to some extent. Complete genomes are preferred over draft assemblies, but until a complete reference is available, it could suffice as a reference. Furthermore, the manner of homology reduction during the creation of the reference database could be improved. The current k-mer based method yields, on occasion, templates, that overlap in their recruited isolates, unnecessarily duplicating the data in the system. A gene-content aware method should reduce the overlap between templates.

In the work presented in **Paper III**, the ”evergreen” concept of the bacterial surveillance pipeline was adapted to consensus sequences. Large quantities of consensus sequences of viral origins are being deposited into Genbank and its equivalents by molecular surveillance projects, and our pipeline, called Krummholz, is capable of providing overview of this data in a matter of minutes. We applied the method to complete sequences of the norovirus major capsid protein, resulting in the data partitioned along the genogroups, with individual trees depicting the evolution of each group of sequences. Krummholz initially accepted only consensus sequences as input, but it can be expanded to work on raw sequencing data as well, taking the differences in sequencing platforms into account. The matter of suitable references discussed above continues to be relevant, however nature of the sequences of interest make it easier to circumvent the problem in this case. The initial reference database is recommended to be built from a multi-entry fasta file of prioritized reference sequences and all available sequences downloaded from Genbank via the supplied script, to provide suitable templates for all sequences. Thereafter, as surveillance over a long period is expected to gather sequences that evolve away from this initial reference set, the

inclusion of new references and repeated genotyping is just a matter of setting an additional flag. Another known limitation of alignment based methods is the exclusion of insertions and repeating regions from the analysis. The pipeline could be upgraded to take insertions (or deletions) into account by evaluating the alignment output, to determine whether they necessitate the inclusion of a new template.

Paper IV shifted focus to a third kingdom in the tree of life, and centred around clinical isolates of Candida species collected in Denmark. After genotyping, the *Candida albicans* and *Candida glabrata* isolates were included in a phylogenetic analysis together with other whole-genome sequenced Candida samples collected worldwide. The Danish isolates fit into globally established clades, and there is no evidence, that any of the isolates are closely related to each other. The analysis pipeline was written in Snakemake, a Python-based workflow manager, thus the work can be repeated with additional samples or applied to samples of a different fungal pathogen.

Nosocomial infections of antimicrobial resistant fungal pathogens are a growing concern, and the genomic epidemiology of these pathogens is less explored. Whole-genome sequencing of clinical Candida samples is in its early stages, with only a few studies conducted to sequence repeated samples or epidemiologically related samples. Therefore, further studies are needed to establish a threshold for genetic relatedness.

In summary, the work in this PhD thesis created computational methods for mass-surveillance of sequencing data, be that for outbreak detection of foodborne bacteria, monitoring of nosocomial infections, or surveying the genetic diversity of sequences of interest. Furthermore, it could contribute to the research into the population structure and evolution of pathogens.

Bibliography

[1] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 560–564, feb 1977.

[2] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, pp. 5463–5467, 1977.

[3] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, "Large-scale DNA sequencing," *Current Opinion in Biotechnology*, vol. 2, no. 1, pp. 92–101, 1991.

[4] G. N. Godson, B. G. Barrell, R. Staden, and J. C. Fiddes, "Nucleotide sequence of bacteriophage G4 DNA," *Nature*, vol. 276, no. 5685, pp. 236–247, 1978.

[5] A. Franck, H. Guilley, G. Jonard, K. Richards, and L. Hirth, "Nucleotide sequence of cauliflower mosaic virus DNA," *Cell*, vol. 21, no. 1, pp. 285–294, 1980.

[6] J. Craig Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Yuan Wang, A. Wang, X. Wang, J. Wang, M. H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. Lai Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch,

E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. Ni Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Deslattes Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The sequence of the human genome," *Science*, vol. 291, pp. 1304–1351, feb 2001.

[7] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, pp. 496–512, jul 1995.

[8] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao, "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, pp. 1453–1462, sep 1997.

[9] N. T. Perna, G. Plunkett, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Pósfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner, "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7," *Nature*, vol. 409, pp. 529–533, jan 2001.

[10] J. Parkhill, G. Dougan, K. D. James, N. R. Thomson, D. Pickard, J. Wain, C. Churcher, K. L. Mungall, S. D. Bentley, M. T. Holden, M. Sebaihia, S. Baker, D. Basham, K. Brooks, T. Chillingworth, P. Connerton, A. Cronin, P. Davis, R. M. Davies, L. Dowd, N. White, J. Farrar, T. Feltwell, N. Hamlin, A. Haque, T. T. Hien, S. Holroyd, K. Jagels, A. Krogh, T. S. Larsen, S. Leather, S. Moule, P. Ó'Gaora, C. Parry, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell, "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18," *Nature*, vol. 413, pp. 848–852, oct 2001.

[11] M. McClelland, K. E. Sanderson, J. Spieth, S. W. Clifton, P. Latreille, L. Courtney, S. Porwollik, J. Ali, M. Dante, F. Du, S. Hou, D. Layman, S. Leonard, C. Nguyen, K. Scott, A. Holmes, N. Grewal, E. Mulvaney, E. Ryan, H. Sun, L. Florea, W. Miller, T. Stoneking, M. Nhan,

R. Waterston, and R. K. Wilson, "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2," *Nature*, vol. 413, pp. 852–856, oct 2001.

[12] N. J. Loman and M. J. Pallen, "Twenty years of bacterial genome sequencing," *Nature Reviews Microbiology*, vol. 13, pp. 787–794, nov 2015.

[13] J. M. Prober, G. L. Trainor, R. J. Dam, F. W. Hobbs, C. W. Robertson, R. J. Zagursky, A. J. Cocuzza, M. A. Jensen, and K. Baumeister, "A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides," *Science*, vol. 238, no. 4825, pp. 336–341, 1987.

[14] "Applied Biosystems Genetic Analysis Systems | Thermo Fisher Scientific - DK."

[15] J. Z. Zhang, Y. Fang, J. Y. Hou, H. J. Ren, R. Jiang, P. Roos, and N. J. Dovich, "Use of Non-Cross-Linked Polyacrylamide for Four-Color DNA Sequencing by Capillary Electrophoresis Separation of Fragments up to 640 Bases in Length in Two Hours," *Analytical Chemistry*, vol. 67, no. 24, pp. 4589–4593, 1995.

[16] S. Iyer, E. Casey, H. Bouzek, M. Kim, W. Deng, B. B. Larsen, H. Zhao, R. E. Bumgarner, M. Rolland, and J. I. Mullins, "Comparison of major and minor viral SNPs identified through single template sequencing and pyrosequencing in acute HIV-1 infection," *PLoS ONE*, vol. 10, aug 2015.

[17] C. J. Houldcroft, M. A. Beale, and J. Breuer, "Clinical and biological insights from viral genome sequencing," feb 2017.

[18] J. Messing, R. Crea, and P. H. Seeburg, "A system for shotgun DNA sequencing," *Nucleic Acids Research*, vol. 9, no. 2, pp. 309–321, 1981.

[19] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich, "Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 51, no. 1, pp. 263–273, 1986.

[20] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim, "Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia," *Science*, vol. 230, no. 4732, pp. 1350–1354, 1985.

