Clinical Microbiology and Infection 24 (2018) 342–349

Contents lists available at ScienceDirect

Clinical Microbiology and Infection

journal homepage: www.clinicalmicrobiologyandinfection.com

Review

A primer on microbial bioinformatics for nonbioinformaticians

J.A. Carriço [1, *], M. Rossi [2], J. Moran-Gilad [3, 4, 5], G. Van Domselaar [6, 7], M. Ramirez [1]

[1] Instituto de Microbiologia, Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal
[2] Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland
[3] Department of Health Systems Management, Faculty of Health Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[4] Public Health Services, Ministry of Health, Jerusalem, Israel
[5] ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD), Basel, Switzerland
[6] National Microbiology Laboratory, Public Health Agency of Canada, 1015 Arlington St, Winnipeg, MB, R3E 3R2, Canada
[7] Department of Medical Microbiology and Infectious Diseases, University of Manitoba, 745 Bannatyne Avenue, Winnipeg, MB, R3E 0J9, Canada

ARTICLE INFO

Article history:
Received 15 June 2017
Received in revised form
13 November 2017
Accepted 22 December 2017
Available online 5 January 2018

Editor: P.T. Tassios

Keywords:
Bioinformatics software
Genomic epidemiology
High-throughput sequencing
Microbial bioinformatics
Microbial typing

ABSTRACT

Background: Presently, the bottleneck in the deployment of high-throughput sequencing technology is the ability to analyse the increasing amount of data produced in a fit-for-purpose manner. The field of microbial bioinformatics is thriving and quickly adapting to technological changes, which creates difficulties for nonbioinformaticians in following the complexity and increasingly obscure jargon of this field.
Aims: This review is directed towards nonbioinformaticians who wish to gain understanding of the overall microbial bioinformatic processes, from raw data obtained from sequencers to final outputs.
Sources: The software and analytical strategies reviewed are based on the personal experience of the authors.
Content: The bioinformatic processes of transforming raw reads to actionable information in a clinical and epidemiologic context is explained. We review the advantages and limitations of two major strategies currently applied: read mapping, which is the comparison with a predefined reference genome, and *de novo* assembly, which is the unguided assembly of the raw data. Finally, we discuss the main analytical methodologies and the most frequently used freely available software and its application in the context of bacterial infectious disease management.
Implications: High-throughput sequencing technologies are overhauling outbreak investigation and epidemiologic surveillance while creating new challenges due to the amount and complexity of data generated. The continuously evolving field of microbial bioinformatics is required for stakeholders to fully harness the power of these new technologies. **J.A. Carriço, Clin Microbiol Infect 2018;24:342**

Introduction

Bioinformatics is an interdisciplinary research field that applies methodologies from computer science, applied mathematics and statistics to the study of biological phenomena. In microbiology, the term is now becoming ubiquitous in many research fields, mainly due to the widespread use and continuous development of high-throughput sequencing (HTS) technologies, also referred to as next-generation sequencing (NGS). These technologies are often referred to as whole genome sequencing (WGS) as a result of their ability to sequence the whole genome in a single process. HTS technologies are critically and irreversibly changing the way microbial DNA is analysed, replacing traditional, targeted molecular methods for microbial detection and typing with an entire draft genome. The term ‘draft genome’ is commonly used because the new technologies that use short reads do not generate a single closed genome but rather a series of sequences (contigs) that may cover up to 95% to 99% of the strain genome.

Indeed, these new sequencing technologies have combined classical microbial typing with modern genomics and created the new field of genomic epidemiology. In genomic epidemiology studies, rather than using a limited number of marker genes or surrogate genetic markers, researchers can directly analyse genome-scale data to better distinguish and more confidently establish the relationships between bacterial strains as well as

[*] Corresponding author. J.A. Carriço, Instituto de Microbiologia, Faculdade de Medicina, Universidade de Lisboa, Av. Professor Egas Moniz, 1649-028, Lisboa, Portugal.
E-mail address: jcarrico@fm.ul.pt (J.A. Carriço).

https://doi.org/10.1016/j.cmi.2017.12.015
1198-743X/

obtain detailed insights into the genetic causes of strain diversity, as well as their potential impact on key traits such as virulence and antimicrobial resistance. To that end, specialized software exists that can be used standing alone or organized into pipelines, where multiple specialized software subcomponents are combined to carry out complex serial or parallel analyses, enabling the automated analysis of large numbers of samples.

