Published in final edited form as:

*Curr Opin Infect Dis*. 2021 August 01; 34(4): 339–345. doi:10.1097/QCO.0000000000000743.

# Techniques in bacterial strain typing: past, present, and future

**Shelby R. Simar**[a,b], **Blake M. Hanson**[a,b,c], **Cesar A. Arias**[a,b,c]

[a]University of Texas Health Science Center at Houston – School of Public Health, University of Texas Health Science Center, Houston, Texas, USA

[b]Center for Antimicrobial Resistance and Microbial Genomics, University of Texas Health Science Center, Houston, Texas, USA

[c]Division of Infectious Diseases, Department of Internal Medicine, McGovern Medical School, University of Texas Health Science Center, Houston, Texas, USA

## Abstract

**Purpose of review—**The advancement of molecular techniques such as whole-genome sequencing (WGS) has revolutionized the field of bacterial strain typing, with important implications for epidemiological surveillance and outbreak investigations. This review summarizes state-of-the-art techniques in strain typing and examines barriers faced by clinical and public health laboratories in implementing these new methodologies.

**Recent findings—**WGS-based methodologies are on track to become the new ‘gold standards’ in bacterial strain typing, replacing traditional methods like pulsed-field gel electrophoresis and multilocus sequence typing. These new techniques have an improved ability to identify genetic relationships among organisms of interest. Further, advances in long-read sequencing approaches will likely provide a highly discriminatory tool to perform pangenome analyses and characterize relevant accessory genome elements, including mobile genetic elements carrying antibiotic resistance determinants in real time. Barriers to widespread integration of these approaches include a lack of standardized workflows and technical training.

**Summary—**Genomic bacterial strain typing has facilitated a paradigm shift in clinical and molecular epidemiology. The increased resolution that these new techniques provide, along with epidemiological data, will facilitate the rapid identification of transmission routes with high confidence, leading to timely and effective deployment of infection control and public health interventions in outbreak settings.

### Keywords

bacteria; strain typing; surveillance; whole-genome sequencing

Correspondence to Cesar A. Arias, MD, PhD, Center for Antimicrobial Resistance and Microbial Genomics (CARMiG) and Division of Infectious Diseases, UTHealth McGovern Medical School, 6431 Fannin, MSB 1.150, Houston, Texas 77030, USA. Tel: +713 500 6500; cesar.arias@uth.tmc.edu.

Conflicts of interest

C.A.A. has received grant support from Merck, MeMed Diagnostics, and Entasis Therapeutics.

## INTRODUCTION

The increasing interconnectedness of society has greatly influenced the transmissibility and subsequent diversification of bacterial pathogens [1], creating a need for improved methods of bacterial characterization and classification. Bacterial strain typing – the practice of microbial characterization used to discriminate between strains of a bacterial species – is a fundamental aspect of epidemiological surveillance and investigation. Strain typing can characterize and confirm epidemiological linkage in an outbreak setting and provide insights into bacterial population dynamics. However, traditional typing methods often target only a small portion of the bacterial genome, limiting the resolution and, thus, the scope of our understanding of the molecular epidemiology of clinically relevant pathogens.

There are a few traditional methods that remain primary choices for strain typing in many clinical and public health laboratories. The first of these methods is pulsed-field gel electrophoresis (PFGE), which uses alternating electric fields applied at differing angles within an agarose gel to separate large DNA molecules, creating size-dependent banding patterns, or 'fingerprints,' based on restriction enzyme cleavage sites [2]. This method is known as the 'gold standard' for subtyping and, until recently, was the primary method used by the Center for Disease Control and Prevention's PulseNet database for tracking outbreaks of foodborne illnesses [3]. However, PFGE has important limitations, including a need for protocols that are standardized for individual pathogens, extreme sensitivity to the selection of restriction enzymes, a time-consuming and labor-intensive workflow, and relatively low throughput. Thus, many large-scale surveillance efforts have transitioned to the use of WGS-based characterization, such as multilocus sequencing typing (MLST), a sequencing-based method that uses allelic permutations of conserved 'housekeeping genes' loci to create MLST schemes known as 'sequence types' (ST) [4]. Although this method provides unambiguous results and allows for easier inter-lab comparisons through a centralized database (PubMLST; pubmlst.org), it can be cost-prohibitive, each species requires a different typing schema, and it lacks the ability to further discriminate relatedness within STs [4,5].