[21] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, "DNA sequencing at 40: past, present and future," *Nature*, vol. 550, pp. 345–353, oct 2017.

[22] R. Staden, "A strategy of DNA sequencing employing computer programs.," *Nucleic acids research*, vol. 6, pp. 2601–10, jun 1979.

[23] M. W. Allard, R. Bell, C. M. Ferreira, N. Gonzalez-Escalona, M. Hoffmann, T. Muruvanda, A. Ottesen, P. Ramachandran, E. Reed, S. Sharma, E. Stevens, R. Timme, J. Zheng, and E. W. Brown, "Genomics of foodborne pathogens for microbial food safety," *Current Opinion in Biotechnology*, vol. 49, pp. 224–229, feb 2018.

[24] M. V. Larsen, S. Cosentino, O. Lukjancenko, D. Saputra, S. Rasmussen, H. Hasman, T. Sicheritz-Ponten, F. M. Aarestrup, D. W. Ussery, and O. Lund, "Benchmarking of Methods for Genomic Taxonomy," *Journal of Clinical Microbiology*, vol. 52, pp. 1529–1539, may 2014.

[25] H. Hasman, D. Saputra, T. Sicheritz-Ponten, O. Lund, C. A. Svendsen, N. Frimodt-Moller, and F. M. Aarestrup, "Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples," *Journal of Clinical Microbiology*, vol. 52, pp. 139–146, jan 2014.

[26] K. G. Joensen, A. M. Tetzschner, A. Iguchi, F. M. Aarestrup, and F. Scheutz, "Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data," *Journal of Clinical Microbiology*, vol. 53, pp. 2410–2426, aug 2015.

[27] E. Zankari, H. Hasman, S. Cosentino, M. Vestergaard, S. Rasmussen, O. Lund, F. M. Aarestrup, and M. V. Larsen, "Identification of acquired antimicrobial resistance genes.," *The Journal of antimicrobial chemotherapy*, vol. 67, pp. 2640–4, nov 2012.

[28] J. J. Davis, S. Boisvert, T. Brettin, R. W. Kenyon, C. Mao, R. Olson, R. Overbeek, J. Santerre, M. Shukla, A. R. Wattam, R. Will, F. Xia, and R. Stevens, "Antimicrobial Resistance Prediction in PATRIC and RAST," *Scientific Reports*, vol. 6, pp. 1–12, jun 2016.

[29] K. G. Joensen, F. Scheutz, O. Lund, H. Hasman, R. S. Kaas, E. M. Nielsen, and F. M. Aarestrup, "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli.," *Journal of clinical microbiology*, vol. 52, pp. 1501–10, may 2014.

[30] S. R. Harris, E. J. Feil, M. T. Holden, M. A. Quail, E. K. Nickerson, N. Chantratita, S. Gardete, A. Tavares, N. Day, J. A. Lindsay, J. D. Edgeworth, H. De Lencastre, J. Parkhill, S. J. Peacock, and S. D. Bentley, "Evolution of MRSA during hospital transmission and intercontinental spread," *Science*, vol. 327, pp. 469–474, jan 2010.

[31] M. L. Metzker, "Sequencing technologies — the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[32] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, pp. 186–194, mar 1998.

[33] H. P. Buermans and J. T. den Dunnen, "Next generation sequencing technology: Advances and applications," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1842, pp. 1932–1941, oct 2014.

[34] A. Ameur, W. P. Kloosterman, and M. S. Hestand, "Single-Molecule Sequencing: Towards Clinical Applications," jan 2019.

[35] N. Kono and K. Arakawa, "Nanopore sequencing: Review of potential applications in functional genomics," *Development, Growth & Differentiation*, vol. 61, pp. 316–326, jun 2019.

[36] H. Teng, M. D. Cao, M. B. Hall, T. Duarte, S. Wang, and L. J. M. Coin, "Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning.," *GigaScience*, vol. 7, no. 5, 2018.

[37] Y. Z. Zhang, A. Akdemir, G. Tremmel, S. Imoto, S. Miyano, T. Shibuya, and R. Yamaguchi, "Nanopore basecalling from a perspective of instance segmentation," *BMC Bioinformatics*, vol. 21, apr 2020.

[38] X. Didelot, R. Bowden, D. J. Wilson, T. E. a. Peto, and D. W. Crook, "Transforming clinical microbiology with bacterial genome sequencing," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 601–612, 2012.

[39] J. R. Stephenson and A. Warnes, eds., *Diagnostic Virology Protocols*, vol. 665 of *Methods in Molecular Biology*. Totowa, NJ: Humana Press, 2011.

[40] P. M. Fratamico, C. DebRoy, Y. Liu, D. S. Needleman, G. M. Baranzoni, and P. Feng, "Advances in Molecular Serotyping and Subtyping of Escherichia coli†," *Frontiers in Microbiology*, vol. 7, p. 644, may 2016.

[41] M. P. Ryan, J. O'Dwyer, and C. C. Adley, "Evaluation of the Complex Nomenclature of the Clinically and Veterinary Significant Pathogen Salmonella.," *BioMed research international*, vol. 2017, p. 3782182, 2017.

[42] D. J. Ingle, M. Valcanis, A. Kuzevski, M. Tauschek, M. Inouye, T. Stinear, M. M. Levine, R. M. Robins-Browne, and K. E. Holt, "In silico serotyping of E. Coli from short read data identifies limited novel o-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages," *Microbial Genomics*, vol. 2, jul 2016.

[43] P. Wattiau, C. Boland, and S. Bertrand, "Methodologies for Salmonella enterica subsp. Enterica Subtyping: Gold Standards and Alternatives," *Applied and Environmental Microbiology*, vol. 77, pp. 7877–7885, nov 2011.

[44] B. Swaminathan, T. J. Barrett, S. B. Hunter, and R. V. Tauxe, "PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States," *Emerging Infectious Diseases*, vol. 7, pp. 382–389, jun 2001.

[45] F. Wu and P. Della-Latta, "Pulsed-field gel electrophoresis," in *Advanced Techniques in Diagnostic Microbiology*, pp. 143–157, Springer US, 2006.

[46] B. R. Jackson, C. Tarr, E. Strain, K. A. Jackson, A. Conrad, H. Carleton, L. S. Katz, S. Stroika, L. H. Gould, R. K. Mody, B. J. Silk, J. Beal, Y. Chen, R. Timme, M. Doyle, A. Fields, M. Wise, G. Tillman, S. Defibaugh-Chavez, Z. Kucerova, A. Sabol, K. Roache, E. Trees, M. Simmons, J. Wasilenko, K. Kubota, H. Pouseele, W. Klimke, J. Besser, E. Brown, M. Allard, and P. Gerner-Smidt, "Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation.," *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, vol. 63, pp. 380–386, aug 2016.

[47] M. C. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt, "Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 3140–3145, mar 1998.

[48] M. C. Maiden, "Multilocus sequence typing of bacteria," 2006.

[49] X. Deng, H. C. den Bakker, and R. S. Hendriksen, "Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens," *Annu Rev Food Sci Technol*, vol. 7, pp. 1–22, feb 2016.

[50] E. J. Feil, B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt, "eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data," *Journal of Bacteriology*, vol. 186, pp. 1518–1530, mar 2004.