Here we aim to provide a primer for nonbioinformaticians focusing on the methods that are currently being applied in clinical and public health microbiology for the study of bacterial infectious diseases. Our goal is to help clinical microbiologists, infectious disease specialists, epidemiologists and other healthcare professionals to navigate this continuously evolving field and to understand its full potential while acknowledging its current limitations but without being overwhelmed by the field's jargon. We will review the concepts needed to understand bioinformatics analyses, starting with the raw data that are generated by the sequencers, focusing on noncommercial, freely available software. Table 1 provides the web links for all the mentioned software. We will not discuss the simultaneous sequencing of genomes from multiple strains or species (metagenomics). Further details on the technology and its general application can be found in other articles in this thematic issue.

HTS basic concepts

The term ‘high-throughput sequencing’ refers to modern technologies that allow the decoding of thousands to millions of DNA sequence fragments at the same time during a sequencing run. A wide variety of sequencers are commercially available, which may be classified into two main types: short-read sequencers, which generate sequence fragments up to around 800 bases, and long-read sequencers, which can generate sequence fragments of up to approximately 100 kilobases. At the time of writing, most HTS data available at the European Nucleotide Archive was short-read data generated from sequencers manufactured by Illumina.

Most whole genome sequences are generated using a shotgun sequencing strategy: many copies of the source DNA are randomly fragmented, and one or both ends of the fragment are sequenced; these sequenced fragments are referred to as reads. The fragments can be sequenced on one end (single-end sequencing) or on both ends. If the reads from both ends span the entire fragment, it is called paired-end sequencing.

Most modern sequencers output their sequencing data in FASTQ files, which is a file format that stores the sequence and associated quality scores—called Phred scores—for each predicted nucleotide, or base call, contained in each read. Phred scores provide a measure of the accuracy of each base call.

The first analysis steps should focus on quality assessment and control. Software such as “FASTQC” is used to evaluate multiple read quality statistics, including the Phred score of each position for all reads. At this point, depending on the intended analysis, the reads are sometimes cleaned or trimmed by removing low-quality regions as well as the sequences inserted during library preparation for read identification (bar coding) and adsorption to the sequencing surface. “TRIMMOMATIC” [1] is a popular choice for this task. Contamination detection of the genomic data from sources other than the intended strain can also be performed at this stage. “Kraken” [2] can be used to rapidly assign reads to their taxonomic origin (i.e. probable genus or species), which can be useful for detecting potential contaminants. After filtering and removal of contaminating reads, the remaining reads can be used to estimate sequence coverage or sequence depth. Coverage refers to the average number of times each nucleotide position in the strain's genome has a read that aligns to that position. Depending on the study goals, bacterial species and the intended analyses, the optimal coverage varies. In public repositories, most FASTQ files have coverages normally ranging from 15 × to 500 × . While the typical downstream analysis will involve assembling the reads into a draft genome, there are also bioinformatics tools that can be applied directly on raw reads without assembly.

De novo assembly

De novo assembly refers to the bioinformatics process whereby reads are assembled into a draft genome using only the sequence information of the reads. This would be akin to putting together a puzzle without advance knowledge of the final picture; the puzzle must be solved exclusively on the basis of the shapes of the tiles. This type of analysis is useful when no appropriate reference genome is available for comparison, to find novel genetic content or to study structural variation. Software for *de novo* assembly uses computationally efficient algorithms to look for overlapping reads and extend them into longer contiguous sequences (contigs). *De novo* assemblies rarely result in completely assembled genomes; instead, they typically generate draft genomes comprising several contigs (tens to hundreds), largely as a result of the presence of large stretches of identical genomic sequences distributed throughout the genome that cannot be extended using the *de novo* approach. Fig. 1 illustrates the basic sequencing concepts from bacterial genome to contigs.

