Sequenced-Based Typing of Prokaryotes

Henrik Christensen and John Elmerdahl Olsen

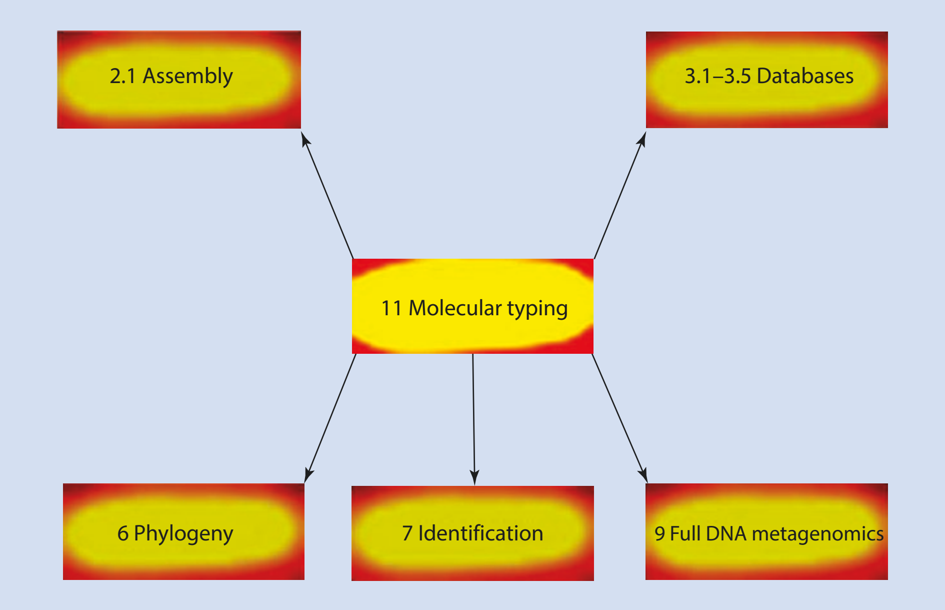
H. Christensen (ed.), *Introduction to Bioinformatics in Microbiology*, Learning Materials in Biosciences,
https://doi.org/10.1007/978-3-319-99280-8_11

What You Will Learn in This Chapter

You will learn some background of why sequence-based typing of microorganisms is performed. You will learn about sequence-based identification and characterization of populations based on multilocus sequence typing (MLST). You will then learn how the MLST concept has been transferred to whole genomic sequence comparison. You will learn about single-nucleotide polymorphism (SNP) analysis based on the whole genomic sequences as well as organism-specific servers enabling the prediction of serotype, antibiotic resistance gene profile and MLST type.

11.1 Background of Prokaryotic Populations and Population Genetics

In this chapter the background of population genetics mainly related to bioinformatics will be presented (◘ Fig. 11.1). Population genetics is the study of the evolutionary change in the genetic composition of populations (Whittam 1995). According to Whittam (1995), population genetics applies both to how the mechanisms (mutation, natural selection, migration, genetic drift) influence the evolutionary rate of change in the populations (◘ Fig. 11.2) as well as to historical investigations of how and when pathogens have evolved. The outcomes of such investigations will be definition of population structures, knowledge of the nature of allelic variation, and the role of different modes of recombination in generating genotypic variation (Milkman 1973; Selander and Levin 1980; Whittam 1995).



◘ **Fig. 11.1** Relationship of the chapter to other chapters of the book. Molecular typing based on sequence comparison relates to many preceding chapters dealing with sequence assembly, databases, phylogeny, and identification. For the future, molecular typing could probably also be linked to full DNA metagenomics