[51] M. Achtman, "Insights from genomic comparisons of genetically monomorphic bacterial pathogens," 2012.

[52] K. A. Jolley and M. C. Maiden, "BIGSdb: Scalable analysis of bacterial genome variation at the population level," *BMC Bioinformatics*, vol. 11, p. 595, dec 2010.

[53] N.-F. Alikhan, Z. Zhou, M. J. Sergeant, and M. Achtman, "A genomic overview of the population structure of Salmonella," *PLOS Genetics*, vol. 14, p. e1007261, apr 2018.

[54] M. Silva, M. P. Machado, D. N. Silva, M. Rossi, J. Moran-Gilad, S. Santos, M. Ramirez, and J. A. Carriço, "chewBBACA: A complete suite for gene-by-gene schema creation and strain identification," *Microbial genomics*, vol. 4, mar 2018.

[55] "Pathogenwatch | A Global Platform for Genomic Surveillance."

[56] M. De Been, M. Pinholt, J. Top, S. Bletz, A. Mellmann, W. Van Schaik, E. Brouwer, M. Rogers, Y. Kraat, M. Bonten, J. Corander, H. Westh, D. Harmsen, and R. J. Willems, "Core genome multilocus sequence typing scheme for high-resolution typing of enterococcus faecium," *Journal of Clinical Microbiology*, vol. 53, pp. 3788–3797, dec 2015.

[57] A. Moura, A. Criscuolo, H. Pouseele, M. M. Maury, A. Leclercq, C. Tarr, J. T. Björkman, T. Dallman, A. Reimer, V. Enouf, E. Larsonneur, H. Carleton, H. Bracq-Dieye, L. S. Katz, L. Jones, M. Touchon, M. Tourdjman, M. Walker, S. Stroika, T. Cantinelli, V. Chenal-Francisque, Z. Kucerova, E. P. Rocha, C. Nadon, K. Grant, E. M. Nielsen, B. Pot, P. Gerner-Smidt, M. Lecuit, and S. Brisse, "Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes," *Nature Microbiology*, vol. 2, oct 2016.

[58] A. Dangel, A. Berger, U. Messelhäußer, R. Konrad, S. Hörmansdorfer, N. Ackermann, and A. Sing, "Genetic diversity and delineation of Salmonella agona outbreak strains by next generation sequencing, Bavaria, Germany, 1993 to 2018," *Eurosurveillance*, vol. 24, no. 18, 2019.

[59] S. R. Leopold, R. V. Goering, A. Witten, D. Harmsen, and A. Mellmann, "Bacterial whole-genome sequencing revisited: Portable, scalable, and standardized analysis for typing and detection of virulence and antibiotic resistance genes," *Journal of Clinical Microbiology*, vol. 52, no. 7, pp. 2365–2370, 2014.

[60] C. Henri, P. Leekitcharoenphon, H. A. Carleton, N. Radomski, R. S. Kaas, J. F. Mariet, A. Felten, F. M. Aarestrup, P. G. Smidt, S. Roussel, L. Guillier, M. Y. Mistou, and R. S. Hendriksen, "An assessment of different genomic approaches for inferring phylogeny of Listeria monocytogenes," *Frontiers in Microbiology*, vol. 8, p. 2351, nov 2017.

[61] C. E. Coipan, T. J. Dallman, D. Brown, H. Hartman, M. van der Voort, R. R. van den Berg, D. Palm, S. Kotila, T. van Wijk, and E. Franz, "Concordance of SNP-and allele-based typing workflows in the context of a large-scale international Salmonella enteritidis outbreak investigation," *Microbial Genomics*, vol. 6, no. 3, 2020.

[62] F. Gona, F. Comandatore, S. Battaglia, A. Piazza, A. Trovato, G. Lorenzin, P. Cichero, A. Biancardi, P. Nizzero, M. Moro, and D. M. Cirillo, "Comparison of core-genome MLST, coreSNP and PFGE methods for klebsiella pneumoniae cluster analysis," *Microbial Genomics*, vol. 6, p. e000347, mar 2020.

[63] L. Uelze, J. Grützke, M. Borowiak, J. A. Hammerl, K. Juraschek, C. Deneke, S. H. Tausch, and B. Malorny, "Typing methods based on whole genome sequencing data," *One Health Outlook*, vol. 2, pp. 1–19, dec 2020.

[64] Z. Zhou, N.-F. Alikhan, M. J. Sergeant, N. Luhmann, C. Vaz, A. P. Francisco, J. A. Carriço, and M. Achtman, "GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens.," *Genome research*, vol. 28, no. 9, pp. 1395–1404, 2018.

[65] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: Architecture and applications," *BMC Bioinformatics*, vol. 10, p. 421, dec 2009.

[66] T. Seemann, "mlst."

[67] L. M. Brinkac, E. Beck, J. Inman, P. Venepally, D. E. Fouts, and G. Sutton, "LOCUST: A custom sequence locus typer for classifying microbial isolates," *Bioinformatics*, vol. 33, pp. 1725–1726, jun 2017.

[68] M. V. Larsen, S. Cosentino, S. Rasmussen, C. Friis, H. Hasman, R. L. Marvig, L. Jelsbak, T. Sicheritz-Pontén, D. W. Ussery, F. M. Aarestrup, and O. Lund, "Multilocus sequence typing of total-genome-sequenced bacteria," *Journal of Clinical Microbiology*, vol. 50, pp. 1355–1361, apr 2012.

[69] M. Inouye, H. Dashnow, L. A. Raven, M. B. Schultz, B. J. Pope, T. Tomita, J. Zobel, and K. E. Holt, "SRST2: Rapid genomic surveillance for public health and hospital microbiology labs," *Genome Medicine*, vol. 6, nov 2014.

[70] P. Feijao, H. T. Yao, D. Fornika, J. Gardy, W. Hsiao, C. Chauve, and L. Chindelevitch, "MentaLiST - A fast MLST caller for large MLST schemes," *Microbial genomics*, vol. 4, feb 2018.

[71] L. S. Katz, T. Griswold, A. J. Williams-Newkirk, D. Wagner, A. Petkau, C. Sieffert, G. V. Domselaar, X. Deng, and H. A. Carleton, "A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens," *Frontiers in Microbiology*, vol. 8, mar 2017.

[72] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1754–60, jul 2009.

[73] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," mar 2013.

[74] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, pp. 357–359, apr 2012.

[75] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, pp. 3094–3100, sep 2018.

[76] P. T. Clausen, F. M. Aarestrup, and O. Lund, "Rapid and precise alignment of raw reads against redundant databases with KMA," *BMC Bioinformatics*, vol. 19, p. 307, dec 2018.

[77] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, pp. 1297–1303, sep 2010.

[78] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson, "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.," *Genome research*, vol. 22, pp. 568–76, mar 2012.

[79] R. S. Kaas, P. Leekitcharoenphon, F. M. Aarestrup, and O. Lund, "Solving the problem of comparing whole bacterial genomes across different sequencing platforms.," *PloS one*, vol. 9, p. e104984, jan 2014.