It is now increasingly common to use methodologies that generate very long reads (10–50 kb or longer) (e.g. Pacific Biosciences or Oxford Nanopore Technologies). A hybrid assembly methodology can then make use of long and short reads together to improve the assembly, resulting in fewer and longer contigs, or even a completely closed genome.

The output of the assemblers is usually provided in FASTA format [3], which is a simple file format containing an identifier and nucleotide sequence of each contig. The quality of the assembly can be assessed by multiple parameters such as the number of contigs produced, size of the assembled genome or average length of the contigs in the draft genome. A more complex but commonly reported statistic is the N50 value, which is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly.

Many software tools are available for performing *de novo* assembly, and several comparisons have been reported elsewhere [4–6]. Specialized software for *de novo* assembly quality check is available, such as “QUAST” [7]. Currently, there are few pipelines such as “INNUca” and “shovill” that automatically perform all steps from reads to a curated draft genome.

Read mapping and single nucleotide variant/single nucleotide polymorphism calling

If a closely related reference sequence is available, a read mapping approach can be used to assist in identifying the differences between it and the newly generated sequence. Mapping refers to the alignment of each read to a position on a reference genome. For each reference nucleotide, multiple reads are typically aligned, and the correct base for the position is inferred from the consensus nucleotide derived from the overlapping reads. Dozens of read mappers are currently available [8]. Popular read mappers for bacterial analysis include “bowtie2” [9] and “BWA” [10]. The read mappers generate pileups that record the position of each read on the provided reference and an alignment score for each read; pileups are often reported in the Sequence Alignment Map (SAM) file format [11]. The SAM file can then be processed to call the single nucleotide variant sites (SNVs), also commonly referred to as single

Table 1
List of bioinformatics software used for microbial bioinformatics data analysis

