Published in final edited form as:
Nat Rev Microbiol. 2013 October ; 11(10): 728–736. doi:10.1038/nrmicro3093.

MLST revisited: the gene-by-gene approach to bacterial genomics

Martin C. J. Maiden, **Melissa J. Jansen van Rensburg**, **James E. Bray**, **Sarah G. Earle**, **Suzanne A. Ford**, **Keith A. Jolley**, and **Noel D. McCarthy**
Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

Abstract

Multilocus sequence typing (MLST) was proposed in 1998 as a portable sequence-based method for identifying clonal relationships among bacteria. Today, in the whole-genome era of microbiology, the need for systematic, standardized descriptions of bacterial genotypic variation remains a priority. Here, to meet this need, we draw on the successes of MLST and 16S rRNA gene sequencing to propose a hierarchical gene-by-gene approach that reflects functional and evolutionary relationships and catalogues bacteria ‘from domain to strain’. Our gene-based typing approach using online platforms such as the Bacterial Isolate Genome Sequence Database (BIGSdb) allows the scalable organization and analysis of whole-genome sequence data.

Advances in nucleotide-sequencing technology have provided unparalleled access to the enormous genetic diversity that has accumulated in the bacterial domain during 3.5–4 billion years of evolution[1]. Numerous sets of whole-genome sequencing (WGS) data for bacterial isolates (BOX 1) are available[2], and metagenomic studies using these technologies continue to reveal further, seemingly boundless, diversity in bacterial communities[3]. Faced with this plethora of information, microbiologists must develop structured means of describing this diversity and of linking phenotype and genotype, thereby facilitating an improved understanding of the microbiological world. Given that we have precise information on the function of only a very small proportion of bacterial genes, and no knowledge at all about most, this is a formidable, if extremely exciting, challenge.

Here, we focus primarily on pathogenic bacteria, although the concepts discussed are applicable more widely to all bacteria and archaea. Bacterial pathogens played a crucial part in the development of experimental microbiology and remain the most intensively studied prokaryotes more than 100 years later[4]. Pathogens have emerged across the diversity of the bacterial — but, interestingly, not the archaeal — domain on many occasions and are both polyphyletic and highly diverse. Thus, although pathogens represent only a tiny subset of the bacterial world, the challenges faced by the clinical microbiology laboratory are representative of those faced by microbiology as a whole.

Taxonomic and functional analyses are based on the observations that diversity among bacteria is not continuous and that distinct, stable types with particular properties exist[5]. These founding principles of microbiology[6] have been upheld by much subsequent research, but the study of such clusters remains largely descriptive, and the evolutionary mechanisms that led to cluster emergence and persistence remain incompletely understood[7,8]. Structuring is also evident within bacterial genomes, as diversity is unevenly distributed among genes

*Correspondence to M.C.J.M.*martin.maiden@zoo.ox.ac.uk.

Competing interests statement: The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

and genomic regions[9]. Further, many bacterial genomes are flexible, containing both 'core' and 'accessory' components[10] (BOX 1). Sequence variation in the core genome represents an important starting point if different organisms are to be compared, but all types of variation, including that in genes, intergenic regions and episomes, must be catalogued effectively to make comprehensive associations between phenotype and genotype.

Pre-WGS cataloguing of diversity

A major advance in defining bacterial diversity was the proposal, by the late Carl Woese and colleagues, of a universal and 'natural' — that is, genealogical — classification system based on small-subunit 16S rRNA gene sequences[11]. The variation in these core genes, which are present in all bacteria, has been extensively exploited for the investigation of taxonomic relationships[12], including defining and identifying species[13,14], and forms the basis of many metagenomic community studies[15]. For these applications, 16S rRNA gene sequences provide an inordinate quantity of information in relation to their small size, and they were seminal in establishing that 'prokaryotes' are not a monophyletic group, but contain two of the three domains of life: Bacteria and Archaea[12]. Despite these successes, 16S rRNA sequences provide only limited resolution among closely related bacteria: many bacterial species exhibit identical 16S rRNA sequences among diverse isolates, and even species groups within genera are often poorly resolved, if at all.

The need for higher-resolution characterization of isolates has led to the development of a wide range of strain-typing methods[16], including multilocus sequence typing (MLST)[17], which has become the method of choice for typing many organisms[18]. MLST was designed to accommodate the conflicting signals of vertical and horizontal genetic transfer (BOX 1) that are present in bacterial populations[19] by examining the genome at multiple 'housekeeping' gene loci[17,20]. This concept had been used previously by multilocus enzyme electrophoresis (MLEE)[21], which examines metabolic enzyme variation, but the indexing of gene sequence variation by MLST dramatically improved resolution, reproducibility and portability. The development of curated online MLST reference data sets (such as those found in the PubMLST database collection) provided both portable nomenclature schemes and the possibility of analysing the sequences to infer evolutionary relationships[17,18].

Like MLEE, MLST uses alleles as the unit of comparison, rather than nucleotide sequences. In allele-based comparisons among isolates, each allelic change is counted as a single genetic event, regardless of the number of nucleotide polymorphisms involved. This provides a simple and effective correction for the fact that in many bacteria, common horizontal genetic transfer events account for many more polymorphisms among specimens than rarer point mutations[22]. The MLST approach retains information at all loci and avoids the need to categorize which changes are recent point mutations and which are due to recombination. As MLST schemes record the sequences of allelic variants, MLST data can also be used for sequence-based analyses when this is appropriate[18,20,23].

In most MLST schemes, seven 'MLST loci' are indexed, for which each unique sequence for each locus is assigned an arbitrary and unique allele number. The designations for each of the loci are incorporated into an allelic profile (for example, 2-3-4-3-8-4-6), or a sequence type (ST), which is also assigned a numerical designation (for example, ST11). The ST and allelic designations are related to their respective allelic profiles and sequences in the MLST databases[20], and each ST thus summarizes thousands of base pairs of information. In those bacteria that were most extensively examined by MLST, many hundreds of alleles at each locus and thousands of STs (see the PubMLST database collection) have been identified. Although an ST represents only a tiny percentage of the 'conserved' parts of the genome in question, the large number of STs in many bacterial populations demonstrates the

importance of an expandable means of summarizing and comparing data. The allele and ST designations can be used in definitions of strains or grouped into clonal complexes or lineages (BOX 1) as an improved understanding of the biological population structure emerges[18,20].

Unfortunately, the variability of house-keeping genes among different bacteria makes it impossible to develop MLST schemes for anything but closely related bacteria. Consequently, even within genera (for example, in the genus *Streptococcus*[24-27]), it is necessary to have more than one MLST scheme, targeting different loci. In addition, MLST does not provide sufficient discrimination for all typing purposes, including resolving differences among variants of single-clone, low diversity, asexual pathogens such as *Bacillus anthracis*[28] and *Yersinia pestis*[29] or isolates of more diverse pathogens that belong to the same lineage. For example, at least two distinct sublineages within the ST11 clonal complex of *Neisseria meningitidis* are indistinguishable by MLST[30]. Therefore, MLST alone is not always sufficient for applications such as contact tracing in epidemics or for characterizing single-clone pathogens[31], and in these cases MLST can be supplemented with additional typing schemes that index more variable loci, such as antigen genes[32,33] or variable-number tandem repeats (VNTRs)[34]. To conclude, before the advent of WGS data, multiple approaches were needed to address the range of isolate identification and typing requirements, many of which were specific for particular organisms.