Strain typing methodologies have recently undergone a paradigm shift as whole-genome sequencing (WGS) has become cheaper and more accessible to clinical and public health laboratories. WGS provides unmatched resolution and discriminatory power for highly related strains, and it has significant potential for outbreak detection, epidemiological surveillance, and infection control strategies.

## NEW TECHNIQUES IN BACTERIAL STRAIN TYPING

The increased resolution provided by new strain typing methodologies has enabled the distinction of bacteria differing at only a few genetic sites, which is a significant advancement from the discriminatory power of traditional strain typing methods (Table 1). The definitions of commonly used terms to classify genetic relatedness among bacterial strains are shown in Table 2. It should be noted that, although these terms are often used interchangeably to define sets of related isolates, this exchangeable use is due to a drift in the terminology used over time amongst scientists studying different pathogens. We have

included the original definitions apart from ‘clonal group,’ which was previously described as an *a priori* defined cluster of bacterial organisms that shared (*n*) alleles of their associated STs [8]. This definition is ambiguous in practice and is often operationalized with the same definition as clonal complex, so we recommend the stricter definition as defined in Table 2.

## Beyond multilocus sequencing typing

The increased use of WGS has enabled the expansion of traditional MLST methods based on 7–8 housekeeping genes to hundreds or thousands of genetic loci, greatly enhancing the precision and discriminatory power of typing and providing relevant clinical and epidemiological information. Here, the two newest expansions of MLST are described in detail.

## Core genome multilocus sequencing typing

This approach is also referred to as the gene-by-gene approach. cgMLST is similar to MLST but utilizes a larger proportion of the genome, defined as the core genome (the set of genes that is found in nearly all strains of a species) to determine genetic relatedness. After WGS, a genome assembly is aligned to a reference-based ‘scheme’ of core genes, and each isolate is characterized based on allelic variations relative to the reference [10■,11■]. Along with delivering higher resolution relative to traditional and MLST typing methods, cgMLST provides the opportunity to investigate organism phylogeny through strategies that include the use of distance-based techniques to create nearest neighbor or minimum-spanning trees [12■,13■]. Since its inception, cgMLST has become a widely used alternative to MLST for those seeking greater resolution through a similar workflow, and typing schemes based on cgMLST have been published for a number of bacterial species, facilitating its use in outbreak investigations [13■,14-16]. In a recent example of cgMLST application, Hansen *et al.* utilized this typing method to identify an outbreak of vancomycin-resistant *Enterococcus faecium* by establishing epidemiological links between patients carrying isolates belonging to the clone ST80-CT993 and distinguishing this clone from similar circulating STs. This analysis identified affected wards, and a targeted infection control intervention was successfully implemented in these areas, saving time and resources with important implications for hospital epidemiology [17■■].

To date, there is no centralized or consistent naming system for cgMLST schemes. In fact, there are three distinct cgMLST schemes available for *Pseudomonas aeruginosa* [14,15,18], which may result in lack of reproducibility in future published data. To create a reliable central database, a large number of high-quality reference genomes would be needed for each species of interest, which is an expensive and computationally intensive undertaking. Most importantly, cgMLST only accounts for the conserved genes within a species and ignores the contribution of the accessory genome—the portion of the genome that varies between strains of a species—to overall intra-clonal diversity [19■].

## Whole-genome multilocus sequencing typing

This approach is an extension of cgMLST that utilizes both the core and accessory genomes (the pangenome), theoretically providing higher resolution than cgMLST for closely related isolates than the cgMLST approach. In a retrospective investigation of listeriosis outbreaks

in small ruminants, wgMLST uncovered a larger breadth of genomic diversity relative to cgMLST, supporting previous findings indicating that wgMLST should be the primary typing method when investigating highly related bacterial groups [20■]. However, several studies have not been able to demonstrate a significant difference in discriminatory power between these two typing schemes [21,22]. A study by Blanc *et al.* even found wgMLST inferior to cgMLST due to homologous recombination of a DNA fragment affecting phylogeny with no epidemiological significance [23■■].

wgMLST shares some limitations with cgMLST since the choice of high-quality references is essential for reliable discriminatory power, and a standardized method of classification is lacking. Additionally, wgMLST requires a higher level of bioinformatic expertise relative to other typing methods, and assembly and alignment of genomic short reads (the output of the most commonly used sequencing platforms) are not robust to permit the reconstruction of complex genomic structures such as mobile genetic elements (MGEs) and long repeat structures [19■,23■■].

