Published in final edited form as:
Euro Surveill. ; 18(4): 20379.

Automated extraction of typing information for bacterial pathogens from whole genome sequence data: *Neisseria meningitidis* as an exemplar

Keith A. Jolley and **Martin C.J. Maiden**
Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom

Abstract

Whole genome sequence (WGS) data are becoming a major means of characterising samples of bacterial pathogens. These data have the advantage of providing detailed information on the genotypes and likely phenotypes of aetiological agents, enabling the relationships of samples from potential disease outbreaks to be established precisely. However, the generation of increasing quantities of sequence data does not, in itself, resolve the problems that a wide variety of microbiological typing methods have addressed over the last 100 years or so; indeed, the provision of very high volumes of unstructured data can confuse rather than resolve these issues. Here we review the nascent field of the storage of WGS data for clinical application and show how curated sequence-based typing schemes on websites such as PubMLST.org, accumulated over the past 14 years or so, has generated an infrastructure that can be used to exploit WGS for bacterial typing efficiently. We review the tools that have been implemented within the PubMLST.org website to extract clinically useful, strain characterisation information which can be provided to physicians and public health scientists and officials in a timely, concise and understandable way. These data can be used to inform medical decisions such as how to treat a patient, whether to institute public health action, and what action might be appropriate. The information is compatible both with previous sequence-based typing data and also with data that can be obtained in the absence of WGS data, for example by real-time PCR tests, providing a flexible infrastructure for WGS-based clinical microbiology.

Keywords

Whole genome sequencing; antimicrobial resistance; MLST; antigen typing; meningococcus; epidemiology

Introduction

The application of whole genome sequencing (WGS) technology to clinical microbiology has been described as revolutionary, and certainly the opportunities are immense, yet so are the challenges of implementing this technology effectively [1]. Above all, clinical microbiology and epidemiology are pragmatic sciences, which require accurate and understandable information to be delivered to those who need to make medical judgements in real time. Often these judgements have to be made in the absence of complete information, and it is essential that widely understood, accepted and reproducible typing

Corresponding author: Martin C.J. Maiden, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, United Kingdom, Telephone: +44 1865 271284, Telefax: +44 1865 281275.

Europe PMC Funders Author Manuscripts

methods are employed to guide these decisions [2]. Just as the advent of molecular techniques challenged phenotypic methodologies over a decade ago, replacing imperfect but at least widely accepted techniques with a plethora of non-standardised alternatives [3], the high volumes of sequence data have to be carefully managed if they are to provide enlightenment rather than confusion.

The multilocus sequence typing (MLST) paradigm was established in 1998 [4], a time that molecular techniques were beginning to be widely used in the clinical laboratory, but when there was no universally agreed way forward [5]. It was intended as a standardised, reproducible, and portable approach that could replace and enhance previous methods, particularly multilocus enzyme electrophoresis (MLEE) [6]. MLST was the first sequence-based approach to the genome-wide characterisation of bacterial isolates to be widely adopted and automated methods for performing the reactions and extracting the sequence information have subsequently been developed [7-9]. At the time MLST was introduced it was impractical to sequence complete genomes on very large numbers of isolates and early analyses showed that in many cases this was not required. The first MLST scheme, for example, was designed to identify major clones within populations of *Neisseria meningitidis*, the meningococcus, and was able to do this reliably and reproducibly with just seven gene fragments, totalling only 3,284 bp, or about 0.15% of the whole genome [10, 11]. Similar numbers and sizes of loci have been successful for MLST schemes covering a wide range of organisms, which is an indication of the high degree of structuring present in many bacterial populations. For many bacteria, including the meningococcus, the extent of genetic diversity present even in this small number of genes under stabilising selection is extensive [12]: as of November 2012 each of the gene fragments used as meningococcal MLST loci had between 424 to 675 distinct alleles recorded on the PubMLST *Neisseria* website [13] with 54-94% (71% mean) sites that were polymorphic. Furthermore, in the representative *abcZ* locus all four bases were present at a given site over the known population in 54/433 (12%) of the nucleotide positions (Figure 1). Much of this variation is at low frequency and transitory, but the precise variants that this is the case for cannot be known without exhaustive, or at least extensive, sampling over time.

The MLST approach catalogues this extreme diversity, which is seen in many microbial populations and which remains only partially explored, by the maintenance of curated libraries of allele sequences for each MLST locus. Each unique sequence (allele) is assigned a unique arbitrary number, effectively compressing 400-600 bp of information into a single integer. Further organisation and compression of genetic variation is attained by combining the data from all MLST loci into allelic profiles or sequence types (STs), which are also assigned arbitrary numeric designations, each of which defines a unique string of several thousand nucleotides [12]. This approach has proved to be both efficient and effective: as of November 2012 there were 9927 STs in the *Neisseria* MLST database for example, each precisely characterising a particular seven locus *Neisseria* genotype. Similar levels of diversity have been observed in other bacteria hosted at PubMLST and on other MLST repositories [14]. The fact that nearly 10,000 distinct variants of only 3,284 bp of coding sequence under stabilising selection are known to exist in one human-associated bacterium with a genome of about 2.2Mbp, indicates the scale of the cataloguing problem facing us in the era of genomic microbiology.