[80] S. Davis, J. B. Pettengill, Y. Luo, J. Payne, A. Shpuntoff, H. Rand, and E. Strain, "CFSAN SNP pipeline: An automated method for constructing snp matrices fromnext-generation sequence data," *PeerJ Computer Science*, vol. 2015, p. e20, aug 2015.

[81] A. Petkau, P. Mabon, C. Sieffert, N. C. Knox, J. Cabral, M. Iskander, M. Iskander, K. Weedmark, R. Zaheer, L. S. Katz, C. Nadon, A. Reimer, E. Taboada, R. G. Beiko, W. Hsiao, F. Brinkman, M. Graham, and G. Van Domselaar, "SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology," *Microbial genomics*, vol. 3, p. e000116, jun 2017.

[82] T. Dallman, P. Ashton, U. Schafer, A. Jironkin, A. Painset, S. Shaaban, H. Hartman, R. Myers, A. Underwood, C. Jenkins, and K. Grant, "SnapperDB: a database solution for routine sequencing analysis of bacterial isolates," *Bioinformatics*, vol. 81, pp. 3946–3952, apr 2018.

[83] P. Leekitcharoenphon, E. M. Nielsen, R. S. Kaas, O. Lund, and F. M. Aarestrup, "Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica," *PLoS ONE*, vol. 9, p. e87991, feb 2014.

[84] A. D. Young and J. P. Gillung, "Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics," *Systematic Entomology*, vol. 45, pp. 225–247, apr 2020.

[85] A. Holmes, L. Allison, M. Ward, T. J. Dallman, R. Clark, A. Fawkes, L. Murphy, and M. Hanson, "Utility of whole-genome sequencing of Escherichia coli o157 for outbreak detection and epidemiological surveillance," *Journal of Clinical Microbiology*, vol. 53, pp. 3565–3573, nov 2015.

[86] X. Didelot and D. J. Wilson, "ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes," *PLOS Computational Biology*, vol. 11, p. e1004041, feb 2015.

[87] J. Felsenstein, *Inferring phylogenies*. Sinauer Associates, 2004.

[88] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, no. 28, pp. 1409–1438, 1958.

[89] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees.," *Molecular biology and evolution*, vol. 4, pp. 406–25, jul 1987.

[90] J. Studier and K. Keppler, "A note on the neighbor-joining algorithm of Saitou and Nei.," *Molecular Biology and Evolution*, vol. 5, pp. 729–731, nov 1988.

[91] M. Hasegawa, H. Kishino, and T. aki Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, pp. 160–174, oct 1985.

[92] C. Lanave, G. Preparata, C. Sacone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of Molecular Evolution*, vol. 20, pp. 86–93, feb 1984.

[93] S. Abadi, D. Azouri, T. Pupko, and I. Mayrose, "Model selection may not be a mandatory step for phylogeny reconstruction," *Nature Communications*, vol. 10, pp. 1–11, dec 2019.

[94] A. Dhar and V. Minin, "Maximum Likelihood Phylogenetic Inference," in *Encyclopedia of Evolutionary Biology*, pp. 499–506, Elsevier, 2016.

[95] A. Stamatakis, "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 1312–3, may 2014.

[96] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.," *Systematic biology*, vol. 52, pp. 696–704, oct 2003.

[97] M. N. Price, P. S. Dehal, A. P. Arkin, E. Nawrocki, D. Kolbe, S. Eddy, M. Price, P. Dehal, A. Arkin, N. Saitou, M. Nei, J. Studier, K. Keppler, J. Felsenstein, S. Roch, S. Guindon, O. Gascuel, W. Hordijk, O. Gascuel, A. Stamatakis, J. Felsenstein, A. Stamatakis, P. Hoover, J. Rougemont, R. Desper, O. Gascuel, M. Nei, S. Kumar, K. Takahashi, A. Stamatakis, Z. Yang, S. Guindon, F. Delsuc, J. Dufayard, O. Gascuel, S. Guindon, J. Dufayard, V. Lefort, MAnisimova, W. Hordijk, O. Gascuel, H. Shimodaira, M. Hasegawa, R. Tatusov, D. Natale, I. Garkavtsev, T. Tatusova, U. Shankavaram, M. Anisimova, O. Gascuel, N. Galtier, A. Jean-Marie, E. DeLong, D. Clarke-Pearson, J. Evans, L. Sheneman, J. Foster, T. Wheeler, R. Desper, O. Gascuel, M. Bordewich, O. Gascuel, K. Huber, V. Moulton, H. Shimodaira, M. Hasegawa, P. Goloboff, J. Farris, K. Nixon, K. Katoh, H. Toh, R. Bradley, A. Roberts, M. Smoot, S. Juvekar, J. D. Ea, E. Sonnhammer, V. Hollich, T. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. Brodie, J. Stoye, D. Evers, and F. Meyer, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments," *PLoS ONE*, vol. 5, p. e9490, mar 2010.

[98] L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh, "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.," *Molecular biology and evolution*, vol. 32, pp. 268–74, jan 2015.

[99] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, pp. 1–26, jan 1979.

[100] J. Felsenstein, "CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP," *Evolution*, vol. 39, pp. 783–791, jul 1985.

[101] B. Q. Minh, M. A. T. Nguyen, and A. Von Haeseler, "Ultrafast approximation for phylogenetic bootstrap," *Molecular Biology and Evolution*, vol. 30, pp. 1188–1195, may 2013.

[102] D. T. Hoang, O. Chernomor, A. Von Haeseler, B. Q. Minh, and L. S. Vinh, "UFBoot2: Improving the ultrafast bootstrap approximation," *Molecular Biology and Evolution*, vol. 35, pp. 518–522, feb 2018.

[103] B. Q. Minh, M. W. Hahn, and R. Lanfear, "New Methods to Calculate Concordance Factors for Phylogenomic Datasets," *Molecular biology and evolution*, vol. 37, pp. 2727–2733, sep 2020.

[104] S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel, "New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0," *Systematic Biology*, vol. 59, pp. 307–321, may 2010.

[105] D. A. Rasko, P. L. Worsham, T. G. Abshire, S. T. Stanley, J. D. Bannan, M. R. Wilson, R. J. Langham, R. S. Decker, L. Jiang, T. D. Read, A. M. Phillippy, S. L. Salzberg, M. Pop, M. N. Van Ert, L. J. Kenefic, P. S. Keim, C. M. Fraser-Liggett, and J. Ravel, "Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, pp. 5027–5032, mar 2011.

[106] D. M. Engelthaler, T. Chiller, J. A. Schupp, J. Colvin, S. M. Beckstrom-Sternberg, E. M. Driebe, T. Moses, W. Tembe, S. Sinari, J. S. Beckstrom-Sternberg, A. Christoforides, J. V. Pearson, J. Carpten, P. Keim, A. Peterson, D. Terashita, and S. Arunmozhi Balajee, "Next-generation sequencing of Coccidioides immitis isolated during cluster investigation," *Emerging Infectious Diseases*, vol. 17, pp. 227–232, feb 2011.

[107] J. L. Gardy, J. C. Johnston, S. J. Ho Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. Jones, F. S. Brinkman, R. C. Brunham, and P. Tang, "Whole-genome sequencing and social-network analysis of a tuberculosis outbreak," *New England Journal of Medicine*, vol. 364, pp. 730–739, feb 2011.