Usage	Software name	Description	URL
Quality measures and read preprocessing	FASTQC	Toolbox for displaying sequence statistics for next-generation sequencing reads	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
	TRIMMOMATIC	Command-line based tool for trimming of short-read paired-end and single-ended data	http://www.usadellab.org/cms/?page=trimmomatic
	FASTX-Toolkit	A collection of command line tools for preprocessing of short-read FASTA/FASTQ files	http://hannonlab.cshl.edu/fastx_toolkit/
	PRINSEQ	Command-line and web-based tool for filtering, reformatting, or trimming genomic and metagenomic sequence data, generates summary statistics in graphical and tabular format	http://prinseq.sourceforge.net/, http://edwards.sdsu.edu/cgibin/prinseq/prinseq.cgi
Contamination detection	Kraken	Taxonomic assignment of reads, useful for metagenomics analysis or detection of contamination in pure culture samples	https://ccb.jhu.edu/software/kraken/
	MIDAS	Taxonomic assignment of reads, useful for metagenomics analysis or detection of contamination in pure culture samples	https://github.com/snayfach/MIDAS
Assembly software and pipelines	Velvet	*De novo* genomic assembler specially designed for short reads	http://github.com/dzerbino/velvet/tree/master
	SPAdes	*De novo* genomic assembler for short reads; it can also provide hybrid assemblies using long-read data together with short-read data	http://cab.spbu.ru/software/spades/
	Canu	*De novo* genomic assembler designed for high-noise single-molecule sequencing such as long reads	http://github.com/marbl/canu
	INNUca	A standardized, fully automated, flexible, portable and pathogen-independent pipeline for bacterial genome assembly and quality control starting from short reads	http://github.com/INNUENDOCON/INNUca
	shovill	A pipeline for bacterial genome assembly which improves SPAdes speed and accuracy	https://github.com/tseemann/shovill
In silico typing	ReMatCh	Software for variant calling based on a read-mapping strategy to selected target sequences; also interacts with European Nucleotide Archive (ENA) repository, easily mining publicly available data	http://github.com/B-UMMI/ReMatCh
	Short Read Sequence Typing for Bacterial Pathogens (SRST2)	It uses short-read data, MLST database and/or database of gene sequences (e.g. resistance genes, virulence genes) and reports the presence of STs and/or reference genes	http://github.com/katholt/srst2
	Microbial InSilico Typer (MIST)	Rapid generation of *in silico* typing data (e.g. MLST, MLVA) from draft bacterial genome assemblies	http://bitbucket.org/peterk87/microbialinsilicotyper
	SISTR	A web- and command line–accessible tool for *Salmonella* typing using draft genome assemblies	http://lfz.corefacility.ca/sistr-app/
	SeqSero	A web-accessible tool for *Salmonella* typing using raw reads or draft genome assemblies	http://www.denglab.info/SeqSero
	RGI-CARD	Curated collection of antimicrobial resistance gene and mutation sequences, bioinformatics models and tools for their detection in bacterial genomes	http://www.card.mcmaster.ca/analyze/rgi
	ResFinder	A web-accessible tool for the detection of acquired antimicrobial resistance genes in bacterial genomes using raw reads or draft genome assemblies	https://cge.cbs.dtu.dk/services/ResFinder/
	VirulenceFinder	A web-accessible tool for the detection of virulence associated genes in *Escherichia coli*, *Listeria* spp., *Staphylococcus aureus*, *Enterococcus* spp. using raw reads or draft genome assemblies	https://cge.cbs.dtu.dk/services/VirulenceFinder/
	MLST1.8	A web-accessible tool for the determination of MLST types from bacterial genomes using publicly available MLST schemas	https://cge.cbs.dtu.dk/services/MLST
	Mlst2.9	Command line–based software which can extract MLST from bacterial genomes using publicly available MLST schemas	https://github.com/tseemann/mlstCFSANSNP
	CFSAN SNP Pipeline	Pipeline for extracting high quality SNV matrices for sequences from closely related pathogens	http://snppipeline.readthedocs.io/en/latest/
	Snippy	A pipeline for rapid identification of haploid variants and construction of phylogeny using core genome SNPs	http://github.com/tseemann/snippy
	SNVPhyl (Single Nucleotide Variant PHYLogenomics)	Pipeline for identifying SNV within a collection of microbial genomes and constructing a phylogenetic tree	http://snvphyl.readthedocs.io/en/latest/
	Lyve-SET	A pipeline for using high-quality SNPs to create a phylogeny, especially for outbreak investigations	https://github.com/lskatz/lyve-SET
Gene-by-gene approaches	BIGSdb	Web-accessible database system designed to store and analyse linked phenotypic and genotypic information, including allele calling engine for gene-by-gene approach; it is the database system for both PubMLST and PasteurMLST	https://github.com/kjolley/BIGSdb, http://pubmlst.org http://bigsdb.pasteur.fr/index.html
	Enterobase	Curated database and online resource for molecular typing of *Salmonella, Escherichia coli, Yersinia* spp. and Moraxella spp. using gene-by-gene approach	http://enterobase.warwick.ac.uk/
	Genome Profiler	Stand-alone gene-by-gene allele calling algorithm which uses conserved gene neighbourhoods to resolve gene paralogy	http://sourceforge.net/projects/genomeprofiler/
	chewBBACA	A comprehensive and highly efficient stand-alone gene-by-gene allele calling algorithm based on coding DNA sequences, including suite of tools for providing overview of schema performance	https://github.com/B-UMMI/chewBBACA

Table 1 (*continued*)