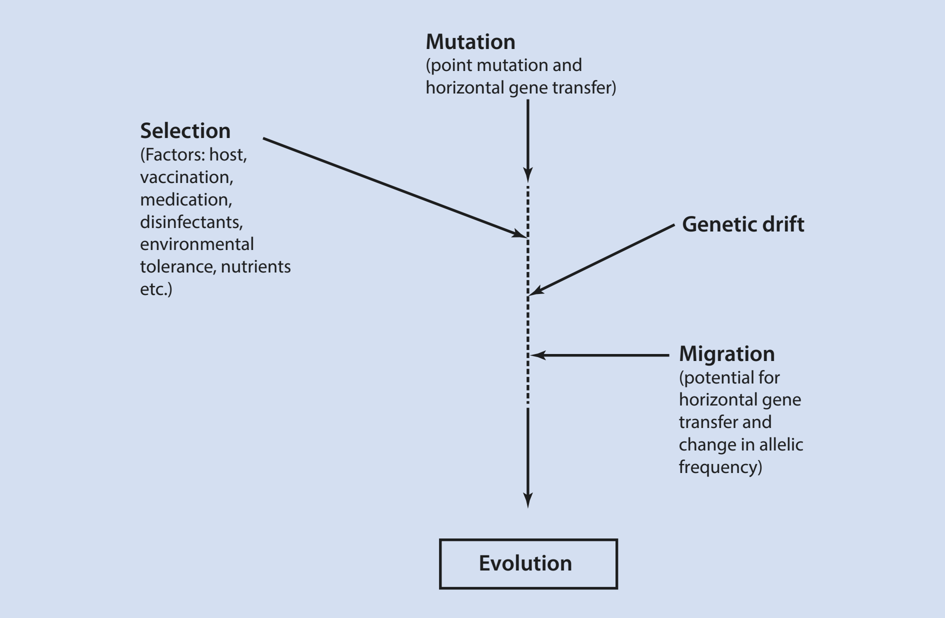
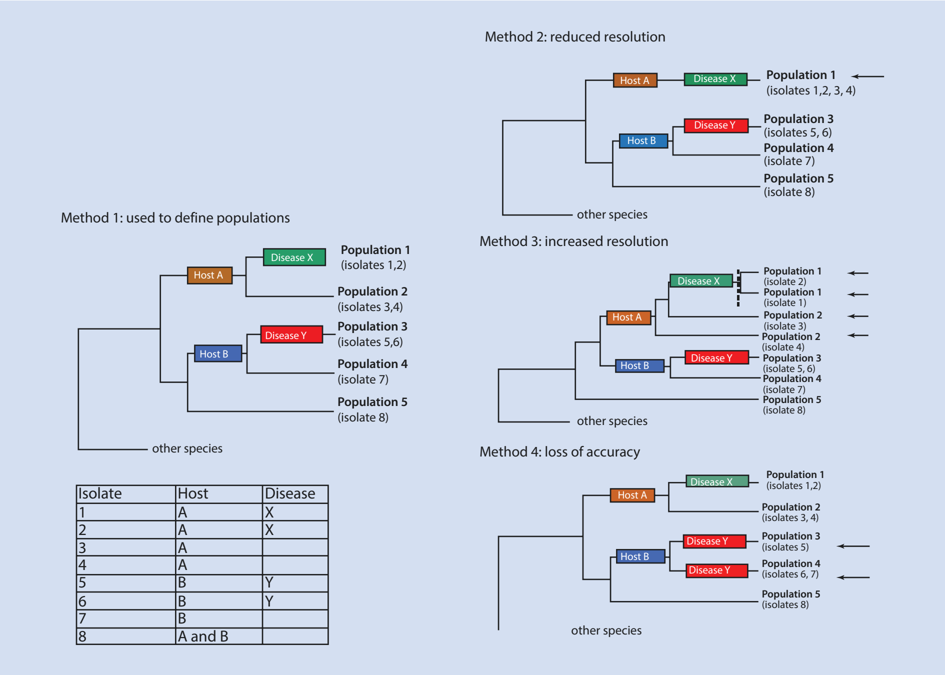


Fig. 11.2 Population genetic mechanisms relevant for the prokaryotes. The mechanisms of selection, mutation, and genetic drift contribute to the evolution of populations by shaping their genetic properties. Migration may also contribute if allelic frequencies are affected

In animals, a population is defined as a group of individuals with a potential for sexual reproduction often limited to a certain geographic region at a certain time point. However, this definition cannot be used for prokaryotes due to the lack of sexual reproduction. Groupings of prokaryotes are better reflected by clonal relationships than physical barriers. A prokaryotic clone was originally defined as a group of prokaryotic isolates "showing so many identical phenotypic and genetic traits that the most likely explanation for this identity is a common origin" (Ørskov and Ørskov 1983). Clonal populations are also called genetic lineages.

The most difficult task with prokaryotes is to set up the limits for the population. A practical solution might be to define the population borders according to the role that their members have in causing disease or other properties useful for epidemiology and for the investigation of pathogenesis. However, populations need to be identified by certain genotyping methods, and their principles are very much determining for how we observe the populations. These methodological problems with identification of clonal populations are illustrated in ▣ Fig. 11.3. Method 1 is optimal in linking genotypic clusters with isolate properties in relation to host and disease association. Method 2 has slightly lower resolution than method 1 and results in the lack of separation of populations 1 and 2. The non-disease-related isolates 3 and 4 are therefore included in this population which is confusing compared to method 1 that allowed the separation of population-associated disease from the nonpathogenic. Method 3 has higher resolution than method 1, and this will not contribute with any error as long as it is taken care of that some clusters might belong to the same population. Method 4 has the same resolution as method 1; however, isolates are clustered differently, and methods 4 will give another interpretation about their role in disease.



Isolate	Host	Disease
1	A	X
2	A	X
3	A	
4	A	
5	B	Y
6	B	Y
7	B	
8	A and B	

Fig. 11.3 Model illustrating dilemmas between methodological and biological relevance of populations. Clonal population are defined according to method 1 providing optimal biological interpretation (hosts A, B, disease associations X, Y) for the isolates investigated in the table

The term "strain" is sometimes misused to equal a population or a clone. In the strict sense, a strain is an isolate that has been further characterized, archived, and documented. The term "strain" refers to the culture or subcultures of it. The misuse is related to the fact that two isolates that share phenotype and genotype in principle must be considered as belonging to the same strain.

11.1.1 Mutation

The most likely starting point for an event affecting the genetics of a population is a mutation (Fig. 11.2). The progeny of a prokaryotic cell should in principle be genetically identical to its ancestor. Two genetic mechanisms, point mutations and horizontal gene transfer (HGT), tend to abolish this identity over time. Mutations will increase genetic diversity between individuals. Point mutations accumulate with a nearly constant rate at random positions in the sequence, and we can analyze them by sequence comparison until a certain level of divergence. HGT, however, is more problematic to analyze, since large fragments of sequences are exchanged between individuals and events of transfer cannot easily be traced since the origin of such fragments are often unknown.