[108] R. S. Hendriksen, L. B. Price, J. M. Schupp, J. D. Gillece, R. S. Kaas, D. M. Engelthaler, V. Bortolaia, T. Pearson, A. E. Waters, B. P. Upadhyay, S. D. Shrestha, S. Adhikari, G. Shaky, P. S. Keim, and F. M. Aarestrupa, "Population genetics of vibrio cholerae from Nepal in 2010: Evidence on the origin of the haitian outbreak," *mBio*, vol. 2, pp. 157–168, sep 2011.

[109] X. Didelot, D. W. Eyre, M. Cule, C. L. Ip, M. A. Ansari, D. Griffiths, A. Vaughan, L. O'Connor, T. Golubchik, E. M. Batty, P. Piazza, D. J. Wilson, R. Bowden, P. J. Donnelly, K. E. Dingle, M. Wilcox, A. S. Walker, D. W. Crook, T. E. Peto, and R. M. Harding, "Microevolutionary analysis of Clostridium difficile genomes to investigate transmission," *Genome biology*, vol. 13, no. 12, p. R118, 2012.

[110] A. Mellmann, D. Harmsen, C. A. Cummings, E. B. Zentz, S. R. Leopold, A. Rico, K. Prior, R. Szczepanowski, Y. Ji, W. Zhang, S. F. McLaughlin, J. K. Henkhaus, B. Leopold, M. Bielaszewska, R. Prager, P. M. Brzoska, R. L. Moore, S. Guenther, J. M. Rothberg, and H. Karch, "Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology," *PLoS ONE*, vol. 6, p. e22751, jul 2011.

[111] H. Rohde, J. Qin, Y. Cui, D. Li, N. J. Loman, M. Hentschke, W. Chen, F. Pu, Y. Peng, J. Li, F. Xi, S. Li, Y. Li, Z. Zhang, X. Yang, M. Zhao, P. Wang, Y. Guan, Z. Cen, X. Zhao, M. Christner, R. Kobbe, S. Loos, J. Oh, L. Yang, A. Danchin, G. F. Gao, Y. Song, Y. Li, H. Yang, J. Wang, J. Xu, M. J. Pallen, J. Wang, M. Aepfelbacher, and R. Yang, "Open-Source Genomic Analysis of Shiga-Toxin–Producing E. coli O104:H4," *New England Journal of Medicine*, vol. 365, pp. 718–724, aug 2011.

[112] E. N. Taboada, M. R. Graham, J. A. Carriço, and G. Van Domselaar, "Food Safety in the Age of Next Generation Sequencing, Bioinformatics, and Open Data Access," *Frontiers in Microbiology*, vol. 8, p. 909, may 2017.

[113] WHO, *Whole genome sequencing for foodborne disease surveillance: landscape paper*. Geneva: World Health Organization, 2018.

[114] R. E. Timme, M. Sanchez Leon, and M. W. Allard, "Utilizing the Public GenomeTrakr Database for Foodborne Pathogen Traceback," in *Foodborne Bacterial Pathogens. Methods in Molecular Biology* (A. Bridier, ed.), vol. 1918, pp. 201–212, Humana, New York, NY, 2019.

[115] A. Moura, M. Tourdjman, A. Leclercq, E. Hamelin, E. Laurent, N. Fredriksen, D. van Cauteren, H. Bracq-Dieye, P. Thouvenot, G. Vales, N. Tessaud-Rita, M. M. Maury, A. Alexandru, A. Criscuolo, E. Quevillon, M. P. Donguy, V. Enouf, H. de Valk, S. Brisse, and M. Lecuit, "Real-time whole-genome sequencing for surveillance of Listeria monocytogenes, France," *Emerging Infectious Diseases*, vol. 23, pp. 1462–1470, sep 2017.

[116] A. K. Jensen, E. M. Nielsen, J. T. Björkman, T. Jensen, L. Müller, S. Persson, G. Bjerager, A. Perge, T. G. Krause, K. Kiil, G. Sørensen, J. K. Andersen, K. Mølbak, and S. Ethelberg, "Whole-genome sequencing used to investigate a nationwide outbreak of listeriosis caused by ready-to-eat delicatessen meat, Denmark, 2014," *Clinical Infectious Diseases*, vol. 63, pp. 64–70, jul 2016.

[117] S. Schjørring, S. Gillesberg Lassen, T. Jensen, A. Moura, J. S. Kjeldgaard, L. Müller, S. Thielke, A. Leclercq, M. M. Maury, M. Tourdjman, M. P. Donguy, M. Lecuit, S. Ethelberg, and E. M. Nielsen, "Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017," *Eurosurveillance*, vol. 22, dec 2017.

[118] T. Inns, P. M. Ashton, S. Herrera-Leon, J. Lighthill, S. Foulkes, T. Jombart, Y. Rehman, A. Fox, T. Dallman, E. De Pinna, L. Browning, J. E. Coia, O. Edeghere, and R. Vivancos, "Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of Salmonella Enteritidis," *Epidemiology and Infection*, vol. 145, pp. 289–298, jan 2017.

[119] A. Holmes, T. J. Dallman, S. Shabaan, M. Hanson, and L. Allison, "Validation of whole-genome sequencing for identification and characterization of Shiga toxin-producing Escherichia coli to produce standardized data to enable data sharing," *Journal of Clinical Microbiology*, vol. 56, mar 2018.

[120] P. Mook, D. Gardiner, N. Q. Verlander, J. McCormick, M. Usdin, P. Crook, C. Jenkins, and T. J. Dallman, "Operational burden of implementing Salmonella Enteritidis and Typhimurium cluster detection using whole genome sequencing surveillance data in England: A retrospective assessment," *Epidemiology and Infection*, vol. 146, pp. 1452–1460, aug 2018.

[121] "British Columbia Integrated Surveillance of Foodborne Pathogens (BCISFP) Annual Summary of Salmonella Findings 2018," 2019.

[122] I. M. Rubin, T. A. Hansen, A. M. Klingenberg, A. M. Petersen, P. Worning, H. Westh, and M. D. Bartels, "A Sporadic Four-Year Hospital Outbreak of a ST97-IVa MRSA With Half of the Patients First Identified in the Community," *Frontiers in Microbiology*, vol. 9, p. 1494, jul 2018.

[123] M. Pinholt, H. Larner-Svensson, P. Littauer, C. E. Moser, M. Pedersen, L. E. Lemming, T. Ejlertsen, T. S. Søndergaard, B. J. Holzknecht, U. S. Justesen, E. Dzajic, S. S. Olsen, J. B. Nielsen, P. Worning, A. M. Hammerum, H. Westh, and L. Jakobsen, "Multiple hospital outbreaks of vanA Enterococcus faecium in Denmark, 2012-13, investigated by WGS, MLST and PFGE.," *The Journal of antimicrobial chemotherapy*, vol. 70, pp. 2474–82, sep 2015.

[124] K. E. Raven, T. Gouliouris, H. Brodrick, F. Coll, N. M. Brown, R. Reynolds, S. Reuter, M. E. Török, J. Parkhill, and S. J. Peacock, "Complex routes of nosocomial vancomycin-resistant Enterococcus faecium transmission revealed by genome sequencing," *Clinical Infectious Diseases*, vol. 64, pp. 886–893, apr 2017.