Post-WGS cataloguing of diversity

The complete closed genome sequence of a bacterium, be it from an isolate or a sequence reconstructed from a metagenomic study, is the ultimate ‘molecular bar-code’ for typing and taxonomic purposes. Rapid, very high-throughput sequencing methods[35] are removing the practical constraints that framed the design of previously used approaches, as it is now possible to generate accurate WGS data for bacterial isolates in a single experiment[36]. Current WGS technologies do not generate complete genome sequences; rather, they produce data that can be used to reconstruct most of the genome sequence[37]. The relative performances of different sequencing platforms have been recently reviewed[38,39] and are not further discussed here, but it is reasonable to suppose that the already very high data quality will continue to improve and that large numbers of very high-quality complete genomes are not too distant a prospect.

Although WGS data are easily and efficiently stored as strings of several million As, Cs, Gs and Ts, systems are required that can summarize and organize the variation present in large numbers of genome sequences within a practical framework. Faced with the wealth of data becoming available, novel systems for the description of genetic diversity can be designed from first principles, and a number of criteria required to guide this enterprise are evident from the successes of rRNA sequencing and MLST. Such systems should be:

- Universal, in that they are applicable to all bacteria
- Natural, reflecting genealogical relationships while retaining the capacity to describe closely related organisms with distinct properties
- Understandable, so that the output and the process by which the system has been arrived at are transparent, easily interpreted and reproducible, and where possible the system should be backwards compatible with previous approaches
- Expandable, to account for the incompleteness of our knowledge of diversity, and flexible enough to accommodate changes in this knowledge

- Portable, because methods need to be easily carried out in any laboratory and the data need to be freely exchanged by the use of generic methodologies, reagents and bioinformatics pipelines
- Technology independent, so that the data used are independent of the means of their collection (this means that schemes adopted now need to retain their validity as data improve)
- Readily available to the entire community
- Scalable, so that methods are sufficiently fast and inexpensive to be useable in real time for large or small numbers of isolates (this scalability is especially important for clinical applications and large-scale bacterial population analyses)
- Able to accommodate a wide range of variation so that they can encompass both close and distant genealogical relationships
- Broadly accepted by those who use them and open to contributions from members of the community.

These apparently straightforward criteria are yet to be met by a broadly adopted approach, and the successes and failures of currently implemented approaches to cataloguing bacterial diversity can be measured against them.

A variety of means have been used to identify WGS variation among bacterial specimens. A popular method has been to construct phylogenies on the basis of SNPs (single-nucleotide differences among samples) that have been identified either by the mapping of short-read sequence data to a reference genome, or by aligning *de novo*-assembled sequences to a reference genome[37,40]. This approach has been used successfully to investigate the epidemiology and evolution of a number of single-clone pathogens or members of the same lineage[41-52], but it is limited by its requirement for a reference sequence or whole-genome alignment[40]. The analysis of diverse or highly recombining organisms in this way will prove challenging because the number of total polymorphisms increases as the number of polymorphisms conforming to a clonal model of descent decreases (BOX 1). Alternatively, methods based on sequence similarity — for example, the use of ‘coloured’ de Bruijn graphs to detect genetic variation — eliminate the need for a reference genome or alignment[53]; however, such approaches are yet to be widely used.

Arguably the most versatile means of examining sequence variation among sets of WGS data, and the most natural to micro biologists, is the gene-by-gene (or ‘MLST-like’) approach to *de novo*-assembled genomes, which can be applied to diverse specimens without the need for high-quality reference genomes[54-56]. Most bacterial genomes are dominated by single-copy functional genes, the majority of which are protein coding[57], whereas pseudogenes, paralogues and intergenic regions represent only a small percentage of the genome[9]. Functional protein-coding genes have numerous advantages as a basis for isolate characterization, as illustrated by the success of MLST, and gene-based typing methods are both easily understood and compatible with previous schemes. Although chromosomal protein-coding genes are a useful unit of analysis, other genomic regions such as pseudogenes and intergenic regions can also be considered as typing loci and indexed in the same way. The basis of the gene-by-gene approach is a *de novo* assembly followed by annotation, and various assembly algorithms are available for this process. The continuing improvements in sequencing and assembly technologies[58], in combination with the efforts of initiatives such as the Genomics Standards Consortium[59], mean that *de novo* assemblies which contain most, if not all, of the genome of a given organism are likely to become the norm in the very near future[60].

An important consideration in the design of typing schemes is the level of resolution required among specimens. This, in turn, depends on the particular question being addressed, as some questions require more discrimination among specimens than others. To use the clinical setting as an example: very high resolution is necessary for the detection of outbreaks and the investigation of within-patient variation[43]; lower resolution is required to determine the membership of a particular clonal complex or lineage[61]; and even lower resolution is sufficient for determining the species causing an infection[37,62,63]. The gene-by-gene approach is inherently hierarchical and scalable, as fewer genes can be used for lower-resolution typing, whereas higher levels of resolution can be attained by increasing the number of genes included in the analysis. However, to establish such a system, it is necessary to be able to store, organize and access genome sequence data, and for this enterprise, databases are essential.

Gene-by-gene typing infrastructure

The need for effective data repositories is well recognized, as is evident from the interest generated by initiatives such as the Global Microbial Identifier[63-65], and the power of such infrastructure is illustrated by the success of the 16S rRNA sequence and MLST databases[20,62]. For example, there are many MLST databases available on a number of websites (see PubMLST, the MLST homepage, the MLST databases at the ERI, UCC (Environmental Research Institute, University College Cork, Ireland) and the Institut Pasteur MLST databases), and these databases enable data generated in different laboratories to be efficiently compared. The first MLST database software to be developed[66-68] stored only isolate information, and allele and ST definitions. Newer platforms such as the Bacterial Isolate Genome Sequence Database (BIGSdb)[69] extend the MLST, or gene-by-gene, approach to WGS data. BIGSdb encompasses all the functionality of the software formerly used for the MLST databases hosted on PubMLST[67,70]. It is open source and runs on the Linux operating system using standard PC hardware and, to our knowledge, is currently the only internet server platform to offer the gene-by-gene approach to genome analysis.

BIGSdb links any type of contiguous sequence data, from single genes to complete closed genomes, with provenance and phenotype data (metadata) for the isolates from which the sequences were derived; it also stores allele and locus definitions. There is no inherent limit on the number of isolate, allele or locus records, and loci can be grouped into an unlimited number of schemes (groups of loci that are analysed together, such as the MLST loci). As all this information is text based, very large data sets can be stored on modest computer equipment. BIGSdb has three main components: an isolate provenance and genome database, a sequence definition database, and typing and analysis schemes (FIG. 1). The isolate database stores assembled contiguous sequences (contigs) for each isolate, along with the designations of each locus that have been defined in the database for that isolate. The sequence definition database contains all known alleles for each locus that has been defined, and these alleles are usually identified by integers. As new variants are identified, this expands to index the diversity of new alleles for both known loci and newly identified genes. Although loci are usually genes (especially protein-coding genes), any sequence string, nucleotide or peptide-coding region can be defined as a locus, so that intergenic regions, for example, can be included in this approach.

The sequence data stored in BIGSdb can be generated by any means and are uploaded as contigs. Sequence quality is controlled at the assembly stage before incorporation into the database, but not in the database itself, although experiment tags can be used to indicate the nature and type of the data included. Different types of data can be stored side-by-side in the sequence bin and can then be retrieved and analysed as required. Data can be uploaded from public archives, such as GenBank or the Integrated Microbial Genomes (IMG) database[71],

and the accession numbers for these repositories provide links between these databases and the BIGSdb record (FIG. 1).