Recombination in prokaryotes has a completely different meaning compared to eukaryotes. In eukaryotes, recombination refers to the results of crossing over in a symmetrical way between chromosomes during the zygotene stage of meiosis. This happens between members of the same species and even with members of the same population.

Recombination in prokaryotes includes HGT, and it can take place between prokaryotes by transformation, transduction, and conjugation and usually occurs as an asymmetrical exchange events between the partners. Some bacteria like *Haemophilus influenzae* are naturally transformable meaning that they can take up DNA directly from the environment. In bacteria with a double-wall structure (Gram negative), DNA is taken up as double stranded across the outer membrane and as a linear single-stranded across the inner membrane, and uptake signal sequences (USS) have been found to favor uptake (Maughan et al. 2008). Transduction is the transfer of genetic material by bacteriophages. There is high variability in the ability of bacteriophages to cross-react with strains within the same species, between different species of the same genus and between different genera of the same family (Jones and Sneath 1970). Conjugation is when genetic material is transferred between two prokaryotes on a conjugational plasmid. This is the most frequent mechanism for horizontal transmission of antibiotic resistance genes.

Luria and Delbruck (1943) pioneered the investigation of prokaryotic population genetics by investigating how bacteria become resistant to bacteriophages. They found resistance to develop as a random process since resistance to bacteriophages developed independent on the presence of bacteriophages in cultures of bacteria sensitive to bacteriophages. Furthermore, they analyzed mutations in a statistical framework and found that the distribution of mutational events followed a special distribution (Luria & Delbruck distribution). Mutations rates are still measured based on the principles laid down by Luria and Delbruck (1943). The most common procedure is the fluctuation test. In this test, the distribution of mutants in a number of parallel cultures is used to estimate the mutation rate based on knowledge of the expected number of mutation events, the number of cultures, and the size of initial inoculum. A range of assumptions are taken in this calculation: constant probability of mutations per cell cycle, that the mutation rate is independent of growth phase, no cell death occur and revertants are not formed, only single mutants arise, wild type and mutants have the same growth rate, a negligible number of mutant cells are initially present compared to final numbers in cultures investigated, and that we are able to detect all mutants (Pope et al. 2008). The mutation rate is the probability of a mutation occurring per cell division. Another measure is the mutation frequency being the proportion of mutant prokaryotes present in a culture. The mutation rate is independent on the age of the prokaryotic culture and a more accurate measure of mutations compared to the mutation frequency that is related to the age of the prokaryotic culture. In principle we should be able to trace mutations directly to the DNA sequence level (Bell 2008). In practice this will require an enormous DNA sequencing effort; however, next-generation sequencing technologies might be able to do this job in order to measure mutation rates directly and not via fluctuation tests.

11.1.2 Selection

Periodic selection is the most prominent population-genetic mechanism of prokaryotes. If a specific allele is favored in the population, and this allele results in higher fitness, then all members of the population with this allele will replace other populations without the allele in the given environment by expansion of this clone. Periodic selection is a form of "bottleneck" (Levin 1981). The higher rate of HGT, the less is the effect of periodic selection (Levin 1981).

The degree of selection acting on a specific coding gene can in theory be predicted by calculation of the so-called dN/dS ratio (Nei 2005). The ratio is the ratio of non-synonymous nucleotide substitutions per non-synonymous site (dN) to that of synonymous nucleotide substitutions per synonymous site (dS). A non-synonymous substitution at DNA level is a change in nucleotide leading to a change in the amino acid translated from the codon where it occurs. It follows that a synonymous substitution is not affecting the amino acid translated from the codon. The theory is based on the "neutral theory": without positive selection, $dN/dS = 1$. Negative selection ($dN/dS < 1$) occurs when deleterious alleles (recognized at amino acid level) are eliminated from the population by purifying selection and leaving only the synonymous changes to be observed. In positive selection ($dN/dS > 1$), polymorphism is assumed to be maintained at the amino acid level, and the changes here will be relatively higher than those observed at synonymous sites. However, high polymorphism at the amino acid level might not necessarily be observed for positive selection to occur (Nei 2005). Textbook examples of positive selection are with peptide-binding sites of MHC genes from humans and mice as well as the antigenic genes of influenzae virus (Nei 2005). To carry out the dN/dS calculations, at least two closely related nucleotide sequences need to be compared pairwise. Computer programs can evaluate the substitutions in regard to all combinations of nucleotides that are legal for each codon. The program DnaSP (Librado and Rozas 2009) (► http://www.ub.edu/dnasp/index_v5.html) can be used to calculate selection and many other parameters from DNA sequence data. MEGA7 (Kumar et al. 2016) introduced in ► Chap. 6 can also calculate the dN/dS parameter and will evaluate if the value is statistically significantly different from neutral.

11.1.3 Genetic Drift

"Genetic drift is a random process that can cause gene frequencies in a population to change over time causing evolution without natural selection" (Madigan et al. 2018). The random distribution of certain genotypes within a population may lead to genetic drift if the population is small. The limit has been suggested as 10^8 cells, and with high population size, this effect may not be important. With low population size, then clones developed by periodic selection may be wiped out more or less by chance. The effect is linked to selection in the way that the weaker and more unpredictable the periodic selection, the higher the effect of genetic drift.