[125] J. Liese, L. Schüle, P. Oberhettinger, L. Tschörner, T. Nguyen, D. Dörfel, W. Vogel, M. Marschal, I. Autenrieth, M. Willmann, and S. Peter, "Expansion of Vancomycin-Resistant Enterococcus faecium in an Academic Tertiary Hospital in Southwest Germany: A large-scale whole-genome-based outbreak investigation," *Antimicrobial Agents and Chemotherapy*, vol. 63, may 2019.

[126] A. Mellmann, S. Bletz, T. Böking, F. Kipp, K. Becker, A. Schultes, K. Prior, and D. Harmsen, "Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting," *Journal of Clinical Microbiology*, vol. 54, pp. 2874–2881, dec 2016.

[127] S. J. Peacock, J. Parkhill, and N. M. Brown, "Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens," *Microbiology*, vol. 164, pp. 1213–1219, oct 2018.

[128] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges*, vol. 1, pp. 33–46, jan 2017.

[129] J. van Beek, M. de Graaf, H. Al-Hello, D. J. Allen, K. Ambert-Balay, N. Botteldoorn, M. Brytting, J. Buesa, M. Cabrerizo, M. Chan, F. Cloak, I. Di Bartolo, S. Guix, J. Hewitt, N. Iritani, M. Jin, R. Johne, I. Lederer, J. Mans, V. Martella, L. Maunula, G. McAllister, S. Niendorf, H. G. Niesters, A. T. Podkolzin, M. Poljsak-Prijatelj, L. D. Rasmussen, G. Reuter, G. Tuite, A. Kroneman, H. Vennema, and M. P. Koopmans, "Molecular surveillance of norovirus, 2005–16: an epidemiological analysis of data collected from the NoroNet network," *The Lancet Infectious Diseases*, vol. 18, pp. 545–553, may 2018.

[130] K. Li, S. Shrivastava, A. Brownley, D. Katzel, J. Bera, A. T. Nguyen, V. Thovarai, R. Halpin, and T. B. Stockwell, "Automated degenerate PCR primer design for high-throughput sequencing improves efficiency of viral sequencing," *Virology Journal*, vol. 9, p. 261, dec 2012.

[131] J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, N. Ouédraogo, B. Afrough, A. Bah, J. H. Baum, B. Becker-Ziaja, J. P. Boettcher, M. Cabeza-Cabrerizo, Á. Camino-Sánchez, L. L. Carter, J. Doerrbecker, T. Enkirch, I. Garciá-Dorival, N. Hetzelt, J. Hinzmann, T. Holm, L. E. Kafetzopoulou, M. Koropogui, A. Kosgey, E. Kuisma, C. H. Logue, A. Mazzarelli, S. Meisel, M. Mertens, J. Michel, D. Ngabo, K. Nitzsche, E. Pallasch, L. V. Patrono, J. Portmann, J. G. Repits, N. Y. Rickett, A. Sachse, K. Singethan, I. Vitoriano, R. L. Yemanaberhan, E. G. Zekeng, T. Racine, A. Bello, A. A. Sall, O. Faye, O. Faye, N. Magassouba, C. V. Williams, V. Amburgey, L. Winona, E. Davis, J. Gerlach, F. Washington, V. Monteil, M. Jourdain, M. Bererd, A. Camara, H. Somlare, A. Camara, M. Gerard, G. Bado, B. Baillet, D. Delaune,

K. Y. Nebie, A. Diarra, Y. Savane, R. B. Pallawo, G. J. Gutierrez, N. Milhano, I. Roger, C. J. Williams, F. Yattara, K. Lewandowski, J. Taylor, P. Rachwal, D. J. Turner, G. Pollakis, J. A. Hiscox, D. A. Matthews, M. K. O'Shea, A. M. D. Johnston, D. Wilson, E. Hutley, E. Smit, A. Di Caro, R. Wolfel, K. Stoecker, E. Fleischmann, M. Gabriel, S. A. Weller, L. Koivogui, B. Diallo, S. Keita, A. Rambaut, P. Formenty, S. Gunther, and M. W. Carroll, "Real-time, portable genome sequencing for Ebola surveillance," *Nature*, vol. 530, pp. 228–232, feb 2016.

[132] J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. De Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, and N. J. Loman, "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples," *Nature Protocols*, vol. 12, pp. 1261–1266, jun 2017.

[133] G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, F. Bielejec, S. L. Caddy, M. Cotten, J. D'Ambrozio, S. Dellicour, A. Di Caro, J. W. Diclaro, S. Duraffour, M. J. Elmore, L. S. Fakoli, O. Faye, M. L. Gilbert, S. M. Gevao, S. Gire, A. Gladden-Young, A. Gnirke, A. Goba, D. S. Grant, B. L. Haagmans, J. A. Hiscox, U. Jah, J. R. Kugelman, D. Liu, J. Lu, C. M. Malboeuf, S. Mate, D. A. Matthews, C. B. Matranga, L. W. Meredith, J. Qu, J. Quick, S. D. Pas, M. V. Phan, G. Pollakis, C. B. Reusken, M. Sanchez-Lockhart, S. F. Schaffner, J. S. Schieffelin, R. S. Sealfon, E. Simon-Loriere, S. L. Smits, K. Stoecker, L. Thorne, E. A. Tobin, M. A. Vandi, S. J. Watson, K. West, S. Whitmer, M. R. Wiley, S. M. Winnicki, S. Wohl, R. Wölfel, N. L. Yozwiak, K. G. Andersen, S. O. Blyden, F. Bolay, M. W. Carroll, B. Dahn, B. Diallo, P. Formenty, C. Fraser, G. F. Gao, R. F. Garry, I. Goodfellow, S. Günther, C. T. Happi, E. C. Holmes, B. Kargbo, S. Keïta, P. Kellam, M. P. Koopmans, J. H. Kuhn, N. J. Loman, N. Magassouba, D. Naidoo, S. T. Nichol, T. Nyenswah, G. Palacios, O. G. Pybus, P. C. Sabeti, A. Sall, U. Ströher, I. Wurie, M. A. Suchard, P. Lemey, and A. Rambaut, "Virus genomes reveal factors that spread and sustained the Ebola epidemic," *Nature*, vol. 544, pp. 309–315, apr 2017.

[134] H. C. Metsky, C. B. Matranga, S. Wohl, S. F. Schaffner, C. A. Freije, S. M. Winnicki, K. West, J. Qu, M. L. Baniecki, A. Gladden-Young, A. E. Lin, C. H. Tomkins-Tinch, S. H. Ye, D. J. Park, C. Y. Luo, K. G. Barnes, R. R. Shah, B. Chak, G. Barbosa-Lima, E. Delatorre, Y. R. Vieira, L. M. Paul, A. L. Tan, C. M. Barcellona, M. C. Porcelli, C. Vasquez, A. C. Cannons, M. R. Cone, K. N. Hogan, E. W. Kopp, J. J. Anzinger, K. F. Garcia, L. A. Parham, R. M. Ramírez, M. C. Montoya, D. P. Rojas, C. M. Brown, S. Hennigan, B. Sabina, S. Scotland, K. Gangavarapu, N. D. Grubaugh, G. Oliveira, R. Robles-Sikisaka, A. Rambaut, L. Gehrke, S. Smole, M. E. Halloran, L. Villar, S. Mattar, I. Lorenzana, J. Cerbino-Neto, C. Valim, W. Degrave, P. T. Bozza, A. Gnirke, K. G. Andersen, S. Isern, S. F. Michael, F. A. Bozza, T. M. Souza, I. Bosch, N. L. Yozwiak, B. L. Macinnis, and P. C. Sabeti, "Zika virus evolution and spread in the Americas," *Nature*, vol. 546, pp. 411–415, jun 2017.