When sequence data are deposited in BIGSdb, they are scanned against the definitions databases using an appropriate algorithm (currently BLAST[72], but other algorithms could be used). When a known allele is identified, this is recorded as a designation for that locus in the isolate record, and the position of the allele in the contig is marked (tagged) (FIG. 1). This process is rapid, and any new alleles identified can be assigned in the definition database; genomic data are periodically rescanned in an iterative process, resulting in an expanding and increasingly comprehensive catalogue of genomic diversity. Thus, as well as linking sequence data to phenotype, BIGSdb is a population-based annotation tool that enables the identification of particular genes and allelic variants within WGS data, maintaining a record of the known variation of that gene across the samples stored in the database. The assignment of alleles to loci can be defined using sequence similarity criteria, and this definition is stored in the locus record.

The grouping of defined loci into schemes is particularly useful for typing and taxonomy applications but has many other possible uses, including functional annotation. Combinations of alleles within schemes can be linked with other data — for example, to associate the sequences of genes with notable phenotypes such as antimicrobial resistance or susceptibility[54]. These schemes are not limited to given groups of organisms, but can be applied to any isolate in the database, which is particularly useful for the annotation of accessory genes that might be widely distributed in diverse organisms, or core genes that are shared among them. When a scheme has been defined, it can be readily applied to other isolates in the database.

The Genome Comparator module of BIGSdb facilitates gene-based comparative genomics by examining groups of shared genes in any number of isolates across any number of genes. For example, it can use an annotated reference genome as the source of comparison sequences to BLAST against other genomes in the database. This generates whole-genome profiles that can be analysed by standard distance methods such as NeighbourNet[73] implemented within the SplitsTree package[74], simultaneously providing lists of loci that are identical, variable, missing or truncated (incomplete genes located at the ends of contigs) among data sets. BIGSdb allows queries through web-based interfaces, making annotated genome content accessible to external databases and analysis software as well as to users querying the data directly (FIG. 1).

Sequence data and nomenclature

With a tool such as BIGSdb, the gene-by-gene approach can be used to catalogue genomic sequence variation and to map this onto existing or novel nomenclature schemes by associating particular alleles or combinations of alleles with names or designations. Nomenclature schemes are important means of communication which have to be agreed on among those who use them. They should be stable and preferably backwards compatible; however, they do require occasional modification as new insights are obtained, because stability should not be allowed to override accuracy in the description of biological diversity. By contrast, if determined correctly, sequence data and data summaries such as STs are absolute and require modification only to correct errors. The relationship between sequence data and nomenclature schemes therefore has to be transparent, and these two things should not be confused (FIG. 2).

Different levels of sequence information can be associated with different taxonomic levels. Analysis of a single locus, for example, is often sufficient to distinguish many groups (phylum, class, order, family or genus), whereas determining speciation and sub speciation

requires higher resolution, which can be attained by increasing the number of loci analysed. Researchers can apply a set of MLST approaches, each using different numbers of loci and each suitable for addressing different levels of isolate discrimination (FIG. 2). The highest level of resolution using a gene-by-gene approach can be termed whole-genome MLST (wgMLST), in which all the loci of a given isolate are compared to equivalent loci in other isolates. The wgMLST approach is applicable to single-clone pathogens with closed genomes or to very closely related variants of more diverse organisms. This analysis can involve the entire genome if loci corresponding to intergenic regions are also defined. Few bacteria share all loci, so comparisons of the core genome of a given group (coregenome MLST (cgMLST)) provides high-resolution data across a group of related but not identical isolates. Other subsets of loci might have particular applications, such as conventional seven-locus MLST. It has been found that a scheme based on the 53 ribosomal protein loci present in most bacteria — ribosomal MLST (rMLST) — is both highly flexible and informative[75].

rMLST for taxonomy and typing

For the purposes of bacterial typing and taxonomy, the ribosomal protein subunit (*rps*) genes have the advantages of being universally present but differentially variable[76]; indeed, although the variability of some *rps* genes has prevented their inclusion in sets of core genes[77], this variation permits high levels of discrimination among closely related isolates[30,75]. The *rps* genes have the further advantage of being distributed across the genome, offering some stability in the face of horizontal genetic transfer. An rMLST allelic profile, or ribosomal sequence type (rST), can accurately summarize the relationships between bacterial genomes, providing a manageable basis for universal bacterial systematics and typing, as exemplified by the use of rMLST to define species groupings and strain types within the genus *Neisseria*[30,75,78].

A database that catalogues variation in the 53 *rps* genes across the bacterial domain has been established (rMLST on PubMLST) using the BIGSdb platform[75]. This database currently includes more than 30,000 sets of assembled WGS data obtained from publicly available sources. All of these data can be indexed with rMLST, providing the basis for an efficient and rapid identification and characterization scheme[76]. The Genome Comparator tool on PubMLST can explore the rMLST loci within very large collections of WGS data from diverse bacterial isolates to identify clusters of isolates rapidly, and these clusters can then be investigated at higher resolution with approaches such as cgMLST and wgMLST as required[30,79] (FIG. 2). As this database contains WGS data, it has applications beyond indexing just the rMLST loci, and schemes for any set of genes present in the *de novo* assemblies can be established within the database, as illustrated below for staphylococci.

Applying gene-by-gene analysis

Several members of the genus *Staphylococcus* are human and animal pathogens, and methicillin-resistant *Staphylococcus aureus* (MRSA) strains prove particularly problematic, largely because of their role in nosocomial outbreaks[80,81]. Consequently, staphylococci have been subjected to many molecular epidemiology studies[80], including WGS analyses[41-46,82]. Many areas of the biology and pathology of these organisms have been investigated, including their evolution (particularly the origins of MRSA) in hospitals and communities[83,84], the spread of clones among animals and humans[85], the dynamics of colonization and infection[86], and outbreak detection and control[87]. Although WGS is increasingly used in these investigations, it is difficult to compare the data generated in different studies, as they are typically available only as short reads. This problem is not confined to staphylococci. Further, the various SNP-based approaches used in different

studies have not been standardized. Here, we illustrate how the gene-by-gene approach, implemented with the BIGSdb platform, can be used for combined analyses of data emanating from different studies of this important group of organisms.

WGS data for 926 staphylococci (889 *S. aureus* isolates and 37 genomes from 20 other species) were obtained from the Sequence Read Archive (SRA), GenBank and the IMG database (BOX 2; Supplementary information S1-S3 (tables)). For data from the SRA, draft genomes were assembled *de novo*, and contigs were uploaded to the PubMLST website. Using the BIGSdb web interface and SplitsTree, species groups were resolved among 52 staphylococci. Nucleotide sequence-based analysis of the rMLST loci defined groups within the genus and provided insights into the species assignments of two recently described staphylococcal isolates, *Staphylococcus* sp. OJ82 (REF. 88) and *Staphylococcus* sp. AL1 (REF. 89) (FIG. 3a). Furthermore, an allele-based rMLST analysis of 669 *S. aureus* isolates identified 229 unique rSTs based on 51 *rps* loci (two para logous loci, *rpsN* and *rpmG*, were excluded for the purposes of this analysis) and resolved the isolates into lineages that corresponded to previously described clonal complexes[83,90], also indicating genealogical relationships among the clonal complexes (FIG. 3b).