11.1.4 Migration

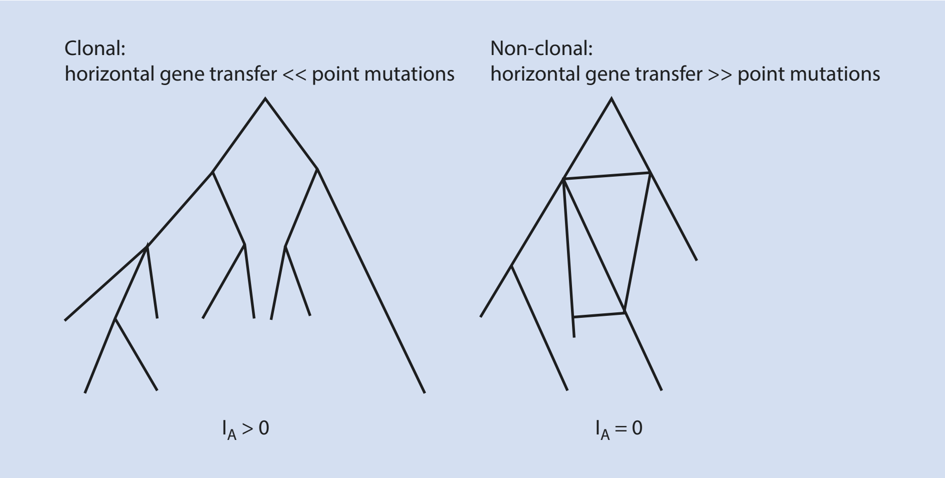
Migration is the assimilation of new individuals into a population from another population. For migration to take effect on evolution, the introduction of individuals should have consequences on population-genetic processes like allelic frequencies and mutation rates. Investigation of the spread of populations including bacterial spread between animals is not population genetics as long as such spread is not affecting the evolution of the organisms. Investigation of such spread is part of epidemiology.

11.1.5 The Biological Consequences of Population Genetics of Prokaryotes

In theory, the population structure of a prokaryote species is determined by the "ratio of genetic changes caused by recombination relative to de novo mutation" (Spratt and Maiden 1999) meaning that if HGT is relatively higher than point mutation rates, the population structure will be very complex and diverse, whereas low degree of recombination relative to point mutation will result in well-defined populations at the sequence level. Only in the last case are we able to recognize clonal populations (◘ Fig. 11.4). The ratio between "recombination and point mutations" can be estimated by comparing recombination between the alleles of the genes used in the MLST (▶ Sect. 11.2) analysis to actual point mutations observed with the same genes.

The distribution of alleles investigated by MLST (▶ Sect. 11.2) can be used to predict if the population structure is clonal or non-clonal (panmictic). With equal assortment of alleles, there should be an equal random distribution of alleles between populations of a species. The expected variance V_E should equal the observed variance V_o. If the populations have evolved like clones, the alleles will be identical or highly related within clones and very different between clones, and V_o will be higher than V_E. To compare V_E and V_o, the index of association is calculated ($I_A = ((V_o/V_E) - 1)$, and it follows that I_A is not significantly different from 0 with a non-clonal (panmictic) population structure but significantly different with a clonal population structure (◘ Fig. 11.4).

Some clonal lineages of a species seem to have adapted to specific hosts. In addition, these lineages often cause disease to a higher extent than other lineages, and for epidemiological investigations, it is therefore of great importance to identify and understand prokaryotes at the population level.


Clonal:
horizontal gene transfer << point mutations
Non-clonal:
horizontal gene transfer >> point mutations
$I_A > 0$
$I_A = 0$

◘ **Fig. 11.4** Population structures are related to the ratio between recombination and point mutations. The expected variance of allele frequency, V_E should equal the observed variance V_O. If the populations have evolved like clones, the alleles will be identical or highly related within clones and very different between clones, and V_O will be higher than V_E. The index of association is calculated as $I_A = ((V_O/V_E) - 1)$

One outcome of population genetics investigations is more realistic diagnostic methods for prokaryotic populations involved in disease. If population structures reflect disease patterns and hosts, we will be able to determine the populations that are really responsible for disease and not their commensal sister groups (◘ Fig. 11.2).

Ecotypes are groups of prokaryotes playing ecological distinct roles defined based on DNA sequences. Ecotypes are analyzed by construction of phylogenetic trees based on housekeeping genes of isolates representing populations of a species. At a certain level of depth of a cluster in the tree, a group of populations that equal an ecotype is defined. Simulations are used to select this level of cutoff for the ecotype as well as to estimate periodic selection and genetic drift. Ecotypes are as a consequence one or more clonal populations if we refer to the current species concept of prokaryotes. Ecotypes can be regarded as species if we redefined the prokaryotic species concept; however, multiple ecotypes are usually recovered within the traditional species (Koeppel et al. 2008).