Nevertheless, there are instances when even the very high levels of diversity routinely seen in MLST datasets do not provide sufficient information for clinical decision making. This is because even populations of diverse organisms, such as the meningococcus, are highly structured, with most isolates belonging to clonal complexes of related bacteria, and many of these share identical sequence types [15]. This detection of population structuring is one of the strengths of the MLST approach, as these clusters are frequently associated with

phenotypes of clinical interest such as virulence or expression of vaccine antigens [16]. This clustering, however, can mean that isolates with the same ST may not have the same point source so ST alone is insufficient to unambiguously identify strains belonging to an outbreak. For this reason, additional highly variable antigenic loci are included in the recommended typing scheme for meningococci [17], and for other organisms such as *Campylobacter* [18] which are regularly typed by MLST. For meningococci, there are also curated sequence based schemes for genes that encode antimicrobial resistance which provide additional clinically valuable information [19, 20]. Other schemes, such as multilocus variable number tandem repeat (VNTR) also provide high discrimination of isolates in outbreak situations [21, 22]. Combining these high-resolution typing approaches with seven locus MLST and spatial and temporal epidemiology techniques permits the proactive identification of outbreaks of infectious disease [23].

For a small number of bacteria, the so-called single clone pathogens, there is insufficient variation in seven locus MLST to provide epidemiological resolution, usually because these pathogens have evolved recently from single clones, undergo little recombination, and contain too little genetic variation [24]. This includes organisms of great medical importance such *Mycobacterium tuberculosis* [25], *Yersinia pestis* [26], *Bacillus anthracis* [27], and *Salmonella enterica* var Typhi [28]. For these bacteria data from the whole genome, often in the form of single nucleotide polymorphisms (SNPs) [29], but also including other types of variation such as VNTRs, is essential for epidemiological purposes. These data will also have to be stored and interpreted in an accessible way that produces data usable by clinical decision makers and which is both forwards and backwards compatible.

One of the motivations that drove the development of MLST was future-proofing. Even at a time when the costs of sequencing were seen by some as prohibitive [30], nucleotide sequence data had major advantages: they might be added to, but they would never become obsolete, as they represented the fundamental level of genetic information and they are readily understood, stored, compared, and distributed [12]. Obtaining WGS is now becoming so inexpensive that it is becoming the fastest and most economical way of obtaining information at multiple loci for determining MLST or other sequence types [31]. When used in this way, these data are directly comparable to the extensive sequence databases that have been established since the first use of MLST [32, 33]. Here we describe how the suite of databases hosted at PubMLST [34] has been updated to accommodate WGS data and describe the tools that are available to rapidly extract typing information from such data. We also describe how these tools can be exploited further to achieve very high resolution from such data on those occasions when this is required.

Database structure

As of November 2012, the majority of the typing databases hosted at PubMLST [34] were using the Bacterial Isolate Genome Sequence Database (BIGS$_{DB}$) platform to archive isolate and sequence diversity data [35]. This software was developed to facilitate the flexible storage and exploitation of the whole range of sequence data that might be available from a clinical specimen, from single Sanger sequencing reads through to whole genomes, which may be either complete or consisting of multiple contiguous sequences (‘contigs’), as assembled from data from the current generation of sequencing instruments. The BIGS$_{DB}$ platform consists of two kinds of database: (i) a definition database that contains the sequences of known alleles of loci under study, as well as allelic profiles (combinations of alleles at specific loci) for schemes such as MLST; and (ii) an isolate database that contains isolate provenance and other metadata along with nucleotide sequences associated with that isolate. An isolate database can interact with any number of definition databases and *vice versa*, allowing networks of authoritative nomenclature servers and partitioning of isolate

datasets and projects, with curator access controlled by specific permissions set by an administrator.

Reference databases

The definition databases are central to genome analysis using the gene-by-gene (MLST-like) analysis approach implemented in BIGS$_{DB}$. By storing all known allelic diversity for any locus of interest, the definition databases provide a centralised queryable repository that provides a common language for expressing sequence differences, making it a trivial process to identify alleles that are different among isolates, and equally importantly, those that are identical. Because sequence differences are linked directly to a particular locus (which can be any definable sequence string, nucleotide or peptide), and with appropriate grouping of loci into 'schemes' (groups of related loci), the context of this locus is immediately apparent: identifying it, for example, as a member of a conventional MLST scheme, as responsible for antimicrobial resistance, as a participant of a biochemical pathway and so on. As of November 2012, the *Neisseria* PubMLST definition database had allelic sequences defined for 1272 loci with 114,469 unique alleles.