The detection and control of *S. aureus* outbreaks remains a high priority in health care systems[41-43]. The gene-by-gene approach described above can rapidly establish very high-resolution relationships among isolates, as demonstrated by an analysis of isolate data from a study that described an MRSA outbreak in a special-care baby unit[43]. Genome Comparator was used to compare all the isolates to a reference genome (from isolate *S. aureus* subsp. *aureus* HO 5096 0412) and provided a web-based visualization of the relationships among isolates from patients and from a health care worker (FIG. 3c); such a visualization would be easily and rapidly generated and interpreted in a health care setting. This analysis demonstrated that isolates from babies, mothers and their partners were closely related to each other and to those from a health-care worker, suggesting that there was an ongoing outbreak. Along with other available data, these results could have been interpreted in such a way that prompted interventions to prevent further transmission. Similar results were presented in the original report, but the gene-by-gene approach is independent of a specialized bioinformatics pipeline. This approach is appropriate for the very high-resolution analysis of isolates from a clinical setting, achieving a resolution equivalent to that obtained by SNP-based approaches.

The analyses presented here are not exhaustive; rather, they are a demonstration of the potential of the gene-by-gene approach as implemented in BIGSdb. This approach can be extended to any set of isolates and any combination of genes using web-based tools. BIGSdb can be used to store, organize and interpret WGS data from many different investigations in a single analysis; however, the database is sufficiently compact that WGS data from tens of thousands of isolates can be stored on a modestly sized personal computer for use in settings where Internet connectivity is poor or sporadic, or where clinically sensitive data need to be stored separately. The analyses relied on straight-forward assumptions and did not require processor- or memory-intensive algorithms, nor did they involve selective removal of data that do not conform to a particular evolutionary model. An additional advantage of using such an approach is that Genome Comparator automatically provides lists of those loci that are identical and those that are variable, enabling genetic differences to be rapidly identified.

Conclusions and future directions

WGS studies of bacteria are currently in a golden age, with many resources and much interest being dedicated to the area as a whole and especially to WGS applications for clinical problems[36-38,63]. However, when this initial excitement has passed, it will be

necessary to move into a more prosaic exploitation phase, which will require the construction of sustainable infrastructures — sustainable both intellectually (that is, the concept of relating WGS data to typing and taxonomy) and physically (that is, the use of inter-operable data sets that are archived in accessible databases). MLST offers mixed precedents for this process. On the positive side, the engagement of communities of scientists working on particular areas has enabled the assembly of many high-quality data sets, and this ‘bottom up’ approach is essential to ensure that appropriate typing schemes are developed and maintained by those people who need to use them. These efforts have largely been ‘amateur’, however, in the sense that they are not directly funded, and most MLST databases rely on enthusiastic community members. The careful curation of sequences and maintenance of typing schemes are essential ‘public goods’ that all researchers in the field wish to use, but to which few are willing, or able, to contribute.

A number of initiatives are currently underway to build such infrastructures, and these initiatives will have to balance community involvement and sustainability (BOX 2). The development of multiple approaches, as long as they are open access, inter-operable and sustainable, is important at this stage, when it is unclear exactly what the most effective and broadly acceptable means of exploiting WGS data will be. However, maintaining databases is challenging in the face of contracting resources as this is not a high-impact activity, although it is essential. The PubMLST model discussed here is one approach to these problems; this approach attempts to balance centralization (in the form of a curated set of reference genes) with affordability, open access and community involvement. The task of relating WGS data to particular nomenclature schemes is of sufficient magnitude and complexity that a whole community of scientists and others will need to engage with the project.

As with all paradigm shifts and ‘disruptive technologies’, the exploitation of WGS is simultaneously exciting and confusing, and presents great opportunities along with major challenges. Ultimately, however, the availability of large volumes of WGS data promises the development of a new type of microbiology research in which population and evolutionary approaches are integrated with functional and structural studies by the organized presentation and interpretation of biodiversity data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the IMG, GenBank and SRA databases, isolates from which were included in the analyses presented in this article. All isolate genomes can be accessed at the rMLST database on PubMLST and are identified under the project names Maiden 2013 Nat Rev SCBU for the outbreak analysis, Maiden 2013 Nat Rev S aureus for the other *S. aureus* analyses and Maiden 2013 Nat Rev rMLST for the genus *Staphylococcus* rMLST analysis, with original database accession numbers provided as available. The authors are grateful to J. S. Bennett, H. B. Bratcher, C. Brehony, A. J. Cody, F. Colles, O. B. Harrison, D. M. Hill, S. K. Sheppard, E. R. Watkins and H. Wimalarathna, as well as many other collaborators, for their contributions to the development of the context of this work and for comments on the manuscript. M.C.J.M. is a Wellcome Trust Senior Fellow in Basic Biomedical Sciences. M.J.J.v.R. is funded by the Clarendon Fund and Merton College, Oxford University, UK, and J.E.B. is funded by the Patho-NGen-Trace consortium. The research from the Patho-NGen-Trace project leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013; grant 278864).

FURTHER INFORMATION

BacMap: http://bacmap.wishartlab.com

Europe PMC Funders Author Manuscripts

BIGSdb: http://pubmlst.org/software/database/bigsdb

Centre for Genomic Epidemiology: http://www.genomicepidemiology.org

ENA: http://www.ebi.ac.uk/ena

EnsemblBacteria: http://bacteria.ensembl.org/index.html

GenBank: http://www.ncbi.nlm.nih.gov/genbank

Genomes Online: http://www.genomesonline.org

GMI: http://www.globalmicrobialidentifier.org

IMG: http://img.jgi.doe.gov

Institut Pasteur MLST databases: http://www.pasteur.fr/mlst

Microbial Genome Database for Comparative Analysis: http://mbgd.genome.ad.jp

MLST Databases at the ERI, UCC: http://mlst.ucc.ie

MLST homepage: http://www.MLST.net

Molecular Biology Database Collection: http://www.oxfordjournals.org/nar/database/a

NCBI Genome: http://www.ncbi.nlm.nih.gov/genome

PATRIC: http://patricbrc.org

PubMLST: http://pubmlst.org

PubMLST database collection: http://pubmlst.org/databases.shtml

rMLST on PubMLST: http://pubmlst.org/rmlst

SRA: http://www.ncbi.nlm.nih.gov/sra

UCSC Microbial Genome Browser: http://microbes.ucsc.edu

Xbase: http://www.xbase.ac.uk

References

1. Ciccarelli FD, et al. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006; 311:1283–1287. [PubMed: 16513982]
2. Medini D, et al. Microbiology in the post-genomic era. Nat. Rev. Microbiol. 2008; 6:419–430. [PubMed: 18475305]
3. Rusch DB, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007; 5:e77. [PubMed: 17355176]
4. Fournier PE, Raoult D. Prospects for the future using genomics and proteomics in clinical microbiology. Annu. Rev. Microbiol. 2011; 65:169–188. [PubMed: 21639792]
5. Stackebrandt E, et al. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 2002; 52:1043–1047. [PubMed: 12054223]
6. Koch R. An address on bacteriological research. Br. Med. J. 1890; 2:380–383.

7. Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. The bacterial species challenge: making sense of genetic and ecological diversity. Science. 2009; 323:741–746. [PubMed: 19197054]
8. Buckee CO, et al. Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. Proc. Natl Acad. Sci. USA. 2008; 105:15082–15087. [PubMed: 18815379]
9. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. Annu. Rev. Genet. 2004; 38:771–792. [PubMed: 15568993]
10. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 2008; 11:472–477. [PubMed: 19086349]
11. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms — proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl Acad. Sci. USA. 1990; 87:4576–4579. [PubMed: 2112744]
12. Fox GE, et al. The phylogeny of prokaryotes. Science. 1980; 209:457–463. [PubMed: 6771870]
13. Clarridge JE. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. Clin. Microbiol. Rev. 2004; 17:840–862. [PubMed: 15489351]
14. Vamosi SM, Heard SB, Vamosi JC, Webb CO. Emerging patterns in the comparative analysis of phylogenetic community structure. Mol. Ecol. 2009; 18:572–592. [PubMed: 19037898]
15. Hugenholtz P, Goebel BM, Pace NR. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 1998; 180:4765–4774. [PubMed: 9733676]
16. Sabat AJ, et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill. 2013; 18:20380. [PubMed: 23369389]
17. Maiden MCJ, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl Acad. Sci. USA. 1998; 95:3140–3145. [PubMed: 9501229]
18. Perez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect. Genet. Evol. 2013; 16:38–53. [PubMed: 23357583]
19. Smith JM, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? Proc. Natl Acad. Sci. USA. 1993; 90:4384–4388. [PubMed: 8506277]
20. Maiden MCJ. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 2006; 60:561–588. [PubMed: 16774461]
21. Selander RK, et al. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. 1986; 51:837–884.
22. Didelot X, Maiden MC. Impact of recombination on bacterial evolution. Trends Microbiol. 2010; 18:315–322. [PubMed: 20452218]
23. Turner KM, Feil EJ. The secret life of the multilocus sequence type. Int. J. Antimicrob. Agents. 2007; 29:129–135. [PubMed: 17204401]
24. Do T, et al. Population structure of Streptococcus oralis. Microbiology. 2009; 155:2593–2602. [PubMed: 19423627]
25. Webb K, et al. Development of an unambiguous and discriminatory multilocus sequence typing scheme for the Streptococcus zooepidemicus group. Microbiology. 2008; 154:3016–3024. [PubMed: 18832307]
26. Coffey TJ, et al. First insights into the evolution of Streptococcus uberis: a multilocus sequence typing scheme that enables investigation of its population biology. Appl. Environ. Microbiol. 2006; 72:1420–1428. [PubMed: 16461695]
27. Enright MC, Spratt BG. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. Microbiology. 1998; 144:3049–3060. [PubMed: 9846740]
28. Priest FG, Barker M, Baillie LW, Holmes EC, Maiden MC. Population structure and evolution of the Bacillus cereus group. J. Bacteriol. 2004; 186:7959–7970. [PubMed: 15547268]
29. Achtman M, et al. Microevolution and history of the plague bacillus, Yersinia pestis. Proc. Natl Acad. Sci. USA. 2004; 101:17837–17842. [PubMed: 15598742]

30. Jolley KA, et al. Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods. J. Clin. Microbiol. 2012; 50:3046–3053. [PubMed: 22785191]
31. Holt KE, et al. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nature Genet. 2008; 40:987–993. [PubMed: 18660809]
32. Jolley KA, Brehony C, Maiden MC. Molecular typing of meningococci: recommendations for target choice and nomenclature. FEMS Microbiol. Rev. 2007; 31:89–96. [PubMed: 17168996]
33. Dingle KE, McCarthy ND, Cody AJ, Peto TE, Maiden MC. Extended sequence typing of Campylobacter spp., United Kingdom. Emerg. Infect. Dis. 2008; 14:1620–1622. [PubMed: 18826829]
34. Adair DM, et al. Diversity in a variable-number tandem repeat from Yersinia pestis. J. Clin. Microbiol. 2000; 38:1516–1519. [PubMed: 10747136]
35. Parkhill J, Wren BW. Bacterial epidemiology and biology — lessons from genome sequencing. Genome Biol. 2011; 12:230. [PubMed: 22027015]
36. Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr. Opin. Microbiol. 2010; 13:625–631. [PubMed: 20843733]
37. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nature Rev. Genet. 2012; 13:601–612. [PubMed: 22868263]
38. Loman NJ, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. 2012; 10:599–606. [PubMed: 22864262]
39. Junemann S, et al. Updating benchtop sequencing performance comparison. Nat. Biotechnol. 2013; 31:294–296. [PubMed: 23563421]
40. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology-present and future. Philos Trans R Soc Lond B Biol Sci. 2013; 368:20120202. [PubMed: 23382424]
41. Köser CU, et al. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. New Engl. J. Med. 2012; 366:2267–2275. [PubMed: 22693998]
42. Eyre DW, et al. A pilot study of rapid benchtop sequencing of Staphylococcus aureus and Clostridium difficile for outbreak detection and surveillance. BMJ Open. 2012; 2:e001124.
43. Harris SR, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a descriptive study. Lancet Infect. Dis. 2013; 13:130–136. [PubMed: 23158674]
44. McAdam PR, et al. Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant Staphylococcus aureus. Proc. Natl Acad. Sci. USA. 2012; 109:9107–9112. [PubMed: 22586109]
45. Holden MT, et al. A genomic portrait of the emergence, evolution and global spread of a methicillin resistant Staphylococcus aureus pandemic. Genome Res. 2013; 23:653–664. [PubMed: 23299977]
46. Young BC, et al. Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc. Natl Acad. Sci. USA. 2012; 109:4550–4555. [PubMed: 22393007]
47. Mutreja A, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature. 2011; 477:462–465. [PubMed: 21866102]
48. Bryant JM, et al. Whole-genome sequencing to identify transmission of Mycobacterium abscessus between patients with cystic fibrosis: a retrospective cohort study. Lancet. 2013; 381:1551–1560. [PubMed: 23541540]
49. He M, et al. Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc. Natl Acad. Sci. USA. 2010; 107:7527–7532. [PubMed: 20368420]
50. Cui YJ, et al. Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. Proc. Natl Acad. Sci. USA. 2013; 110:577–582. [PubMed: 23271803]
51. Walker TM, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect. Dis. 2013; 13:137–146. [PubMed: 23158499]
52. Grad YH, et al. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. Proc. Natl Acad. Sci. USA. 2012; 109:3065–3070. [PubMed: 22315421]

53. Iqbal Z, Turner I, McVean G. High-throughput microbial population genomics using the Cortex variation assembler. Bioinformatics. 2013; 29:275–276. [PubMed: 23172865]
54. Jolley KA, Maiden MC. Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill. 2013; 18:20379. [PubMed: 23369391]
55. Sheppard SK, Jolley KA, Maiden MCJ. A gene-by-gene approach to bacterial population genomics: whole genome MLST of Campylobacter. Genes. 2012; 3:261–277. [PubMed: 24704917]
56. Bratcher HB, Bennett JS, Maiden MCJ. Evolutionary and genomic insights into meningococcal biology. Future Microbiol. 2012; 7:873–885. [PubMed: 22827308]
57. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc. Natl Acad. Sci. USA. 2004; 101:3160–3165. [PubMed: 14973198]
58. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
59. Chain PS, et al. Genomics. Genome project standards in a new era of sequencing. Science. 2009; 326:236–237. [PubMed: 19815760]
60. Hunt M, et al. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013; 14:R47. [PubMed: 23710727]
61. Enright MC, et al. The evolutionary history of methicillin-resistant Staphylococcus aureus (MRSA). Proc. Natl Acad. Sci. USA. 2002; 99:7687–7692. [PubMed: 12032344]
62. Harmsen D, Rothganger J, Frosch M, Albert J. RIDOM: ribosomal differentiation of medical microorganisms database. Nucleic Acids Res. 2002; 30:416–417. [PubMed: 11752353]
63. Koser CU, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog. 2012; 8:e1002824. [PubMed: 22876174]
64. Aarestrup FM, et al. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg. Infect. Dis. 2012; 18:e1. [PubMed: 23092707]
65. Carrico JA, Sabat AJ, Friedrich AW, Ramirez M, ESCMID Study Group for Epidemiological Markers (ESGEM). Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill. 2013; 18:20382. [PubMed: 23369390]
66. Chan MS, Maiden MC, Spratt BG. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics. 2001; 17:1077–1083. [PubMed: 11724739]
67. Jolley KA, Chan MS, Maiden MC. mlstdbNet — distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics. 2004; 5:86. [PubMed: 15230973]
68. Aanensen DM, Spratt BG. The multilocus sequence typing network: mlst.net. Nucleic Acids Res. 2005; 33:W728–W733.mlst.net [PubMed: 15980573]
69. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010; 11:595. [PubMed: 21143983]
70. Jolley KA, Maiden MC. AgdbNet — antigen sequence database software for bacterial typing. BMC Bioinformatics. 2006; 7:314. [PubMed: 16790057]
71. Markowitz VM, et al. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res. 2010; 38:D382–D390. [PubMed: 19864254]
72. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]
73. Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 2004; 21:255–265. [PubMed: 14660700]
74. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics. 1998; 14:68–73. [PubMed: 9520503]
75. Jolley KA, et al. Ribosomal multi-locus sequence typing: universal characterization of bacteria from domain to strain. Microbiology. 2012; 158:1005–1015. [PubMed: 22282518]