11.2 Multilocus Sequence Typing (MLST)

11.2.1 MLST

Multilocus sequence typing (MLST) is based on the comparison of DNA sequences of conserved genes in strains of a species (Maiden et al. 1998). For each gene (locus), all different versions of sequences are scored as alleles and designated a random number. This number is not reflecting the quantitative difference between the sequences but just stating that the sequences are different. An allele profile, for instance, 1, 3, 4, 5, 2, 3, 1 is defined as a sequence type (ST). If there only is one difference in the allelic profiles between two STs, these STs are single-locus variants (SLV). This will happen if, for instance, the one ST is 1, 1, 1, 1, 1, 1, 1 and the other 1, 1, 1, 1, 1, 1, 2. Double-locus variants (DLV) share five alleles and have two variants, and as the name indicates then triple-locus variants (TLV) have three different alleles. If a group of STs are linked as SLVs or DLVs, they form a clonal complex (CC). The clonal complexes are named after the ST estimated to be the ancestor of the complex – the founder.

MLST typing has mainly been performed on two servers ► http://www.mlst.net/ and ► https://pubmlst.org/general.shtml, respectively, each dedicated to a series of species. The databases on these servers have been maintained by curators who have uploaded new sequence types and information of isolates. Users are able to upload sequences to the servers and to obtain a sequence type.

ClonalFrame (► http://www.xavierdidelot.xtreemhost.com/clonalframe.htm) (Didelot and Falush 2007) is a program that can analyze population-genetic parameters based on MLST sequence data. ParseCF is used to do the final calculations of the population-genetic parameters that can be extracted from ClonalFrame such as the relative effect of recombination compared to mutation.

11.2.2 Multilocus Sequence Analysis

The information in the DNA sequences compared for MLST can also be directly analyzed by phylogeny (► Chap. 6) and is then called multilocus sequence analysis (MLSA). For MLSA the DNA sequences of the genes (loci) used for each ST are concatenated meaning that they are joined end by end. For instance, if a seven-locus scheme includes partial regions of 500 nt., the concatenated sequence will be 3500 nt. For most of the MLST servers, concatenated datasets can be downloaded for this type of phylogenetic analysis. The information gained from such an analysis differs from the MLST analysis since the phylogeny will show the actual evolutionary diversity of the STs. The MLSA analysis works best for prokaryotes with a clonal population structure (◘ Fig. 11.4). For a true clonal population structure, the phylogeny of all seven genes should be the same (congruent). More frequent horizontal gene transfer will lead to more linkage equilibrium and will result in a different phylogeny for each gene. In this case a concatenation of all genes will make little sense.

11.3 Whole-Genome-Based Typing

The relative low cost of generating whole genomic draft sequences has enabled the use of the most common typing methods based on the direct comparison of the whole genomic sequences. Even for MLST typing which only relies on a few kilo bases of sequence, it is now cheaper to sequence the whole genome and use that to determine the allelic profile than to sequence the seven genes by the traditional Sanger method. For the more common species of clinical importance, servers have become available that can predict serotypes, antibiotic resistance gene profiles, and virulence gene profiles.

11.3.1 Whole Genomic Multilocus Sequence Typing (wgMLST)

Whole genomic MLST (wgMLST) is an extension of the MLST concept to more than seven conserved genes of the species. In principle all conserved gene sequences can be included in such wgMLST scheme. We are aware that the term wgMLST in some other texts is used for the pangenome and cgMLST reserved only for the conserved part of the pangenome. For simplicity we will use wgMLST here and only consider to include the conserved genes of a species in such a scheme. However, if all conserved genes of the species are included, the number of variations in alleles and sequence types will be very high, and there will be a risk of too high resolution of the typing system (◘ Fig. 11.2 method 3). To focus on clones with certain properties, a set of genes can be selected which includes genes predicted to encode for virulence or other important functional factors. A wgMLST typing system can be set up on the BIGSdb platform ► https://pubmlst.org/software/database/bigsdb/ (Jolley and Maiden 2010). This platform can be used both for MLST and for wgMLST. To set up the platform, you will need assistance from a professional system administrator. The platform allows users to upload whole genomic sequences and extract the sequence type. Curators of the database associated with BIGSdb are able to add new sequence types and information of isolates to the database. The platform has, for instance, been set up for various microorganisms: ► http://199.133.98.43/dbase_list.html.

ClonalOrigin is a parallel to ClonalFrame for whole genomic sequences (Didelot et al. 2010). This program should be able to model the source of specific recombination events.

11.3.2 Single-Nucleotide Polymorphisms (SNP)

Single-nucleotide polymorphisms (SNP) are point mutations that only occur as single-nucleotide changes in sequences. The SNP concept will be introduced here in relation to the comparison of raw reads of isolates obtained by high-throughput sequencing to a reference sequence. The reference can be a fully assembled genome of a well-characterized strain of the bacterial species investigated. It is important that this reference is closely related to the isolates that are being typed. The isolates for SNP typing are whole genome sequenced, and the reads are mapped to the reference. All positions where SNPs are identified between the reads and reference are then scored. The positions can be scored as a multiple alignment, and it can be used to construct a phylogeny as described in ▶ Chap. 6. To construct such a phylogeny, many isolates are compared at one time to the same reference.