## Single Nucleotide Polymorphism-based methods

Although cgMLST and wgMLST-based methods of strain typing are currently being applied as higher-resolution replacements for traditional MLST and PFGE, there is a considerable amount of genomic variability that cannot be accounted for with these methodologies. Indeed, regions in the accessory genome are often not considered with these approaches, resulting in an inability to differentiate closely related outbreak strains for the detection of recent transmission events when paired with traditional epidemiological metadata. The identification of single-nucleotide polymorphisms (SNPs) between bacterial isolates is one of the most commonly used analyses on WGS data and can be performed with or without the use of a reference genome. Here, two methods and applications of SNP calling are discussed.

## Reference-based single-nucleotide polymorphism calling

SNP calling is most often reference-based and involves the alignment (frequently referred to as mapping) of sequenced isolates to a closely related reference genome to detect SNPs and quantify the genetic relatedness between strains [24]. Though MGEs and regions of recombination are generally excluded, over 95% of the genome is accounted for in these analyses [25]. Reference-based SNP calling is particularly useful when a relatively small number of isolates are available for analysis. Indeed, Hoang *et al.* used this strategy to identify region-specific lineages of six *Bacillus anthracis* strains isolated in northern Vietnamese provinces, finding that all strains could be classified into a single lineage that has not been previously reported in Asia [26■■].

Reference-based SNP calling as a typing method is relatively straightforward and can yield highly accurate results, but it has an important limitation; the selection of a reference genome is of paramount importance, and a closed genome that is highly related to the sequences of interest is highly desired [25]. Thus, reference-based SNP calling becomes a difficult task when analyzing nonmodel organisms for which there are no well-established reference genomes. Varied reference choices and SNP calling workflows

can greatly influence the number of SNP differences identified, limiting reproducibility and comparability across studies and laboratories and resulting in incorrect epidemiological inferences [25,27]. For example, a study by Valiente-Mullor *et al.* examined the effect of using different reference genomes for SNP calling and phylogenetic analyses of five bacterial species and found that the choice of the reference strain had an impact on all parameters considered, including SNP calling and phylogenetic tree construction [28■■].

## Reference-agnostic single-nucleotide polymorphism calling

To circumvent the need for an appropriate reference strain, methods based on k-mer comparisons have been developed for WGS data. K-mers are defined blocks of nucleotides of length (k) that can be compared in a pairwise fashion between sets of genomes of interest to model intra-sample diversity and taxonomy [29■]. This approach has been applied to a number of retrospective outbreak investigations, including the first WGS-based characterization of *Bacillus cereus* isolates linked to a foodborne outbreak, where investigators compared a number of reference-based SNP calling tools using a reference-free, k-mer based approach (kSNP3) [30]. This investigation found that kSNP3 produced consistent results that were not affected by the choice of reference genome. These findings also support existing literature that reference-free methods are most reliable in suspected outbreak situations, or where isolates are expected to be relatively similar [31■■]. Another study by Cremers *et al.*, that used this approach to study an outbreak of methicillin-resistant *Staphylococcus aureus* in a neonatal intensive care unit, found that a k-mer-based pairwise SNP analysis substantially outperformed other typing methods, including cgMLST and wgMLST. Maximizing the amount of genetic material utilized for comparison from both the core and accessory genomes produced highly precise insights into potential chains of transmission among neonates [32■■].

## Other considerations

Though SNP-based typing methods arguably deliver the highest discriminatory power of all the methods described thus far, there is still considerable debate among the scientific community regarding SNP thresholds for genetic relatedness (‘clonality’) that impacts the interpretation of outbreak and infection control investigations. Determination of clusters and significance thresholds is often based on substitution and recombination rates along with quantification of SNPs and is therefore not universally applicable to all bacterial species [24]. However, several recent studies have made efforts to define thresholds of genetic relatedness that indicate transmission events in an outbreak setting for several organisms. These include cutoffs of 25 whole-genome/15 core genome SNPs for methicillin-resistant *S. aureus* [33■], and 6 and 21 core genome SNPs for *E. faecium* and carbapenem-resistant *Klebsiella pneumoniae*, respectively [34■,35■]. There is a need for further studies on this subject, as these cut-offs may be dynamic and dependent on host and environmental factors [36■■]. Lastly, the importance of epidemiological evidence and context should be taken into consideration, as genomic data alone is not sufficient for determination of outbreak transmission dynamics with full confidence [20■,36■■].