Extracting typing information

Web-based and stand-alone tools have been developed which facilitate identification of sequence types directly from short read data [36, 37]. These methods are, of course, dependent on the sequence and profile definitions made available on PubMLST.org, which also has functionality to extract typing information directly from submitted assembled genomes which are routinely scanned for known alleles. As the locations of these loci are 'tagged' in the sequence data for future reference within BIGS$_{DB}$, this means that the genome sequences are automatically annotated for those loci for which definition databases exist. The definition database can also be queried using genome data not uploaded to the isolate database to identify a strain direct from sequence data. The BIGS$_{DB}$ platform also has functionality that enables an administrator to define scanning rules and report formatting. This uses a built-in script interpreter so that analysis paths can be taken by following a decision tree defined by the rules. This has been implemented within the PubMLST *Neisseria* sequence definition database to automatically extract the strain typing information for the meningococcus (ST, clonal complex, and antigen sequence type comprising PorA variable regions and FetA variable region) [17, 33], along with antibiotic resistance information from sequence data that is pasted in to a web form (Figure 2A). The script instructs the software to first scan the MLST alleles and, if these are all identified, to identify the ST and clonal complex by querying the reference data tables. It then scans the typing antigens and formats the results of these with the MLST results in to a standardised strain designation [17]. Following this, the sequences of the *penA* and *rpoB* genes are extracted and then compared with isolates with matching sequences within the PubMLST isolate database to determine the most likely penicillin and rifampicin sensitivity. All of this is displayed in a plain language report (Figure 2B). The whole analysis is extremely rapid taking about 40 seconds within the web interface.

Comparing genomes

Because genomic diversity is stored within BIGS$_{DB}$ as allele numbers, WGS analysis is possible using the highly-scalable techniques developed for seven-locus MLST. Once loci have been defined and alleles identified, they can be used essentially as a whole-genome MLST scheme, or any chosen subset of predefined loci combined to form a scheme. This is the principle behind the Genome Comparator analysis [38], which can use either the defined loci or extract coding sequences from an annotated reference genome to perform

comparisons against genomes within the database. Using a reference genome, or set of predefined reference loci, each of the coding sequences are compared against the test genomes using BLAST. Allele sequences that are the same as the reference are designated allele 1, while each unique allele different from the reference is assigned a sequential number. Once each locus has been tested, a distance matrix is then generated based on allelic identities between each pair of isolates. This can then be visualised using standard algorithms - the PubMLST website incorporates the NEIGHBORNET algorithm [39] implemented in SPLITSTREE4 [40]. Because analysis relies only on using BLAST to compare each locus within a genome in turn, either against the single annotated reference sequence or against all known alleles if using defined loci, the analysis is again very rapid, allowing multiple genomes to be compared within minutes, with the time taken to analyse only increasing linearly, not geometrically, with additional genomes.

The Genome Comparator approach is generic and any number of loci in any groups can be used for this type of analysis. Many loci have been defined for the meningococcus, including the 53 ribosomal genes that are used as a basis of rMLST [41-44]. The full complement of ribosomal genes has a number of advantages for indexing variation. These genes are: universally present in members of the domain; protein encoding and therefore generally assemble well from short-read sequences; are distributed around the genome; encode proteins which form part of a coherent, macromolecular structure; and they contain variation that is informative at a wide range of levels of discrimination. These data can be used within and among members of the same genus being useful for both species and strain definition [42].

Analysis of whole genome sequence data for meningococci

The *Neisseria* PubMLST database is continually expanding, but as of November 2012 there were 221 isolate records with deposited genome sequence data linked to published studies [11, 45-51]. Of these 221 genomes, 170 were meningococci, with the remainder belonging to other species within the genus [42]. The data consisted of a mixture of complete closed genomes, multiple contigs generated from *de novo* assembly, contigs generated by mapping to a reference sequence and sets of predicted coding sequences. These are treated identically by BIGSDB to identify and tag sequences of known loci, and where these loci are members of existing typing schemes, such as MLST or antigen typing, these genomes could be compared to legacy data (Table 1).

NEIGHBORNET visualisation of distance matrices generated with Genome Comparator from allelic ribosomal MLST (rMLST) data [44], provides a highly scalable, rapid, and easily understood way of placing isolates within the known diversity of a bacterial species. For example, the interrelationships of 139 *N. meningitidis* isolates present in the PubMLST *Neisseria* database [13] can be efficiently represented by this method. Since rMLST alleles are automatically tagged within the database, this analysis is rapid, and the NEIGHBORNET trees can be generated in a few minutes. The rMLST analysis differentiates clonal complexes; however, in addition it provides much higher resolution than conventional seven locus MLST, robustly indicating both relationships among and diversity within clonal complexes (Figure 3) [38].

Conclusions and future prospects

Nucleotide sequences are a universal language which can be interpreted in a number of ways. For clinical and epidemiological purposes, sequences from clinical specimens have to be rapidly and effectively translated into a meaningful term or set of terms that define those properties of the aetiological agents of disease which affect effective medical and public

health action. One of the factors behind the success of seven-locus MLST was the introduction of standard sets of nomenclature that reflected the structure of microbial populations and their phenotypic properties. For organisms with well-established and accepted MLST and other typing schemes in place, the impact of the application of WGS data will be to rapidly identify properties such as strain type. In some cases novel nomenclatures may be required, but this is a process that has to be approached with care if confusion in the wider clinical community is to be avoided.