The analysis can be performed at Center for Genomic Epidemiology (▶ http://www.genomicepidemiology.org). The files with all reads of a genome can be uploaded to ▶ https://cge.cbs.dtu.dk//services/CSIPhylogeny/ (Leekitcharoenphon et al. 2012). A reference genome needs to be defined, and then files with reads each representing isolates for typing can be uploaded. The programs on the server will extract all SNPs identified between the isolates compared to the reference. The SNPs can be downloaded in the form of a multiple alignment, and a phylogeny can be visualized by MEGA7 (Kumar et al. 2016). The benefit with this typing method is that is relatively easy to perform and has high resolution (Schürch et al. 2018). A database is not needed for this type of SNP like for wgMLST. Unfortunately, there are at least two drawbacks with the SNP concept. One is that rates of SNPs differ in different species. When *Listeria monocytogenes* was subcultured from a frozen stock every 3 months during a 3-year period, only one SNP was detected (Kwong et al. 2016) and almost at the same level of SNPs was found in *Mycobacterium tuberculosis* with four SNPs over 4 years, whereas *Helicobacter pylori* accumulated 30 SNPs per year (Schürch et al. 2018). The other problem is that SNP results are only valid in comparison with a specific reference. If another reference is selected, another result is obtained. For species of *Salmonella*, wgMLST was preferred compared to SNP in a large comparative study (Alikhan et al. 2018).

11.3.3 Typing of Virulence, Serotype, and Antibiotic Resistance Based on the Whole Genomic Sequence

This approach is only possible for certain well-characterized species where all or most virulence genes and antibiotic resistance genes have been identified and characterized and where the genetic background for the antigenic profile is well characterized. With this information, a typing system can be established on a server with Internet access. Such a tool is available Center for Genomic Epidemiology (DTU) (▶ http://www.genomicepidemiology.org/) (Larsen et al. 2012; Joensen et al. 2014, 2015). Here the prediction of virulence genes, MLST (multilocus sequence type), and serotype on the basis of the whole genomic sequences can be done for selected prokaryotic species. The whole genomic sequences can be uploaded in different formats including assembled genomes or raw reads.

11.4 Organism-Specific Platforms for Whole-Genome Sequence-Based Typing

SuperPhy (► https://lfz.corefacility.ca/superphy) (Whiteside et al. 2016) is dedicated for genomics investigation of *E. coli* that aims to link genomic information to phenotypic knowledge.

Enterobase is dedicated work with *Escherichia*, *Salmonella*, *Clostridioides* (*Clostridium difficile*), and *Yersinia* (► https://enterobase.warwick.ac.uk/) (Alikhan et al. 2018). At this server the raw Illumina reads can be uploaded and compared to traditional MLST (legacy MLST) and different variants of wgMLST databases. It is also possible only to perform MLST based on genes encoding for the ribosomal proteins (rMLST) (Jolley et al. 2012).

At NCBI, the pathogen detection portal (► https://www.ncbi.nlm.nih.gov/pathogens/) is available for identification and typing of the 22 most common human prokaryotic pathogens.

11.5 Activity

11.5.1 MLST Typing of *Pasteurella multocida*

This species is mainly an important animal pathogen; however, it can also affect humans if bitten by animals. A well-curated MLST scheme has been set up at ► https://pubmlst.org/bigsdb?db=pubmlst_pmultocida_seqdef (Davies et al. 2003; Jolley and Maiden 2010; Subaaharan et al. 2010). Select **Sequence query all loci,** and upload a genome of *P. multocida*. To get a genome, follow the instruction in Activity 3.8.2, and use acc. no. LT906458. The result should be ST13 in the **RIRDC MLST** scheme and ST3 in the **Multiple host scheme**. The reason for two sequence types is that there exist two MLST typing systems for this species. For most species, only one MLST scheme is available.

11.5.2 Graphics

PHYLOViZ (phylogenetic inference and data visualization) is used for visualization of results generated by sequence-based typing (► http://www.phyloviz.net/) (Francisco et al. 2012). It is a free-ware program that can visualize population genetics structures as a system of lines and circles with sizes of circles proportional to number of isolates in each ST (◘ Figs. 11.5 and 11.6). Similar facilities are only available from costly programs like Bionumerics (AppliedMaths) (► http://www.applied-maths.com/bionumerics).

The program can both be used online from the URL and as a downloaded program. Two files are needed, one with the MLST profiles and one with the isolate information (auxiliary data). At least one of the columns in the isolate files needs to be labeled the same way as in the profiles. This can, for example, be “ST” (◘ Fig. 11.5).

The program can be used online, but to obtain graphics for publications, you need to download and install the program from the URL. For Windows, save the zip file to your computer, unzip to a folder on c:, for instance, **c:/phyloviz**. Open the folder and the “**bin**” folder, and press the **phyloviz64.exe** icon which will activate the program.