[135] N. D. Grubaugh, J. T. Ladner, M. U. Kraemer, G. Dudas, A. L. Tan, K. Gangavarapu, M. R. Wiley, S. White, J. Thézé, D. M. Magnani, K. Prieto, D. Reyes, A. M. Bingham, L. M. Paul, R. Robles-Sikisaka, G. Oliveira, D. Pronty, C. M. Barcellona, H. C. Metsky, M. L. Baniecki, K. G. Barnes, B. Chak, C. A. Freije, A. Gladden-Young, A. Gnirke, C. Luo, B. MacInnis, C. B. Matranga, D. J. Park, J. Qu, S. F. Schaffner, C. Tomkins-Tinch, K. L.

West, S. M. Winnicki, S. Wohl, N. L. Yozwiak, J. Quick, J. R. Fauver, K. Khan, S. E. Brent, R. C. Reiner, P. N. Lichtenberger, M. J. Ricciardi, V. K. Bailey, D. I. Watkins, M. R. Cone, E. W. Kopp, K. N. Hogan, A. C. Cannons, R. Jean, A. J. Monaghan, R. F. Garry, N. J. Loman, N. R. Faria, M. C. Porcelli, C. Vasquez, E. R. Nagle, D. A. Cummings, D. Stanek, A. Rambaut, M. Sanchez-Lockhart, P. C. Sabeti, L. D. Gillis, S. F. Michael, T. Bedford, O. G. Pybus, S. Isern, G. Palacios, and K. G. Andersen, "Genomic epidemiology reveals multiple introductions of Zika virus into the United States," *Nature*, vol. 546, pp. 401–405, jun 2017.

[136] N. R. Faria, J. Quick, I. M. Claro, J. Thézé, J. G. De Jesus, M. Giovanetti, M. U. Kraemer, S. C. Hill, A. Black, A. C. Da Costa, L. C. Franco, S. P. Silva, C. H. Wu, J. Raghwani, S. Cauchemez, L. Du Plessis, M. P. Verotti, W. K. De Oliveira, E. H. Carmo, G. E. Coelho, A. C. Santelli, L. C. Vinhal, C. M. Henriques, J. T. Simpson, M. Loose, K. G. Andersen, N. D. Grubaugh, S. Somasekar, C. Y. Chiu, J. E. Muñoz-Medina, C. R. Gonzalez-Bonilla, C. F. Arias, L. L. Lewis-Ximenez, S. A. Baylis, A. O. Chieppe, S. F. Aguiar, C. A. Fernandes, P. S. Lemos, B. L. Nascimento, H. A. Monteiro, I. C. Siqueira, M. G. De Queiroz, T. R. De Souza, J. F. Bezerra, M. R. Lemos, G. F. Pereira, D. Loudal, L. C. Moura, R. Dhalia, R. F. França, T. Magalhães, E. T. Marques, T. Jaenisch, G. L. Wallau, M. C. De Lima, V. Nascimento, E. M. De Cerqueira, M. M. De Lima, D. L. Mascarenhas, J. P. Neto, A. S. Levin, T. R. Tozetto-Mendoza, S. N. Fonseca, M. C. Mendes-Correa, F. P. Milagres, A. Segurado, E. C. Holmes, A. Rambaut, T. Bedford, M. R. Nunes, E. C. Sabino, L. C. Alcantara, N. J. Loman, and O. G. Pybus, "Establishment and cryptic transmission of Zika virus in Brazil and the Americas," *Nature*, vol. 546, pp. 406–410, jun 2017.

[137] R. J. Rockett, A. Arnott, C. Lam, R. Sadsad, V. Timms, K. A. Gray, J. S. Eden, S. Chang, M. Gall, J. Draper, E. M. Sim, N. L. Bachmann, I. Carter, K. Basile, R. Byun, M. V. O'Sullivan, S. C. Chen, S. Maddocks, T. C. Sorrell, D. E. Dwyer, E. C. Holmes, J. Kok, M. Prokopenko, and V. Sintchenko, "Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling," *Nature Medicine*, vol. 26, pp. 1398–1404, sep 2020.

[138] B. B. Oude Munnink, D. F. Nieuwenhuijse, M. Stein, Á. O'Toole, M. Haverkate, M. Mollers, S. K. Kamga, C. Schapendonk, M. Pronk, P. Lexmond, A. van der Linden, T. Bestebroer, I. Chestakova, R. J. Overmars, S. van Nieuwkoop, R. Molenkamp, A. A. van der Eijk, C. GeurtsvanKessel, H. Vennema, A. Meijer, A. Rambaut, J. van Dissel, R. S. Sikkema, A. Timen, M. Koopmans, G. J. Oudehuis, J. Schinkel, J. Kluytmans, M. Kluytmans-van den Bergh, W. van den Bijllaardt, R. G. Berntvelsen, M. M. van Rijen, P. Schneeberger, S. Pas, B. M. Diederen, A. M. Bergmans, P. A. van der Eijk, J. Verweij, A. G. Buiting, R. Streefkerk, A. P. Aldenkamp, P. de Man, J. G. Koelemal, D. Ong, S. Paltansing, N. Veassen, J. Sleven, L. Bakker, H. Brockhoff, A. Rietveld, F. Slijkerman Megelink, J. Cohen Stuart, A. de Vries, W. van der Reijden, A. Ros, E. Lodder, E. Verspui-van der Eijk, I. Huijskens, E. M. Kraan, M. P. van der Linden, S. B. Debast, N. A. Naiemi, A. C. Kroes, M. Damen, S. Dinant, S. Lekkerkerk, O. Pontesilli, P. Smit, C. van Tienen, P. C. Godschalk, J. van Pelt, A. Ott, C. van der Weijden, H. Wertheim, J. Rahamat-Langendoen, J. Reimerink, R. Bodewes, E. Duizer, B. van der Veer, C. Reusken, S. Lutgens, P. Schneeberger, M. Hermans, P. Wever, A. Leenders, H. ter Waarbeek, and C. Hoebe, "Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands," *Nature Medicine*, vol. 26, pp. 1405–1410, sep 2020.

[139] L. M. Gonçalves Rossi, A. Escobar-Gutierrez, and P. Rahal, "Advanced molecular surveillance of hepatitis C virus," 2015.

[140] S. Goya, M. Galiano, I. Nauwelaers, A. Trento, P. J. Openshaw, A. S. Mistchenko, M. Zambon, and M. Viegas, "Toward unified molecular surveillance of RSV: A proposal for genotype definition," *Influenza and Other Respiratory Viruses*, vol. 14, pp. 274–285, may 2020.