## FUTURE DIRECTIONS

Defining the core, or conserved, regions of bacterial genomes is an important first step in most strain typing and phylogenetic analyses. However, the bacterial 'mobilome' (the repertoire of acquired MGEs) is a primary driver of adaptive evolution through horizontal gene transfer and a major determinant of bacterial resistance and virulence phenotypes [37■]. Indeed, the mobilome is generally disregarded in most strain typing methodologies, as MGE structures are difficult to reconstruct with widely used WGS methods like short-read sequencing, and it is difficult to distinguish between transient gene acquisition and stable assimilation into genetic lineages [38]. Thus, more extensive research is critical to enable proper consideration of the role of the mobilome in the context of genomic diversification and its impacts on strain typing and outbreak investigations.

The advent of long-read sequencing technologies, such as those from Oxford Nanopore Technologies (ONT) and Pacific Biosciences, allow for accurate reconstruction of complicated MGEs—particularly plasmids harboring AMR and virulence determinants—due to their ability to generate sequencing reads that span the entirety of these complex genetic structures. However, the high error rates often associated with these technologies pose a challenge for accurate genomic analysis, particularly SNP calling. To overcome the limitations of both short- and long-read sequencing, hybrid assemblies may be created using highly accurate short-read data to 'polish' the less-accurate long-read sequences, generating closed, reference-quality genomes [39]. Neal-McKinney *et al.* compared Illumina short-read and Pacific Biosciences long-read sequencing alone to hybrid assemblies generated by this technique in *Campylobacter jejuni* and found that the latter created the most contiguous assemblies and was the superior method for SNP typing and definitive isolate characterization [40■]. This technique was also used by Prussing *et al.* to identify the potential transfer of a plasmid harboring $bla_{\text{KPC-2}}$ across bacterial species in epidemiologically linked patients [41■].

Despite these advances in sequencing technology and strain typing methodologies, most WGS-based outbreak investigations and surveillance efforts are still performed retrospectively, limiting the impact these methods can have on clinical decision-making and infection control interventions at the time they are most needed. Thus, there is a critical need to place more emphasis on developing tools and workflows for real-time sequencing and data analysis. Currently, there is only one methodology available for such applications—the long-read nanopore sequencing platform developed by ONT (www.nanoporetech.com). Since the release of its first sequencer, the MinION, in 2014, ONT sequencing platforms have been increasingly utilized in environments ranging from small-scale research studies that have uncovered new classes of antibiotic resistance plasmids [42■] to the implication of contaminated detergent as the source of an extended-spectrum beta-lactamase- producing *Klebsiella michiganensis* outbreak in an Australian neonatal unit [43■■]. However, while this technology has enormous potential for advancing the fields of real-time bacterial identification, strain typing, and outbreak and surveillance efforts, there is much work to be done to optimize and standardize long-read sequencing library preparation and analysis workflows before this technology can be scaled to larger datasets.

## IMPLEMENTATION IN THE CLINICAL AND PUBLIC HEALTH LABORATORY

WGS strain-typing workflows are increasingly being adopted by clinical and public health laboratories as these technologies become more accessible and cost-effective. Sequencing the entire genome of an infecting/colonizing organism provides not only unparalleled discriminatory power for highly related lineages, but also delivers insight into drug susceptibility and virulence potential, which would otherwise require a combination of laboratory methods and tools. Timely accessibility of this breadth of information is crucial for effective outbreak management and infection control efforts. Yet there remain barriers to widespread integration of WGS-based bacterial typing into clinical and public health laboratory workflows.

### Standardization

A significant barrier to implementation of WGS methodologies is the lack of standardized workflows. Protocols and analysis methods (from the quality of DNA extracted to the choice of SNP calling pipeline) vary considerably between laboratories, resulting in differing interpretations, quality control issues, and decreased reproducibility [44■,45■]. In 2017, a Swiss trial of nine laboratories aimed at fostering harmonization of WGS-based bacterial strain typing found that, whereas MLST typing, phylogenetic tree construction, and cluster identification were relatively harmonious across laboratories, differing interpretations of sequencing data based on SNP counts led to diverse inferences regarding strain relatedness during outbreak investigations, highlighting the need for standardized definitions and interpretation criteria to reach reproducible conclusions [12■]. In order for WGS-based methods to become the new standard in strain typing, there must be full confidence in the accuracy and robustness of the data generated across different sequencing platforms and laboratories. Every step of the WGS workflow—sample preparation, sequencing, and downstream analysis and interpretation—needs to be standardized and validated with a variety of bacterial species against current ‘gold standard’ typing methods. Furthermore, analysis tools and pipelines must be version-controlled, and parameters used for each workflow must be standardized and validated. This is not a trivial task, as sequencing technologies and data analysis methods are constantly changing. It may be helpful to look to human genetics for insight, as this field has made considerable progress in the creation of well-established references and tools for widescale laboratory use [46].