Profiles file

ST	gene1	gene2	gene3	gene4	gene5	gene6	gene7
1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	2
3	3	1	1	1	1	1	2
4	3	1	1	2	1	1	2
5	3	1	2	1	2	1	1
6	1	2	1	3	2	2	1
7	3	3	2	3	2	2	1
8	3	3	1	2	1	2	3
9	3	3	2	3	2	2	3
10	3	2	2	1	3	2	1

Isolates file (note that the ST column links this file to the Profiles file)

ST	isolate	country	host
1	A1	USA	Mouse
2	B2	Germany	Mouse
3	C4	Australia	Mouse
3	D5	Germany	Mouse
3	E6	Germany	Mouse
4	G2	Australia	Rat
5	H3	Netherlands	Mouse
6	I1	Netherlands	Mouse
7	J6	Netherlands	Mouse
8	K9	Denmark	Mouse
9	L3	Germany	Mouse
9	M7	Netherlands	Mouse
10	N4	Australia	Rat

Fig. 11.5 Input files for PHYLOViZ. This example is showing ten MLST types in the "profiles file" and 13 isolates linked to the STs in the "isolates file"

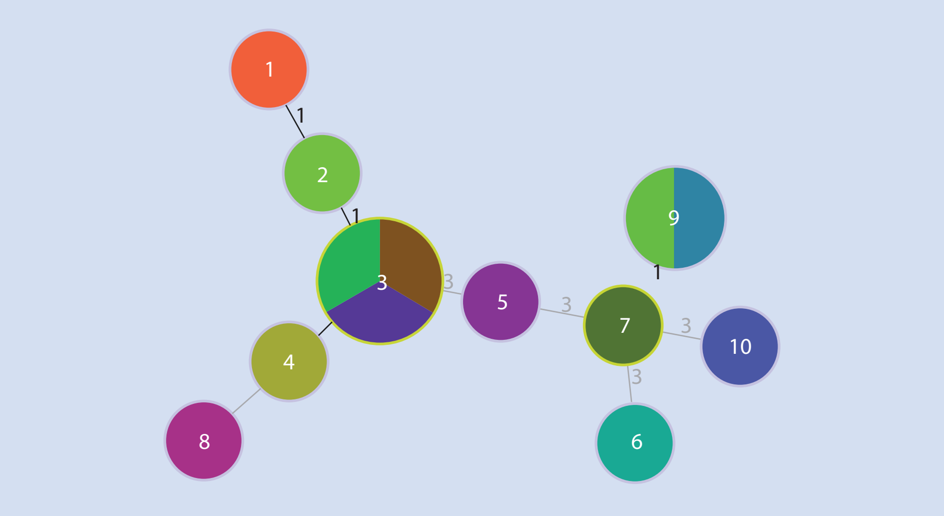


Fig. 11.6 Output from PHYLOViZ based on the data in Fig. 11.5. The nodes are labeled with the STs, and the numbers at the branches are the number of allele differences between STs. The size of nodes reflect the number of isolates in each ST, and the color can be used for a legend indicating the origin of the isolates

Prepare the two input files (◘ Fig. 11.5) with a text editor like **WordPad**. Separate the columns with one tab stroke and save each file in txt format. Start the program, and select **File | Load dataset | Dataset name** where you can select any name. Select dataset type: **MLST**, select **Next,** and upload the profiles file, press Open and **Finish,** and then the isolate files, and again press **Open** and **Finish**. At **Datasets** in the top left corner, you will now see the name of the dataset just uploaded. Double-click on it, and you will see the name of the two files with the isolate and profile (MLST) information. Right-click on the MLST set and select **Compute | geoBURST** and select **Next**. Select the level of links between STs, SLV, DLV, or TRV, for instance, TRV. Click on the key to left to the MLST name which will show the file **geoBURST**. Click on it, and you will see the graphics. Now go the upper left corner of the window, double-click on isolates which will show the table with isolates information, and press **Select** and **View**. Format the information from the **Options** panel below left to the graphics by unselecting **Group,** and select all other options. Use **control** to adjust the line lengths and node sizes. You can arrange the nodes and lines by dragging with the mouse. When done press the pause (||) sign. It should result in the output like ◘ Fig. 11.6. If there are problems, then control the input files with respect to profiles and isolate information, and make sure that columns have been separated by exactly one tab field. If the nodes are not nice pie-shaped, try to open the isolate table from the program, and press **Select** and **View** again. When the graphics is satisfactory, pause the viewer, save the result by clicking on the camera, and select appropriate format, for instance, jpg. Insert the jpg in power point as a picture (◘ Fig. 11.6).

You can now use this setup with the fictive dataset as a learning tool by changing the allelic profiles in the profile file and see the effect on the graphics. If there too few common alleles between the STs (less than 4), they will no longer be linked in the graphic structure.

Take-Home Messages

- Sequence-based typing of prokaryotes is performed to group isolates by common properties like generic genotypic relatedness or specific genotypic properties linked to virulence, serotype, or antimicrobial resistance.
- Multilocus sequence typing (MLST) is based on the comparison of nucleotide sequences of seven or more conserved genes in a set of related isolates.
- In MLST analysis the nucleotide differences within each gene (locus) of the isolates compared is scored as alleles, and a combination of alleles (profile) for all loci is designated a sequence type (ST).
- MLST information can be compared between different laboratories and analyzed in well-curated databases on servers via the Internet.
- Whole genomic multilocus sequence typing (wgMLST) is extending the MLST concept by the analysis of the core genome which may include hundreds of genes.
- In single-nucleotide polymorphism (SNP) analyses of whole genomes, the reads from high-throughput sequencing of isolates are compared to a common references sequence allowing a very detailed comparison at the single-nucleotide level in order to trace single isolates.
- Typing of virulence, serotype, and antibiotic resistance based on the whole genomic sequence can be performed for certain, mainly human pathogenic bacteria like *Escherichia coli* and *Salmonella enterica*, on dedicated servers that can be accessed via the Internet.