[141] C. Kuiken, K. Yusim, L. Boykin, and R. Richardson, "The Los Alamos hepatitis C sequence database," *Bioinformatics*, vol. 21, pp. 379–384, feb 2005.

[142] R. A. Neher and T. Bedford, "nextflu: real-time tracking of seasonal influenza virus evolution in humans," *Bioinformatics*, vol. 31, pp. 3546–3548, nov 2015.

[143] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, pp. 4121–4123, dec 2018.

[144] M. Muñoz, L. M. Wintaco, S. A. Muñoz, and J. D. Ramírez, "Dissecting the Heterogeneous Population Genetic Structure of Candida albicans: Limitations and constraints of the multilocus sequence typing scheme," *Frontiers in Microbiology*, vol. 10, no. MAY, 2019.

[145] K. Wu, T. Luo, L. Li, Q. Zhang, J. Zhu, Q. Gao, M. Chen, and M. Zhu, "Multilocus sequence typing of pathogenic Candida albicans isolates collected from a teaching hospital in Shanghai, China: A molecular epidemiology study," *PLoS ONE*, vol. 10, p. e0125245, apr 2015.

[146] X. Hou, M. Xiao, S. C. Chen, F. Kong, H. Wang, Y. Z. Chu, M. Kang, Z. Y. Sun, Z. D. Hu, R. Y. Li, J. Lu, K. Liao, T. S. Hu, Y. X. Ni, G. L. Zou, G. Zhang, X. Fan, Y. P. Zhao, and Y. C. Xu, "Molecular epidemiology and antifungal susceptibility of Candida glabrata in China (August 2009 to July 2014): A multi-center study," *Frontiers in Microbiology*, vol. 8, may 2017.

[147] K. W. Chen, Y. C. Chen, H. J. Lo, F. C. Odds, T. H. Wang, C. Y. Lin, and S. Y. Li, "Multilocus sequence typing for analyses of clonality of Candida albicans strains in Taiwan," *Journal of Clinical Microbiology*, vol. 44, pp. 2172–2178, jun 2006.

[148] A. S, S.-G. M, G. M, and R.-A. M, "Genotyping of Clinical Isolates of Candida Glabrata From Iran by Multilocus Sequence Typing and Determination of Population Structure and Drug Resistance Profile," *Medical mycology*, vol. 56, no. 2, 2018.

[149] A. Tavanti, A. D. Davidson, M. J. Fordyce, N. A. Gow, M. C. Maiden, and F. C. Odds, "Population structure and properties of Candida albicans, as determined by multilocus sequence typing," *Journal of Clinical Microbiology*, vol. 43, pp. 5601–5613, nov 2005.

[150] F. C. Odds and M. D. Jacobsen, "Multilocus sequence typing of pathogenic Candida species," jul 2008.

[151] M. E. Bougnoux, D. Diogo, N. François, B. Sendid, S. Veirmeire, J. F. Colombel, C. Bouchier, H. Van Kruiningen, C. D'Enfert, and D. Poulain, "Multilocus sequence typing reveals intrafamilial transmission and microevolutions of Candida albicans isolates from the human digestive tract," *Journal of Clinical Microbiology*, vol. 44, pp. 1810–1820, may 2006.

[152] A. R. Dodgson, C. Pujol, D. W. Denning, D. R. Soll, and A. J. Fox, "Multilocus Sequence Typing of Candida glabrata Reveals Geographically Enriched Clades," *Journal of Clinical Microbiology*, vol. 41, pp. 5709–5717, dec 2003.

[153] M. E. Bougnoux, A. Tavanti, C. Bouchier, N. A. Gow, A. Magnier, A. D. Davidson, M. C. Maiden, C. D'Enfert, and F. C. Odds, "Collaborative Consensus for Optimized Multilocus Sequence Typing of Candida albicans," *Journal of Clinical Microbiology*, vol. 41, pp. 5265–5266, nov 2003.

[154] N. A. Chow, J. F. Muñoz, L. Gade, E. L. Berkow, X. Li, R. M. Welsh, K. Forsberg, S. R. Lockhart, R. Adam, A. Alanio, A. Alastruey-Izquierdo, S. Althawadi, A. B. Araúz, R. Ben-Ami, A. Bharat, B. Calvo, M. Desnos-Ollivier, P. Escandón, D. Gardam, R. Gunturu, C. H. Heath, O. Kurzai, R. Martin, A. P. Litvintseva, and C. A. Cuomo, "Tracing the Evolutionary History and Global Expansion of Candida auris Using Population Genomic Analyses," *mBio*, vol. 11, apr 2020.

[155] C. Biswas, V. R. Marcelino, S. Van Hal, C. Halliday, E. Martinez, Q. Wang, S. Kidd, K. Kennedy, D. Marriott, C. O. Morrissey, I. Arthur, K. Weeks, M. A. Slavin, T. C. Sorrell, V. Sintchenko, W. Meyer, and S. C. Chen, "Whole Genome Sequencing of Australian Candida glabrata Isolates Reveals Genetic Diversity and Novel Sequence Types," *Frontiers in Microbiology*, vol. 9, mar 2018.

[156] E. Sitterlé, C. Maufrais, N. Sertour, M. Palayret, C. D'Enfert, and M. E. Bougnoux, "Within-Host Genomic Diversity of Candida albicans in Healthy Carriers," *Scientific Reports*, vol. 9, pp. 1–12, dec 2019.

[157] L. Carreté, E. Ksiezopolska, C. Pegueroles, E. Gómez-Molero, E. Saus, S. Iraola-Guzmán, D. Loska, O. Bader, C. Fairhead, and T. Gabaldón, "Patterns of Genomic Variation in the Opportunistic Pathogen Candida glabrata Suggest the Existence of Mating and a Secondary Association with Humans," *Current Biology*, vol. 28, pp. 15–27.e7, jan 2018.

[158] J. Ropars, C. Maufrais, D. Diogo, M. Marcet-Houben, A. Perin, N. Sertour, K. Mosca, E. Permal, G. Laval, C. Bouchier, L. Ma, K. Schwartz, K. Voelz, R. C. May, J. Poulain, C. Battail, P. Wincker, A. M. Borman, A. Chowdhary, S. Fan, S. H. Kim, P. Le Pape, O. Romeo, J. H. Shin, T. Gabaldon, G. Sherlock, M.-E. Bougnoux, and C. D'Enfert, "Gene flow contributes to diversification of the major fungal pathogen Candida albicans," *Nature Communications*, vol. 9, p. 2253, dec 2018.

[159] A. E. Barber, M. Weber, K. Kaerger, J. Linde, H. Gölz, D. Duerschmied, A. Markert, R. Guthke, G. Walther, and O. Kurzai, "Comparative genomics of serial candida glabrata isolates and the rapid acquisition of echinocandin resistance during therapy," *Antimicrobial Agents and Chemotherapy*, vol. 63, feb 2019.

[160] L. Carreté, E. Ksiezopolska, E. Gómez-Molero, A. Angoulvant, O. Bader, C. Fairhead, and T. Gabaldón, "Genome Comparisons of Candida glabrata Serial Clinical Isolates Reveal Patterns of Genetic Variation in Infecting Clonal Populations," *Frontiers in Microbiology*, vol. 10, p. 112, feb 2019.