### Analysis training and expertise

Another barrier to integration of WGS in many laboratories is the absence of bioinformatics expertise needed to analyze WGS data. As bioinformatic analysis approaches are not commonly utilized in most diagnostic or public health laboratories, emphasis must be placed on developing tools that are user friendly, otherwise, laboratories would need to hire bioinformaticians to aid in interpretation of data. Lastly, there remains a critical need to train the next generation of clinical microbiologists in WGS and bioinformatics practices to meet these needs and further the advancement of WGS analysis tools.

## CONCLUSION

The increasing accessibility and cost-effectiveness of WGS have catalyzed the innovation of new, higher-resolution bacterial strain typing methods that are likely to replace traditional typing methods as the new ‘gold standard’ in the coming years. However, significant work will need to be done regarding standardization of sequencing and analysis workflows, personnel training, and increasing cost-effectiveness before such methodologies can be widely implemented in clinical and public health laboratories.

## Financial support and sponsorship

S.R.S. was partially funded under an NIH predoctoral T32 training grant (5T32AI055449-15 to Theresa M. Koehler). B.M.H. was partially funded by a National Institute of Allergy and Infectious Disease (NIAID) of the National Institutes of Health under Award Number K01AI148593. CAA was partially funded by the NIH/NIAID Award Numbers K24AI121296, R01AI134637, R01AI148342-01, R21AI143229, and P01AI152999-01.

## REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

■ of special interest

■■ of outstanding interest

1. Berndtson AE. Increasing globalization and the movement of antimicrobial resistance between countries. Surg Infect 2020; 21:579–585.
2. Herschleb J, Ananiev G, Schwartz DC. Pulsed-field gel electrophoresis. Nat Protoc 2007; 2:677–684. [PubMed: 17406630]
3. National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Foodborne, Waterborne, and Environmental Diseases (DFWED). Pulsed-field Gel Electrophoresis (PFGE) [Internet]. Centers for Disease Control and Prevention [updated 2016]. Available from: https://www.cdc.gov/pulsenet/pathogens/pfge.html. [Accessed 25 February 2021]
4. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 2013; 11:728–736. [PubMed: 23979428]
5. Kovanen SM, Kivistö RI, Rossi M, et al. Multilocus sequence typing (MLST) and whole-genome MLST of campylobacter jejuni isolates from human infections in three districts during a Seasonal Peak in Finland. J Clin Microbiol 2014; 52:4147–4154. [PubMed: 25232158]
6. Baum D Phylogenetic trees and monophyletic groups. Nat Educ 2008; 1:190.
7. Spratt BG. Exploring the concept of clonality in bacteria. Methods Mol Biol 2004; 266:323–352. [PubMed: 15148426]
8. Feil EJ, Li BC, Aanensen DM, et al. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J Bacteriol 2004; 186:1518–1530. [PubMed: 14973027]
9. Dijkshoorn L, Ursing BM, Ursing JB. Strain, clone and species: comments on three basic concepts of bacteriology. J Med Microbiol 2000; 49: 397–401. [PubMed: 10798550]
10 ■. Uelze L, Grützke J, Borowiak M, et al. Typing methods based on whole genome sequencing data. One Health Outlook 2020; 2:3. [PubMed: 33829127] An interesting and thorough review of the tools and techniques used for WGS-based bacterial strain typing.
11 ■. Liang KYH, Orata FD, Islam MT, et al. A vibrio cholerae core genome multilocus sequence typing scheme to facilitate the epidemiological study of cholera. J Bacteriol 2020; 202:e00086–20. [PubMed: 32540931] This study was the first to propose a cgMLST typing scheme for V. cholerae.