The suite of database subsites on PubMLST, which now includes a site that catalogues the ribosomal diversity across the whole domain for the purposes of rMLST typing [44] [52], provides an example of how WGS data can be used to efficiently designate specimens to current strain types. It can be also used to establish additional typing schemes which can co-exist with each other side-by-side, as there is no limit to the number of loci and schemes that can be defined. As the database stores the sequence information that is available for an isolate, be that a single read or a whole genome, it means that it is possible to seamlessly compare isolates for which different types of information are available, achieving backwards compatibility with previous typing schemes, as well as compatibility with diagnostic tests that may target only one or a few loci. The extent to which isolates can be compared depends only on the quality of the sequence data available for the locus in question, but given that clinical specimens are often imperfect it is important for clinical and epidemiological purposes that incomplete or partial information can be used. While many studies place short read data in a sequence read archive (SRA) this is not easily accessible or readily analysed. PubMLST curators do proactively assemble short read data and incorporate the resultant contigs in to the database where metadata are available. Links are made to the SRA within PubMLST isolate records so that original data can be retrieved and analysed when required. While the *Neisseria* databases described are exemplars, databases for other species can be hosted on request and the open-source BIGSdb software is freely available for local installation.

The first analyses of WGS data on bacterial specimens relied on single nucleotide polymorphism (SNP) analysis of closely related bacteria, with mapping of sequence reads to a pre-defined reference genome. These have required pre-analysis of the samples by an approach such as MLST to limit the extent of variation being analysed [53-58]. This approach is also appropriate and can be very effective for ‘single clone’ pathogens [25-28]; however, it is not feasible for the general analysis for diagnosis or surveillance of bacteria such as the meningococcus which exhibit more typical levels of sequence diversity. Indeed, the use of the term SNP when discussing bacterial genome variation outside the examples described above, is unfortunate and can be misleading. The concept of the ‘SNP’ has been ported from human medical genomics to microbial genomics: while in the human it is in some cases appropriate to discuss single nucleotide polymorphisms, when they are associated with a particular genetic disease, genetic variation in terms of sequence polymorphism is much more complex in the bacteria. As seen here, the great majority of microbial populations contain tens of thousands of polymorphisms even within organisms that are closely related – not to mention large amounts of variation due to insertions, deletions and rearrangements which cannot even remotely be described as ‘SNPs’. The term sequence variation is more appropriate as individual polymorphisms, especially in the bacteria, are invariably embedded with many other variants into alleles and it is these alleles, each often with many variable sites that are associated with particular phenotypes.

Although the typing of bacterial specimens with existing schemes is a valuable contribution of WGS data to clinical microbiology and epidemiology it is not, of course, the only use for these data. There are many other possible applications for both research and detailed investigation of outbreaks [38]; however, it is important that the analysis of these data is

driven by the question that is being asked. If an outbreak can be resolved with a few loci then there is no need to pursue the data further and certainly no need to report more detail than necessary to a hard-pressed front-line clinician or epidemiologist who, in general, will only require the information necessary to resolve the medical problem at hand. In other cases resolution of a particular outbreak may require data from the whole genome [53]. For this reason it will be increasingly necessary to store WGS data from clinical specimens in understandable form, that is as assembled sequences, within flexible structures, such as that offered by the PubMLST platform powered by BIGS$_{DB}$, where WGS information can be hierarchically queried in real time by individuals with limited bioinformatics expertise to generate the data at the resolution required to address their problem. In this context these data will provide an exciting opportunity to extend our understanding of infectious disease caused by bacteria and will enhance our ability to combat it.

References

1. Pallen MJ, Loman NJ, Penn CW. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Curr Opin Microbiol. 2010; 13(5):625–31. [PubMed: 20843733]
2. van Belkum A, Tassios PT, Dijkshoorn L, Haeggman S, Cookson B, Fry NK, Fussing V, Green J, Feil E, Gerner-Smidt P, Brisse S, Struelens M. Guidelines for the validation and application of typing methods for use in bacterial epidemiology. Clin Microbiol Infect. 2007; 13(s3):1–46. [PubMed: 17716294]
3. Achtman M. A surfeit of YATMs? J Clin Microbiol. 1996; 34(7):1870. [PubMed: 8964891]
4. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA. 1998; 95(6):3140–3145. [PubMed: 9501229]
5. van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. J Clin Microbiol. 2001; 14(3):547–60.
6. Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS. Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol. 1986; 51:837–884.
7. Sullivan CB, Jefferies JM, Diggle MA, Clarke SC. Automation of MLST using third-generation liquid-handling technology. Mol Biotechnol. 2006; 32(3):219–26. [PubMed: 16632888]
8. Platt S, Pichon B, George R, Green J. A bioinformatics pipeline for high-throughput microbial multilocus sequence typing (MLST) analyses. Clin Microbiol Infect. 2006; 12(11):1144–6. [PubMed: 17002618]
9. O'Farrell B, Haase JK, Velayudhan V, Murphy RA, Achtman M. Transforming microbial genotyping: a robotic pipeline for genotyping bacterial strains. PLoS One. 2012; 7(10):e48022. [PubMed: 23144721]
10. Holmes EC, Urwin R, Maiden MCJ. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol Biol Evol. 1999; 16(6):741–749. [PubMed: 10368953]
11. Parkhill J, Achtman M, James KD, Bentley SD, Churcher C, Klee SR, Morelli G, Basham D, Brown D, Chillingworth T, Davies RM, Davis P, Devlin K, Feltwell T, Hamlin N, Holroyd S, Jagels K, Leather S, Moule S, Mungall K, Quail MA, Rajandream MA, Rutherford KM, Simmonds M, Skelton J, Whitehead S, Spratt BG, Barrell BG. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. Nature. 2000; 404(6777):502–506. [PubMed: 10761919]
12. Maiden MC. Multilocus Sequence Typing of Bacteria. Annu Rev Microbiol. 2006; 60:561–588. [PubMed: 16774461]
13. *Neisseria* MLST website. http://pubmlst.org/neisseria/http://pubmlst.org/neisseria/
14. MLST databases. http://pubmlst.org/databases.shtmlhttp://pubmlst.org/databases.shtml