76. Yutin N, Puigbo P, Koonin EV, Wolf YI. Phylogenomics of prokaryotic ribosomal proteins. PLoS ONE. 2012; 7:e36972. [PubMed: 22615861]
77. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008; 9:R151. [PubMed: 18851752]
78. Bennett JS, et al. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. Microbiology. 2012; 158:1570–1580. [PubMed: 22422752]
79. Cody AJ, et al. Real-time genomic epidemiology of human Campylobacter isolates using whole genome multilocus sequence typing. J. Clin. Microbiol. 2013; 51:2526–2534. [PubMed: 23698529]
80. Stefani S, et al. Meticillin-resistant Staphylococcus aureus (MRSA): global epidemiology and harmonisation of typing methods. Int. J. Antimicrob. Agents. 2012; 39:273–282. [PubMed: 22230333]
81. Widerstrom M, Wistrom J, Sjostedt A, Monsen T. Coagulase-negative staphylococci: update on the molecular epidemiology and clinical presentation, with a focus on Staphylococcus epidermidis and Staphylococcus saprophyticus. Eur. J. Clin. Microbiol. Infect. Dis. 2012; 31:7–20. [PubMed: 21533877]
82. Harris SR, et al. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327:469–474. [PubMed: 20093474]
83. Lindsay JA. Genomic variation and evolution of Staphylococcus aureus. Int. J. Med. Microbiol. 2010; 300:98–103. [PubMed: 19811948]
84. Uhlemann AC, Otto M, Lowy FD, Deleo FR. Evolution of community- and healthcare-associated methicillin-resistant Staphylococcus aureus. Infect. Genet. Evol. 2013http://dx.doi.org/10.1016/j.meegid.2013.04.030
85. Fitzgerald JR. Livestock-associated Staphylococcus aureus: origin, evolution and public health threat. Trends Microbiol. 2012; 20:192–198. [PubMed: 22386364]
86. Fitzgerald JR. Evolution of Staphylococcus aureus during human colonization and infection. Infect. Genet. Evol. 2013http://dx.doi.org/10.1016/j.meegid.2013.04.020
87. Lindsay JA. Evolution of *Staphylococcus aureus* and MRSA during outbreaks. Infect. Genet. Evol. 2013http://dx.doi.org/10.1016/j.meegid.2013.04.017
88. Sung JS, Chun J, Choi S, Park W. Genome sequence of the halotolerant Staphylococcus sp. strain OJ82, isolated from Korean traditional salt-fermented seafood. J. Bacteriol. 2012; 194:6353–6354. [PubMed: 23105083]
89. Chong TM, Tung HJ, Yin WF, Chan KG. Insights from the genome sequence of quorumquenching Staphylococcus sp. strain AL1, isolated from traditional Chinese soy sauce brine fermentation. J. Bacteriol. 2012; 194:6611–6612. [PubMed: 23144375]
90. Holt DC, et al. A very early-branching Staphylococcus aureus lineage lacking the carotenoid pigment staphyloxanthin. Genome Biol. Evol. 2011; 3:881–895. [PubMed: 21813488]
91. Milkman, R.; McKane, M. Population Genetics of Bacteria. Baumberg, S.; Young, JPW.; Wellington, EMH.; Saunders, JR., editors. Cambridge Univ. Press; 1995. p. 127-142.
92. Lapage, SP., et al. International Code of Nomenclature of Bacteria: Bacteriological Code, 1990 Revision. American Society for Microbiology; 1992.
93. Smith JM, Dowson CG, Spratt BG. Localized sex in bacteria. Nature. 1991; 349:29–31. [PubMed: 1985260]
94. Spratt BG. Hybrid penicillin-binding proteins in penicillin-resistant strains of Neisseria gonorrhoeae. Nature. 1988; 332:173–176. [PubMed: 3126399]
95. Fernandez-Suarez XM, Galperin MY. The 2013 nucleic acids research database issue and the online molecular biology database collection. Nucleic Acids Res. 2013; 41:D1–D7. [PubMed: 23203983]

Box 1

A narrative glossary of terms

Much of experimental bacteriology depends on the isolate (a bacterial specimen separated from its original environment and cultured in the laboratory). Ideally, all the cells in an isolate culture are clones and thus the direct descendants of a single cell and genetically identical; however, this might not be the case if mutations occur during the propagation of the isolate or there is more than one variant present in the culture. The term meroclone has been used to describe a group of organisms that are descended from a single cell but have started to diversify by recombination[91].

Very similar isolates that share important characteristics can be grouped into strains. Although this term is used interchangeably with isolate by some microbiologists, it is useful to reserve it for a group of very similar bacteria that share an important set of properties, such as the propensity to cause disease. Thus, in public health settings, it is often necessary to define the ‘outbreak strain’ and assign isolates to it.

Comparisons of isolates from many different bacteria show that they can be assigned to distinct clusters or groups, variously referred to as lineages or clonal complexes, the members of which are inferred to share a recent common ancestor. The genetic distances within these clusters are appreciably smaller than those between clusters, and clusters can be sufficiently distinct to be referred to as subspecies. Bacterial nomenclature at the level of subspecies and above is regulated by the Bacteriological Code[92], which provides guidelines for the naming of clusters for groupings such as species, genus and family for the domains Archaea and Bacteria, the highest levels of taxonomy. However, there remains no generally agreed way of defining these groups across the whole domain, and groups are not necessarily equivalent.

Bacterial isolates have genomes that comprise a core genome (genes present in all members of a given population subset), often with an accessory genome (genes that are variably present in isolates from that population). All of the genes available to the population represent the pan-genome[10]. Genome evolution proceeds by a combination of horizontal (or lateral) and vertical genetic transfer, the balance of which varies widely among bacteria[22]. Vertical transfer is the passing of genetic material by descent, whereas horizontal transfer (sometimes called localized sex) is the movement of genetic material among bacteria that do not necessarily share a mother cell[93]. This occurs by transformation (the uptake of DNA by a cell), conjugation (transfer facilitated by conjugative elements) and phage-mediated transduction. Transferred DNA can comprise large multigene fragments, distinct genes or genetic elements, or fragments of genes, which can be incorporated into genes that are already in the chromosome by means of homologous recombination, yielding mosaic genes, different parts of which have different evolutionary histories[94].