12 ■. Dylus D, Pillonel T, Opota O, et al. NGS-based *S. aureus* typing and outbreak analysis in clinical microbiology laboratories: lessons learned from a swiss-wide proficiency test. Front Microbiol 2020; 11:591093. [PubMed: 33424794] This pilot study demonstrated the strengths and weaknesses associated with wide-scale implementation of WGS-based bacterial strain typing methods in the clinical microbiology laboratory. This trial provides a blueprint for implementation and quality assessment of WGS workflows in similar lab settings.

13 ■. Liu S, Li X, Guo Z, et al. A core genome multilocus sequence typing scheme for *Streptococcus mutans*. mSphere 2020; 5:e00348–20. [PubMed: 32641425] This study was the first to develop a cgMLST typing scheme for S. mutans.

14. Tönnies H, Prior K, Harmsen D, Mellmann A. Establishment and evaluation of a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Pseudomonas aeruginosa*. J Clin Microbiol 2021; 59:e01987–20. [PubMed: 33328175]

15. de Sales RO, Migliorini LB, Puga R, et al. A core genome multilocus sequence typing scheme for *Pseudomonas aeruginosa*. Front Microbiol 2020; 11:1049. [PubMed: 32528447]

16. Hsu C-H, Harrison L, Mukherjee S, et al. Core genome multilocus sequence typing for food animal source attribution of human campylobacter jejuni infections. Pathog Basel Switz 2020; 9:532.

17 ■■. Hansen SK, Andersen L, Detlefsen M, et al. Using core genome MLST typing for vancomycin-resistant *Enterococcus faecium* isolates to guide infection control interventions and end an outbreak. J Glob Antimicrob Resist 2021; 24:418–423. [PubMed: 33618041] This is an interesting study that used cgMLST to identify and end an outbreak of VRE. This is a motivating example of the use of WGS-based strain typing to guide and streamline infection control response to an outbreak.

18. Stanton RA, McAllister G, Daniels JB, et al. Development and application of a core genome multilocus sequence typing scheme for the healthcare-associated pathogen *Pseudomonas aeruginosa*. J Clin Microbiol 2020; 58:.

19 ■. Tümmler B Molecular epidemiology in current times. Environ Microbiol 2020; 22:4909–18. [PubMed: 32945108] This review provides numerous examples of applications of WGS-based bacterial strain typing methodologies.

20 ■. Papić B, Kušar D, Zdovc I, et al. Retrospective investigation of listeriosis outbreaks in small ruminants using different analytical approaches for whole genome sequencing-based typing of Listeria monocytogenes. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis 2020; 77:104047. This study shows that WGS-based methods of typing have superior discriminatory power compared to traditional typing methods when applied to highly related groups of bacterial organisms.

21. Miro E, Rossen JWA, Chlebowicz MA, et al. Core/whole genome multilocus sequence typing and core genome SNP-based typing of OXA-48-producing *Klebsiella pneumoniae* Clinical Isolates From Spain. Front Microbiol 2019; 10:2961. [PubMed: 32082262]

22. Henri C, Leekitcharoenphon P, Carleton HA, et al. An assessment of different genomic approaches for inferring phylogeny of listeria monocytogenes. Front Microbiol 2017; 8:2351. [PubMed: 29238330]

23 ■■. Blanc DS, Magalhães B, Koenig I, et al. Comparison of whole genome (wg-) and core genome (cg-) MLST (BioNumericsTM) versus SNP variant calling for epidemiological investigation of *Pseudomonas aeruginosa*. Front Microbiol 2020; 11:1729. [PubMed: 32793169] This is an interesting study that shows the impact of genomic recombination on WGS-based typing schemes in P. aeruginosa. This may have important implications for future interpretations of results from this technology.

24. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 2011; 12:443–51. [PubMed: 21587300]

25. Jagadeesan B, Gerner-Smidt P, Allard MW, et al. The use of next generation sequencing for improving food safety: translation into practice. Food Microbiol 2019; 79:96–115. [PubMed: 30621881]

26 ■■. Hoang TTH, Dang DA, Pham TH, et al. Epidemiological and comparative genomic analysis of *Bacillus anthracis* isolated from northern Vietnam. PLoS One 2020; 15:e0228116. [PubMed: 32084143] An interesting application of reference-based SNP calling to understand the genetic

and epidemiologic background of Vietnamese B. anthracis strains that had not previously been investigated.