15. Caugant DA, Maiden MC. Meningococcal carriage and disease--population biology and evolution. Vaccine. 2009; 27(Suppl 2):B64–70. [PubMed: 19464092]
16. Yazdankhah SP, Kriz P, Tzanakaki G, Kremastinou J, Kalmusova J, Musilek M, Alvestad T, Jolley KA, Wilson DJ, McCarthy ND, Caugant DA, Maiden MC. Distribution of serogroups and genotypes among disease-associated and carried isolates of *Neisseria meningitidis* from the Czech Republic, Greece, and Norway. J Clin Microbiol. 2004; 42(11):5146–53. [PubMed: 15528708]
17. Jolley KA, Brehony C, Maiden MC. Molecular typing of meningococci: recommendations for target choice and nomenclature. FEMS Microbiol Rev. 2007; 31(1):89–96. [PubMed: 17168996]
18. Dingle KE, McCarthy ND, Cody AJ, Peto TE, Maiden MC. Extended sequence typing of *Campylobacter* spp., United Kingdom. Emerg Infect Dis. 2008; 14(10):1620–2. [PubMed: 18826829]
19. Taha MK, Hedberg ST, Szatanik M, Hong E, Ruckly C, Abad R, Bertrand S, Carion F, Claus H, Corso A, Enriquez R, Heuberger S, Hryniewicz W, Jolley KA, Kriz P, Mollerach M, Musilek M, Neri A, Olcen P, Pana M, Skoczynska A, Sorhouet Pereira C, Stefanelli P, Tzanakaki G, Unemo M, Vazquez JA, Vogel U, Wasko I. Multicenter study for defining the breakpoint for rifampin resistance in Neisseria meningitidis by rpoB sequencing. Antimicrob Agents Chemother. 2010; 54(9):3651–8. [PubMed: 20606072]
20. Taha MK, Vazquez JA, Hong E, Bennett DE, Bertrand S, Bukovski S, Cafferkey MT, Carion F, Christensen JJ, Diggle M, Edwards G, Enriquez R, Fazio C, Frosch M, Heuberger S, Hoffmann S, Jolley KA, Kadlubowski M, Kechrid A, Kesanopoulos K, Kriz P, Lambertsen L, Levenet I, Musilek M, Paragi M, Saguer A, Skoczynska A, Stefanelli P, Thulin S, Tzanakaki G, Unemo M, Vogel U, Zarantonelli ML. Target gene sequencing to characterize the penicillin G susceptibility of *Neisseria meningitidis*. Antimicrob Agents Chemother. 2007; 51(8):2784–92. [PubMed: 17517841]
21. Schouls LM, van der Ende A, Damen M, van de Pol I. Multiple-locus variable-number tandem repeat analysis of *Neisseria meningitidis* yields groupings similar to those obtained by multilocus sequence typing. J Clin Microbiol. 2006; 44(4):1509–18. [PubMed: 16597884]
22. Elias J, Schouls LM, van de Pol I, Keijzers WC, Martin DR, Glennie A, Oster P, Frosch M, Vogel U, van der Ende A. Vaccine Preventability of Meningococcal Clone, Greater Aachen Region, Germany. Emerg Infect Dis. 2010; 16(3):464–472.
23. Elias J, Harmsen D, Claus H, Hellenbrand W, Frosch M, Vogel U. Spatiotemporal analysis of invasive meningococcal disease, Germany. Emerg Infect Dis. 2006; 12(11):1689–95. [PubMed: 17283618]
24. Achtman M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu Rev Microbiol. 2008; 62:53–70. [PubMed: 18785837]
25. Wirth T, Hildebrand F, Allix-Beguec C, Wolbeling F, Kubica T, Kremer K, van Soolingen D, Rusch-Gerdes S, Locht C, Brisse S, Meyer A, Supply P, Niemann S. Origin, spread and demography of the *Mycobacterium tuberculosis* complex. PLoS Pathogens. 2008; 4(9):e1000160. [PubMed: 18802459]
26. Haensch S, Bianucci R, Signoli M, Rajerison M, Schultz M, Kacki S, Vermunt M, Weston DA, Hurst D, Achtman M, Carniel E, Bramanti B. Distinct Clones of Yersinia pestis Caused the Black Death. Plos Pathogens. 2010; 6(10)
27. Pearson T, Okinaka RT, Foster JT, Keim P. Phylogenetic understanding of clonal populations in an era of whole genome sequencing. Infect Genet Evol. 2009; 9(5):1010–9. [PubMed: 19477301]
28. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. Nat Genet. 2008; 40(8):987–93. [PubMed: 18660809]
29. Baker L, Brown T, Maiden MC, Drobniewski F. Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. Emerg Infect Dis. 2004; 10(9):1568–77. [PubMed: 15498158]
30. Olive DM, Bean P. Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol. 1999; 37(6):1661–9. [PubMed: 10325304]

31. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. Microbiology in the post-genomic era. Nat Rev Microbiol. 2008; 6(6):419–430. [PubMed: 18475305]
32. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One. 2011; 6(7):e22751. [PubMed: 21799941]
33. Vogel U, Szczepanowski R, Claus H, Junemann S, Prior K, Harmsen D. Ion Torrent Personal Genome Machine Sequencing for Genomic Typing of *Neisseria meningitidis* for Rapid Determination of Multiple Layers of Typing Information. J Clin Microbiol. 2012; 50(6):1889–94. [PubMed: 22461678]
34. PubMLST Home Page. http://pubmlst.org/http://pubmlst.org/
35. Jolley KA, Maiden MC. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics. 2010; 11(1):595. [PubMed: 21143983]
36. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012; 50(4):1355–61. [PubMed: 22238442]
37. Inouye M, Conway TC, Zobel J, Holt KE. Short read sequence typing (SRST): multi-locus sequence types from short reads. BMC Genomics. 2012; 13:338. [PubMed: 22827703]
38. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC. Resolution of a meningococcal disease outbreak from whole genome sequence data with rapid web-based analysis methods. J Clin Microbiol. 2012; 50(9):3046–53. [PubMed: 22785191]
39. Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 2004; 21(2):255–65. [PubMed: 14660700]
40. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006; 23(2):254–67. [PubMed: 16221896]
41. Read DS, Woodcock DJ, Strachan NJ, Forbes KJ, Colles FM, Maiden MC, Clifton-Hadley F, Ridley A, Vidal A, Rodgers J, Whiteley AS, Sheppard SK. Evidence for phenotypic plasticity amongst multi-host *Campylobacter jejuni* and *C. coli* lineages using ribosomal MLST and Raman spectroscopy. Appl Environ Microbiol. 2013; 79(3):965–73. [PubMed: 23204423]
42. Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MC. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. Microbiology. 2012; 158(Pt 6):1570–80. [PubMed: 22422752]
43. Ussery DW, Gordon SV. Two novel methods for using genome sequences to infer taxonomy. Microbiology. 2012; 158(Pt 6):1414. [PubMed: 22504434]
44. Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony CM, Colles FM, Wimalarathna HM, Harrison OB, Sheppard SK, Cody AJ, Maiden MC. Ribosomal Multi-Locus Sequence Typing: universal characterisation of bacteria from domain to strain. Microbiology. 2012; 158:1005–15. [PubMed: 22282518]
45. Hao W, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, Alexander DC. Extensive genomic variation within clonal complexes of Neisseria meningitidis. Genome Biol Evol. 2011; 3:1406–18. [PubMed: 22084315]
46. Budroni S, Siena E, Hotopp JCD, Seib KL, Serruto D, Nofroni C, Comanducci M, Riley DR, Daugherty SC, Angiuoli SV, Covacci A, Pizza M, Rappuoli R, Moxon ER, Tettelin H, Medini D. *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc Natl Acad Sci USA. 2011; 108(11):4494–4499. [PubMed: 21368196]
47. Schoen C, Blom J, Claus H, Schramm-Gluck A, Brandt P, Muller T, Goesmann A, Joseph B, Konietzny S, Kurzai O, Schmitt C, Friedrich T, Linke B, Vogel U, Frosch M. Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in *Neisseria meningitidis*. Proc Natl Acad Sci USA. 2008; 105(9):3473–8. [PubMed: 18305155]
48. Katz LS, Humphrey JC, Conley AB, Nelakuditi V, Kislyuk AO, Agrawal S, Jayaraman P, Harcourt BH, Olsen-Rasmussen MA, Frace M, Sharma NV, Mayer LW, Jordan IK. *Neisseria*

Base: a comparative genomics database for *Neisseria meningitidis*. Database (Oxford). 2011; 2011:bar035. [PubMed: 21930505]