References

Alikhan, N.-F., Zhou, Z., Sergeant, M.J. & Achtman, M. (2018) A genomic overview of the population structure of *Salmonella*. PLoS Genet 14: e1007261.

Bell, G. 2008. Selection. The mechanism of evolution. 2nd ed. Oxford Univ. Press.

Davies, R.L., MacCorquodale, R., Baillie, S., Caffrey, B., 2003. Characterization and comparison of Pasteurella multocida strains associated with porcine pneumonia and atrophic rhinitis. J. Med. Microbiol. 52, 59–67.

Didelot X, Falush D. 2007. Inference of bacterial microevolution using multilocus sequence dta. Genetics 175:1251–66.

Didelot X, Lawson D, Darling A, Falush D. 2010. Inference of homologous recombination in bacteria using whole-genome sequences. Genetics. 186:1435–1449.

Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carriço JA. 2012. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics 13:87.

Joensen, K.G., Scheutz, F., Lund, O., Hasman, H., Kaas, R.S., Nielsen, E.M., Aarestrup, F.M., 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J. Clin. Microbiol. 52, 1501–1510.

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., Scheutz, F., 2015. Rapid and easy in silico serotyping of Escherichia coli using whole genome sequencing (WGS) data. J. Clin. Microbiol. 5, 2410–2426.

Jolley, K. A. & Maiden, M. C. J. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics **11:595**

Jolley, K.A., Bliss, C.M., Bennett, J.S., Bratcher, H.B., Brehony, C., Colles, F.M., Wimalarathna, H., Harrison, O.B., Sheppard, S.K., Cody, AJ. & Maiden, M.C. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158, 1005–1015.

Jones, D. and Sneath, P. H. 1970. Genetic transfer and bacterial taxonomy. Bacteriol Rev. 34, 40–81.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM, Rooney AP, Brambilla E, Connor N, Ratcliff RM, Nevo E, Cohan FM. 2008. Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics. Proc Natl Acad Sci U S A. 105: 2504–9.

Kwong, J.C., Mercoulia, K., Tomita, T., Easton, M., Li, H.Y., Bulach, D.M., Stinear, T.P., Seemann, T., Howden, B.P., 2016. Prospective whole-genome sequencing enhances national surveillance of Listeria monocytogenes. J Clin Microbiol. 54, 333–342.

Larsen, M.V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M., Lund, O., 2012. Multilocus sequence typing of total genome sequenced bacteria. J. Clin. Microbiol. 50, 1355–1361.

Leekitcharoenphon, P., Kaas, R.S., Thomsen, M.C.F., Friis, C., Rasmussen, S., Aarestrup, F.M., 2012. snpTree-a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics 13 Suppl 7:S6

Levin, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure of E. coli populations. Genetics 99, 1–12.

Librado P & Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25:1451–1452.

Luria, S.E. & Delbrück M. 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28, 491–511.

Madigan, M. T., Bender, K. S., Buckley, D. H., Sattley, W. M. & Stahl, D. A. 2018. Brock biology of microorganisms. 15th ed. Global Edition. Pearson, New York.

Milkman R. 1973. Electrophoretic variation in Escherichia coli from natural sources. Science182, 1024–6.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. & Spratt, B.G. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 95, 3140–3145.

Maughan, H., Sinha, S., Wilson L. and Redfield, R. 2008. Competence, DNA Uptake and Transformation in the Pasteurellaceae. In Kuhnert, P and Christensen, H. Pasteurellaceae, Biology, genomics and molecular aspects. Caister Acad. Press. In press.

Nei, M. (2005). Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22, 2318–42.

Pope et al. 2008. A practical guide to measuring mutation rates in antibiotic resistance. Antimicrob. Agents Chemotherapy 52, 1209–14.

Schürch, A.C., Arredondo-Alonso, S., Willems, R.J.L., Goering, R.V. 2018. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. Clin. Microbiol. Infect. 24, 350–354.

Selander RK, Levin BR. Genetic diversity and structure in Escherichia coli populations. 1980. Science 210, 545–7.

Spratt, B. G., M. C. J. Maiden (1999): Bacterial population genetics, evolution and epidemiology. Phil. Trans. R. Soc. Lond. B, 354, 701–10.

Subaaharan, S., Blackall, L.L., Blackall, P.J., 2010. Development of a multi locus sequence typing scheme for avian isolates of Pasteurella multocida. Vet. Microbiol. 141, 354–361.

Whiteside MD, Laing CR, Manji A, Kruczkiewicz P, Taboada EN, Gannon VP. 2016. SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*. BMC Microbiol. 16:65.

Whittam, T. S. 1995. Genetic population structure and pathogenicity in enteric bacteria. In Population genetics of bacteria. Eds. Baumberg, S., Young, J. P. W., Wellington, E. M. H. and Saunders, J. R. 1995. 52nd Symposium of the Society for General Microbiology. pp. 217–245. Cambridge Univ. Press.

Ørskov, F., I. Ørskov (1983): Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the Enterobacteriaceae and other bacteria. J. Infect. Dis. 148, 346–357.