27. Besser JM, Carleton HA, Trees E, et al. Interpretation of whole-genome sequencing for enteric disease surveillance and outbreak investigation. Foodborne Pathog Dis 2019; 16:504–12. [PubMed: 31246502]

28 ■■. Valiente-Mullor C, Beamud B, Ansari I, et al. One is not enough: on the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol 2021; 17:e1008678. [PubMed: 33503026] This study demonstrates the impact that the choice of bacterial reference genome can have on WGS-based strain typing analysis and interpretation.

29 ■. Anyansi C, Straub TJ, Manson AL, et al. Computational methods for strain-level microbial detection in colony and metagenome sequencing data. Front Microbiol 2020; 11:1925. [PubMed: 33013732] A thorough and clear overview of k-mer based strain typing methods.

30. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 2015; 31:2877–2878. [PubMed: 25913206]

31 ■■. Carroll LM, Wiedmann M, Mukherjee M, et al. Characterization of emetic and Diarrheal Bacillus cereus strains from a 2016 foodborne outbreak using whole-genome sequencing: addressing the microbiological, epidemiological, and bioinformatic challenges. Front Microbiol 2019; 10:144. [PubMed: 30809204] This is the first study to use reference-agnostic k-mer based strain typing to characterize B. cereus isolates linked to a foodborne outbreak.

32 ■■. Cremers AJH, Coolen JPM, Bleeker-Rovers CP, et al. Surveillance-embedded genomic outbreak resolution of methicillin-susceptible *Staphylococcus aureus* in a neonatal intensive care unit. Sci Rep 2020; 10:2619. [PubMed: 32060342] An interesting study that used WGS-based strain typing to pinpoint the sources of a MSSA outbreak in a NICU that could not be resolved with traditional strain typing methods.

33 ■. Coll F, Raven KE, Knight GM, et al. Definition of a genetic relatedness cutoff to exclude recent transmission of meticillin-resistant *Staphylococcus aureus*: a genomic epidemiology analysis. Lancet Microbe 2020; 1:e328–35. [PubMed: 33313577] This study provides a genetic relatedness cutoff for MRSA based on core and whole genome SNPs to define recent transmission events.

34 ■. Gouliouris T, Coll F, Ludden C, et al. Quantifying acquisition and transmission of *Enterococcus faecium* using genomic surveillance. Nat Microbiol 2021; 6:103–11. [PubMed: 33106672] This study proposes a genetic relatedness cutoff for *E. faecium* based on core and whole genome SNPs to define recent transmission events.

35 ■. David S, Reuter S, Harris SR, et al. Epidemic of carbapenem-resistant Klebsiella pneumoniae in Europe is driven by nosocomial spread. Nat Microbiol 2019; 4:1919–1929. [PubMed: 31358985] This study provides a SNP cutoff based on core genome SNPs in K. pneumoniae to discriminate between hospital clusters and identify transmission events.

36 ■■. Jia H, Chen Y, Wang J, et al. Emerging challenges of whole-genome-sequencing–powered epidemiological surveillance of globally distributed clonal groups of bacterial infections, giving Acinetobacter baumannii ST195 as an example. Int J Med Microbiol 2019; 309:151339. [PubMed: 31451388] This interesting study uses A. baumannii to explain the limitations of WGS-based strain typing and provides important considerations for future use of these techniques.

37 ■. Carr VR, Shkoporov A, Hill C, et al. Probing the mobilome: discoveries in the dynamic microbiome. Trends Microbiol 2021; 29:158–70. [PubMed: 32448763] A thorough review that defines various mobile genetic elements and bioinformatics tools used to identify them.

38. Brockhurst MA, Harrison E, Hall JPJ,R, et al. The ecology and evolution of pangenomes. Curr Biol 2019; 29:R1094–103. [PubMed: 31639358]

39. Chen Z, Erickson DL, Meng J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. BMC Genom 2020; 21:631.

40 ■. Neal-McKinney JM, Liu KC, Lock CM, et al. Comparison of MiSeq, MinION, and hybrid genome sequencing for analysis of Campylobacter jejuni. Sci Rep 2021; 11:5676. [PubMed: 33707610] This study demonstrates the high accuracy and resolution of hybrid assembly using long- and short-read sequencing data.

41 ■. Prussing C, Snavely EA, Singh N, et al. Nanopore MinION sequencing reveals possible transfer of blaKPC-2 plasmid across bacterial species in two healthcare facilities. Front Microbiol 2020; 11:2007. [PubMed: 32973725] This investigation used short- and long-read sequencing to identify the possible transfer of a multidrug-resistant plasmid across bacterial species in epidemiologically linked patients.