49. Rusniok C, Vallenet D, Floquet S, Ewles H, Mouze-Soulama C, Brown D, Lajus A, Buchrieser C, Medigue C, Glaser P, Pelicic V. NeMeSys: a biological resource for narrowing the gap between sequence and function in the human pathogen *Neisseria meningitidis*. Genome Biol. 2009; 10(10):R110. [PubMed: 19818133]
50. Bentley SD, Vernikos GS, Snyder LA, Churcher C, Arrowsmith C, Chillingworth T, Cronin A, Davis PH, Holroyd NE, Jagels K, Maddison M, Moule S, Rabbinowitsch E, Sharp S, Unwin L, Whitehead S, Quail MA, Achtman M, Barrell B, Saunders NJ, Parkhill J. Meningococcal Genetic Variation Mechanisms Viewed through Comparative Analysis of Serogroup C Strain FAM18. PLoS Genet. 2007; 3(2):e23. [PubMed: 17305430]
51. Peng J, Yang L, Yang F, Yang J, Yan Y, Nie H, Zhang X, Xiong Z, Jiang Y, Cheng F, Xu X, Chen S, Sun L, Li W, Shen Y, Shao Z, Liang X, Xu J, Jin Q. Characterization of ST-4821 complex, a unique *Neisseria meningitidis* clone. Genomics. 2008; 91(1):78–87. [PubMed: 18031983]
52. Ribosomal MLST web site. http://rmlst.org/http://rmlst.org/
53. Köser CU, Holden MT, Ellington MJ, Cartwright EJ, Brown NM, Ogilvy-Stuart AL, Hsu LY, Chewapreecha C, Croucher NJ, Harris SR, Sanders M, Enright MC, Dougan G, Bentley SD, Parkhill J, Fraser LJ, Betley JR, Schulz-Trieglaff OB, Smith GP, Peacock SJ. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N Engl J Med. 2012; 366(24): 2267–75. [PubMed: 22693998]
54. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011; 331(6016):430–4. [PubMed: 21273480]
55. Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. Evolution of MRSA during hospital transmission and intercontinental spread. Science. 2010; 327(5964):469–74. [PubMed: 20093474]
56. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, Godfrey P, Haas BJ, Murphy CI, Russ C, Sykes S, Walker BJ, Wortman JR, Young S, Zeng Q, Abouelleil A, Bochicchio J, Chauvin S, Desmet T, Gujja S, McCowan C, Montmayeur A, Steelman S, Frimodt-Moller J, Petersen AM, Struve C, Krogfelt KA, Bingen E, Weill FX, Lander ES, Nusbaum C, Birren BW, Hung DT, Hanage WP. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc Natl Acad Sci USA. 2012; 109(8):3065–70. [PubMed: 22315421]
57. Young BC, Golubchik T, Batty EM, Fung R, Larner-Svensson H, Votintseva AA, Miller RR, Godwin H, Knox K, Everitt RG, Iqbal Z, Rimmer AJ, Cule M, Ip CL, Didelot X, Harding RM, Donnelly P, Peto TE, Crook DW, Bowden R, Wilson DJ. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. Proc Natl Acad Sci USA. 2012; 109(12):4550–5. [PubMed: 22393007]
58. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, Ip CL, Wilson DJ, Didelot X, O'Connor L, Lay R, Buck D, Kearns AM, Shaw A, Paul J, Wilcox MH, Donnelly PJ, Peto TE, Walker AS, Crook DW. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. BMJ Open. 2012; 2(3)

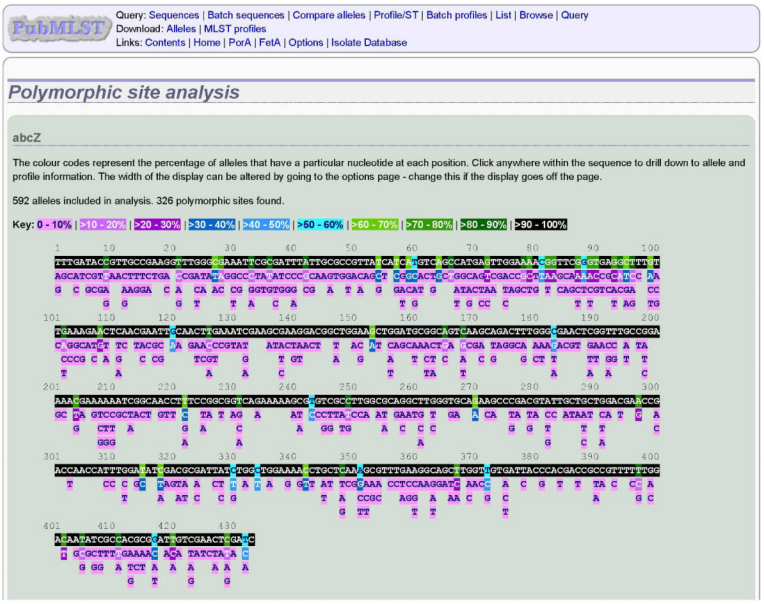

Figure 1.

A schematic of one of the MLST loci showing the number and positions of known polymorphic sites within the gene fragment (unmodified PubMLST.org screenshot).