Europe PMC Funders Author Manuscripts

Box 2

A short guide to other microbial genome databases

Databases are an essential means of sharing the ever increasing volumes of bacterial whole-genome sequencing (WGS) and related data. There are advantages to having many different systems available, because they each provide different services to different communities, but as the number of such databases expands (see, for example, the Molecular Biology Database Collection[95]), the need for database inter-operability increases[65]. The US National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Databank of Japan (DDBJ) are members of the International Nucleotide Sequence Database Collaboration and have provided sequence databases for many years, including microbial genome databases with annotation tools (for example, NCBI Genome and EnsemblBacteria). These resources also provide the sequence read archives (SRAs; for example, the NCBI SRA database and the European Nucleotide Archive (ENA)), which are crucial repositories for WGS data. A number of databases have been developed specifically for storing and analysing complete genomes and metagenomes, including Integrated Microbial Genomes (IMG), Genomes Online, the Microbial Genome Database for Comparative Analysis, the UCSC Microbial Genome Browser, Xbase and BacMap. Enterprises more specifically developed for systematics and epidemiological applications include the Centre for Genomic Epidemiology, the Pathosystems Resource Integration Center (PATRIC) and the planned Global Microbial Identifier (GMI).

Europe PMC Funders Author Manuscripts

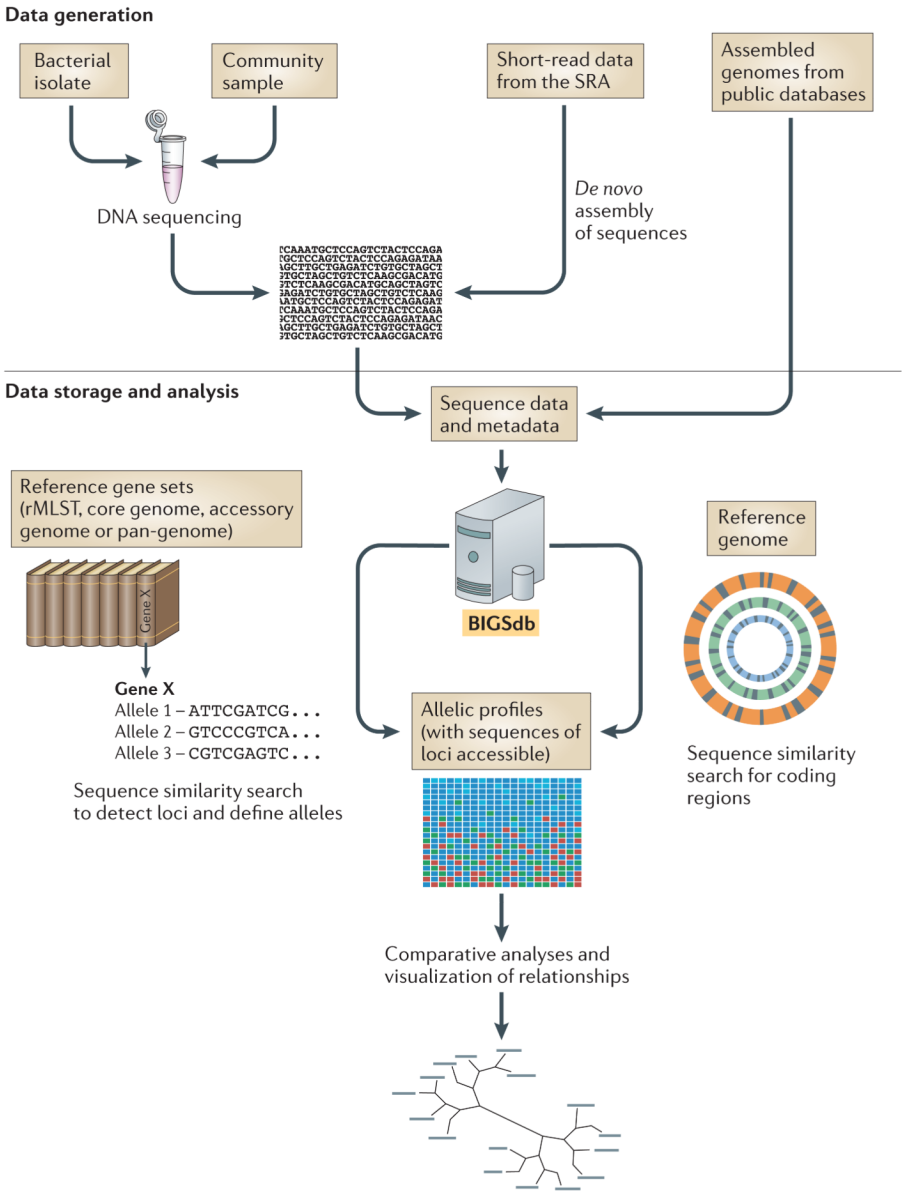

Figure 1. Schematic illustration of the gene-by-gene approach to the analysis of genome sequences using the Bacterial Isolate Genome Sequence Database platform

The gene-by-gene approach can be used to integrate whole-genome sequencing (WGS) data with isolate data, facilitating easy storage and retrieval for downstream analyses. WGS data can be obtained in several ways: DNA can be isolated from a bacterial isolate or community sample and sequenced on an appropriate platform, short-read data can be obtained from the Sequence Read Archive (SRA), and assembled genomes can be downloaded from public databases (for example, GenBank). In the cases of sequencing and SRA data, short-read data are assembled with an appropriate algorithm, directed by the sequencing platform used. Assembled contigs are uploaded to a Bacterial Isolate Genome Sequence Database (BIGSdb), and they can then be compared against either sets of reference genes or a reference genome using algorithms such as BLAST. Reference gene sets can be tailored to meet particular requirements and thus range from collections of loci that are useful for epidemiological investigations to subsets of genes with functional relevance, for example in

a metabolic pathway. For comparisons based on both reference gene sets and reference genomes, the nucleotide sequences remain accessible, but loci are assigned allele designations to generate an allelic profile as for multilocus sequence typing (MLST). The Genome Comparator module in BIGSdb can be used to produce a distance matrix based on allelic profiles, and this matrix can in turn be used to visualize relationships between isolates using an appropriate algorithm, such as NeighbourNet in SplitsTree. Alternatively, downstream analyses of the aligned sequence data can be carried out by exporting sequence data to external packages. In addition to a distance matrix and sequence alignments, the Genome Comparator outputs also include a table showing which loci are identical and which are different among the isolates examined. rMLST, ribosomal MLST.