42 ■. Liu H, Moran RA, Chen Y, et al. Transferable Acinetobacter baumannii plasmid pDETAB2 encodes OXA-58 and NDM-1 and represents a new class of antibiotic resistance plasmids. J Antimicrob Chemother 2021; 76:1130–1134. [PubMed: 33501980] This study discovered a novel MDR plasmid in a rare A. baumannii lineage using long- and short-read sequencing.

43 ■■. Chapman P, Forde BM, Roberts LW, et al. Genomic investigation reveals contaminated detergent as the source of an extended-spectrum-β-lactamase-producing Klebsiella michiganensis outbreak in a neonatal unit. J Clin Microbiol 2020; 58:e01980–e01919. [PubMed: 32102855] This investigation is a motivating example of the use of long-read sequencing and SNP-based strain typing to implicate K. michiganensis from contaminated detergent as the cause of a NICU outbreak in Queensland.

44 ■. Nouws S, Bogaerts B, Verhaegen B, et al. Impact of DNA extraction on whole genome sequencing analysis for characterization and relatedness of Shiga toxin-producing *Escherichia coli* isolates. Sci Rep 2020; 10:14649. [PubMed: 32887913] This study demonstrates the importance of presequencing isolate handling on downstream plasmid reconstruction in Shiga toxin-producing E. coli.

45 ■. Bush SJ, Foster D, Eyre DW, et al. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism–calling pipelines. GigaScience 2020; 9:giaa007. [PubMed: 32025702] This study demonstrates the impact of reference choice on the accuracy of bacterial SNP calling.

46. Marshall CR, Chowdhury S, Taft RJ, et al. Best practices for the analytical validation of clinical whole-genome sequencing intended for the diagnosis of germline disease. Npj Genom Med 2020; 5:1–12. [PubMed: 31969989]

## KEY POINTS

- Whole-genome sequencing has enabled a paradigm shift in bacterial strain typing methodologies.
- Single nucleotide polymorphism (SNP) calling provides the highest discriminatory power relative to other WGS-based typing techniques but is subject to important limitations that include the lack of standardization in thresholds to define relatedness in bacterial species.
- There remain important barriers to wide-scale implementation of WGS-based strain typing methodologies in clinical and microbiological labs – namely, an absence of harmonized workflows and appropriate analytic training.

**Table 1.**

Features of molecular strain typing methods for bacterial organisms

| Method | Type of markers used for differentiation | Discriminatory power | Reproducibility | Bioinformatic knowledge needed | Cost |
|---|---|---|---|---|---|
| Pulsed-field gel electrophoresis (PFGE) | Number of bands depending on restriction enzyme | • | •• | • | •• |
| Multilocus sequence typing (MLST) | 7–8 housekeeping genes | •• | ••••[a] | •• | •• |
| Core genome MLST (cgMLST) | Hundreds to thousands of core genes | ••• | ••••[a] | ••• | ••• |
| Whole genome MLST (wgMLST) | Hundreds to thousands of core plus accessory genes | ••• | •••• | ••• | ••• |
| Reference-based single nucleotide polymorphism (SNP) calling | Depends on organism of interest plus reference choice | •••• | ••• | •••• | •••• |
| Reference-agnostic/k-mer based SNP calling | Depends on organism of interest | •••• | •••• | •••• | •••• |

• low, •• medium, ••• high, •••• very high.

[a]Generally high, but depends on organism of interest and chosen reference.

**Table 2.**

Definitions of terms commonly used to classify genetic relatedness among bacterial strains

| Term | Definition |
|---|---|
| Clade | A group of organisms that contains a single ancestor and its descendants; a monophyletic group [6] |
| Clade | A group of isolates that are genetically indistinguishable [though not necessarily identical] based on a particular molecular typing method and are presumed to be descendants of a common ancestor [7] |
| Sequence type (ST) | Organisms that possess identical allelic profiles of fragments of predetermined housekeeping genes [4] |
| Clonal group | All isolates that belong to a particular ST [8,9] |
| Clonal complex | A cluster of bacterial organisms that originate from a common ancestor and generally share at least 6/7 alleles of their associated ST with another member of the group [8] |
| Strain | Isolate(s) that are distinct from other isolates of the same genus and species based on phenotypic and/or genotypic features [9] |