Europe PMC Funders Author Manuscripts

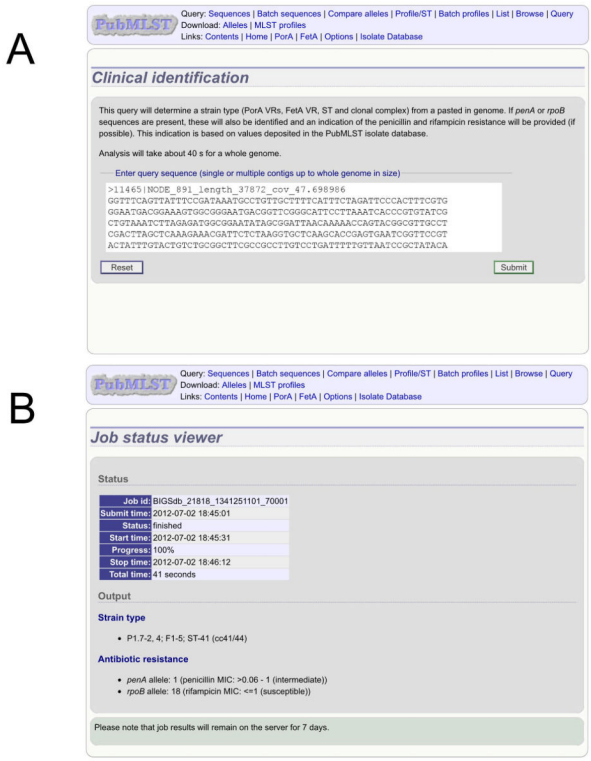

Figure 2.
Extracting antigen and antibiotic resistance data from whole genome sequences. A whole genome sequence, which may consist of multiple contigs, can be pasted in to the *Neisseria* PubMLST website (A) with typing and antibiotic resistance data for penicillin and rifampicin rapidly extracted (B) (unmodified PubMLST.org screenshots).

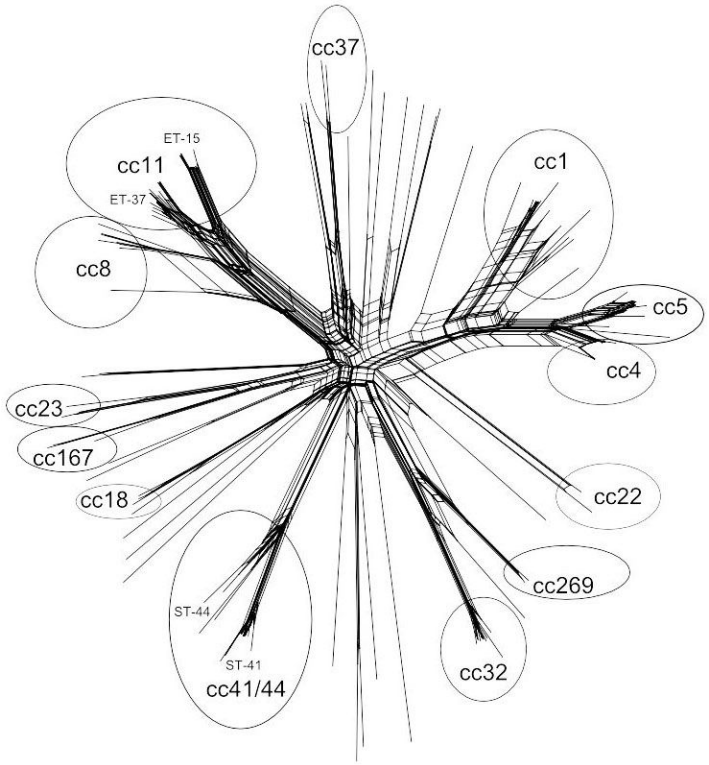

Figure 3.
The relationships of 139 *Neisseria meningitidis* genomes stored in the PubMLST.org *Neisseria* database, generated with Genome Comparator and NeighborNet from allelic profiles data for rMLST loci. The locations of isolates belonging to major clonal complexes identified by conventional MLST are indicated (cc1, etc). The figure illustrates relationships not apparent from sevenlocus MLST, including the diversity of some clonal complexes (e.g. cc1) and the interrelationships of others, e.g. cc8 and cc11 clonal complexes, and the relationships of the 'ET-15 and 'ET-37' variants within cc11.

Europe PMC Funders Author Manuscripts

Table 1

Breakdown of meningococcal WGS data linked to published studies, deposited in the PubMLST *Neisseria* database in November 2012, showing the clonal complex and indicating the diversity of sequence type, serogroup (NG: non-groupable; NA: not available) and typing antigens. Only clonal complexes represented by two or more genomes are included.

Clonal complex	Number of genome sequences	Number of STs	Serogroups	PorA variant combinations	FetA variants
cc11	31	6	C (22), W (4), B(2), NG (1), NA (2)	8	8
cc41/44	20	12	B (14), NA (5), NG (1)	10	5
cc32	17	4	B (14), C (1), NG (1), NA (1)	10	5
cc5	16	5	A (16)	3	5
cc4	14	1	A (14)	4	1
cc1	13	3	A (13)	4	5
cc8	9	5	B (5), C (3), NA (1)	6	5
cc18	5	4	B (4), C (1)	5	4
cc23	5	2	Y (5)	3	2
cc22	4	1	W (4)	1	2
cc167	4	4	Y (4)	1	2
cc269	4	3	B (2), NA (2)	4	3
cc37	2	2	B (2)	1	2