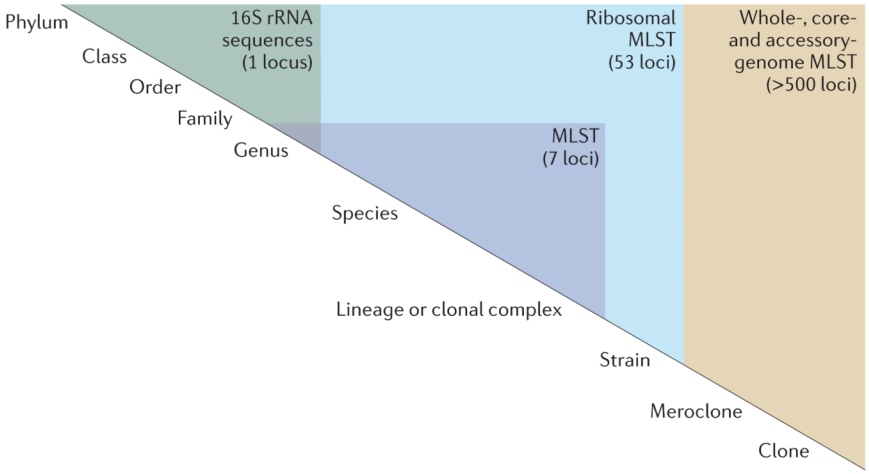

Figure 2. Relating sequence data to nomenclature schemes

Hierarchical nomenclature schemes are artificial constructs that are developed to facilitate communication and are subject to change as new information becomes available. Nomenclature schemes are dependent on various factors, including sequence relationships, and are ideally genealogically based. The challenge is to map whole-genome sequencing (WGS) data to nomenclature schemes transparently but flexibly at a range of resolutions. The highest discrimination is required for studies of bacterial isolates from one patient or from very closely related transmission chains; these isolates can be thought of as having undergone microevolution. Such studies will require comparisons of whole genomes using whole-genome multilocus sequence typing (MLST)[30]. Progressively lower resolution is required for studies of isolates with more distant common ancestors and, therefore, with more genetic differences. These relationships are best studied using the core genome common to the set of isolates of interest. Genes encoding ribosomal proteins are a particularly useful subset of core genes, and ribosomal MLST[75] accommodates many levels of genealogical relationships, from clonal complexes and lineages to species[78], genera and beyond[75]. In a database such as the Bacterial Isolate Genome Sequence Database (BIGSdb), multiple gene-by-gene schemes can be implemented alongside other, more conventional sequence-based schemes[69]. Particular genotype summaries of genes or collections of genes can be associated with particular nomenclature schemes, enabling the database to deliver a plain-language report to the user.

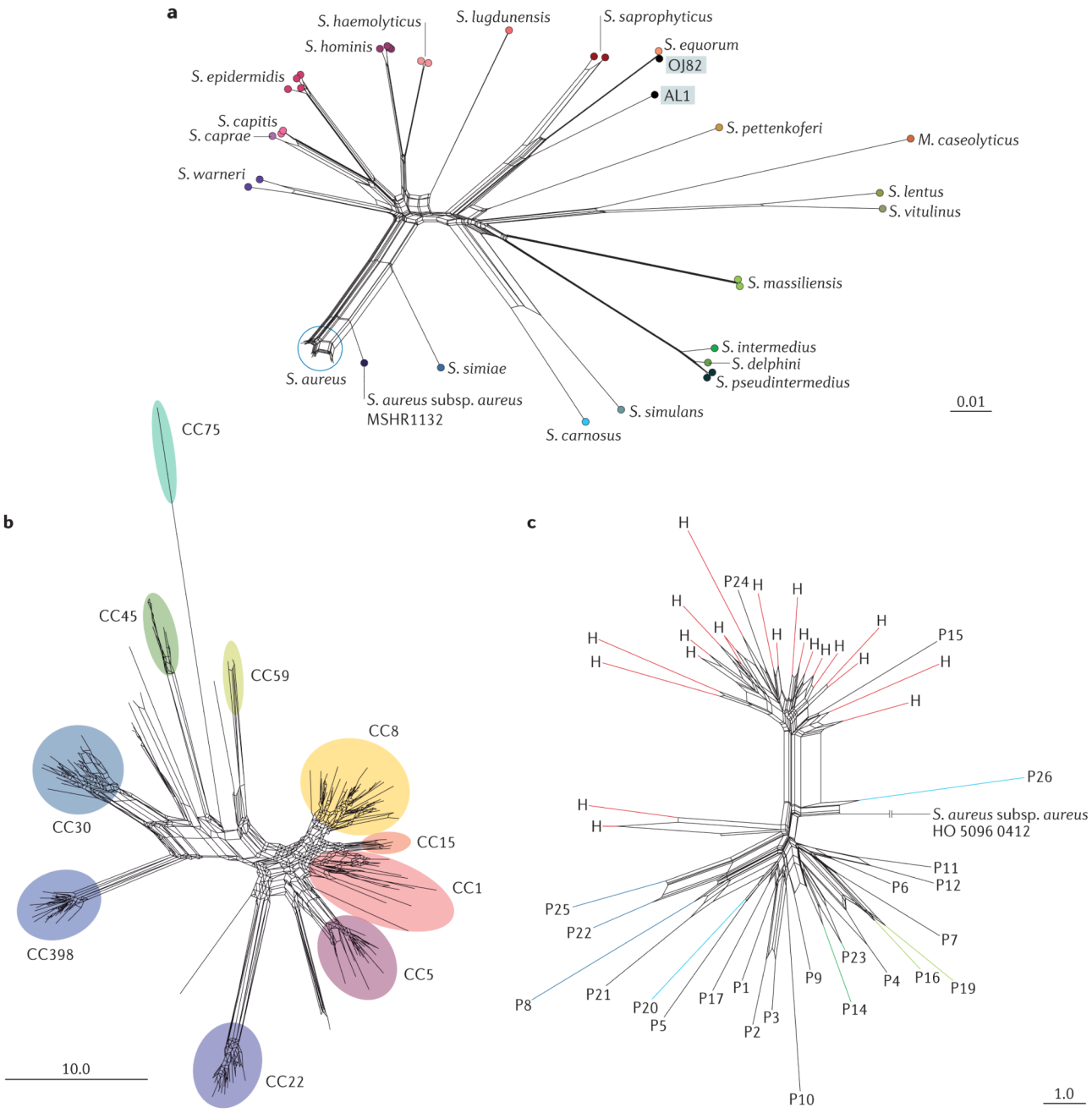

Figure 3. Ribosomal multilocus sequence typing-based analysis of *Staphylococcus* spp. whole-genome sequence data

These analyses were carried out using both the Genome Comparator module within the Bacterial Isolate Genome Sequence Database (BIGSdb) platform and data publicly available within the PubMLST website. The phylogenetic networks were generated using the NeighbourNet algorithm in SplitsTree (v4.12.3). **a** | Resolution of 52 staphylococcal isolates on the basis of nucleotide sequence diversity at 51 ribosomal multilocus sequence typing (rMLST) loci, permitting the determination of the species assignment of two recently described isolates, *Staphylococcus* sp. OJ82 and *Staphylococcus* sp. AL1. *Staphylococcus* sp. OJ82 probably corresponds to *Stapylococcus equorum*, whereas *Staphylococcus* sp. AL1

Europe PMC Funders Author Manuscripts

is related to, but distinct from, *S. equorum* and *Staphylococcus saprophyticus*. All species shown are in the genus *Staphylococcus*, except for *Macrococcus caseolyticus*. **b** | The diversity of 669 *Staphylococcus aureus* isolates on the basis of allelic diversity at 51 rMLST loci. The extensive diversity of *S. aureus* is illustrated here; the rMLST clustering is congruent with MLST clonal complexes (CCs; indicated) and indicates relationships among the isolates. **c** | Resolution of multidrug-resistant *S. aureus* (MRSA) isolates from an outbreak in a special-care baby unit[41], using a gene-by-gene comparison to a reference genome (*S. aureus* subsp. *aureus* HO 5096 0412). Twenty isolates obtained from a health care worker are indicated with the letter H and shown in red, whereas patient isolates are indicated with a letter P. Groups of isolates from patients who were members of the same family are shown in the same colour. Reticulations in the diagrams indicate departures from a strictly tree-like phylogeny; this can have a number of causes, including homoplasy as a result of recombination, mutation or lack of resolution. Such graphs are rapidly produced and represent the relationships among sets of genome data; these relationships can then be readily used to resolve isolate relationships in clinical and other settings. Scale bars represent distances calculated from the nucleotide sequence alignment (part **a**) and number of loci (parts **b**,**c**)