Usage	Software name	Description	URL
Gene annotation	Prodigal	Protein-coding gene prediction software tool for bacterial and archaeal genomes	http://github.com/hyattpd/prodigal/wiki
	Prokka	Quick functional annotation of bacterial genomes producing standards-compliant output file	http://github.com/tseemann/prokka
	RAST	Fully automated service for annotating bacterial and archaeal genomes	http://rast.nmpdr.org/
	MicroScope	Comprehensive analytical platform for genome annotation and analysis of bacterial genomes	http://www.genoscope.cns.fr/agc/microscope/home/index.php
	NCBI prokaryotic genome annotation pipeline (PGAP)	Automatic prokaryotic genome annotation pipeline that combines ab initio gene prediction algorithms with homology-based methods	https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
	NCBI Pathogen Detection	An online platform for sharing and comparing data on outbreak strains; currently contains databases for 20 bacterial species, focusing on food-borne pathogens and healthcare-associated infections	https://www.ncbi.nlm.nih.gov/pathogens/
Genome alignments	Harvest	A suite of core genome alignment and visualization tools for quick and high-throughput analysis of intraspecific bacterial genomes	http://harvest.readthedocs.io/en/latest/
	Mauve	Aligner for comparative analysis of full bacterial genomes	http://darlinglab.org/mauve/mauve.html
Homology clustering and Association studies	Roary	High speed stand-alone pan-genome pipeline for bacterial genomes	http://sanger-pathogens.github.io/Roary/
	Scoary	Pan-genome–wide association studies using Roary output	https://github.com/AdmiralenOla/Scoary
	Neptune	Software designed for detecting genomic signatures within bacterial populations	https://github.com/phac-nml/neptune
Phylogenetic inference	RAxML	Sequential and parallel maximum-likelihood phylogeny estimation that operates on nucleotide and protein sequence alignments	https://sco.h-its.org/exelixis/software.html
	FastTree	Compute approximately maximum likelihood phylogenetic trees from large nucleotide or protein multiple sequence alignments	http://www.microbesonline.org/fasttree/
	Gubbins	Compute maximum likelihood from alignment after removing regions containing elevated densities of base substitutions	https://github.com/sangerpathogens/gubbins
	ClonalFrameML	A maximum likelihood implementation of ClonalFrame designed for genomes sequences	https://github.com/xavierdidelot/ClonalFrameML
	PHYLOViZ	Online Web-based tool for phylogenetic inference, visualization, analysis and sharing of sequence-based typing methods that generate allelic profiles and associated epidemiologic data	http://online.phyloviz.net
	PHYLOViZ 2.0	Stand-alone Java software for phylogenetic inference, visualization and analysis of sequence-based typing methods that generate allelic profiles and their associated epidemiologic data	http://www.phyloviz.net/
Visualization tools	Microreact	A web-based tool for genomic epidemiology data visualization and sharing	http://microreact.org
	Phandango	Interactive web-based tool for fast exploration of large-scale population genomics data sets combining output from multiple genomic analysis methods	https://github.com/jameshadfield/phandango
	iTOL	Web-based tool for display, annotation and management of phylogenetic trees	http://itol.embl.de/
	GenGIS 2	Application including 3-D graphical and Python interfaces allowing users to combine digital map data and sequences	http://kiwi.cs.dal.ca/GenGIS/Main_Page
Multipurpose analytical platforms and pipelines	Centre for Genomic Epidemiology Toolbox	A suite of web-based tools and service for pathogen molecular typing, genome assembly, phenotypic prediction (e.g. resistance prediction) and phylogeny construction	http://cge.cbs.dtu.dk/services/
	Integrated Rapid Infectious Disease Analysis (IRIDA) Platform	A Galaxy-based platform for real-time infectious disease outbreak investigation using genomic data including a sequence data management module and workflows, ontology framework (GenEpiO) and data visualization tools	https://irida.corefacility.ca/documentation/downloads/index.html, http://irida.ca/
	Integration genomics in surveillance of food-borne pathogens (INNUENDO) platform	A platform for real-time disease outbreak investigation and surveillance of food-borne pathogens using genomic data including sequence-data management module, assembly modules with QA/QC measures, gene-by-gene analytical pipeline, ontology framework (GenEpiO) and visualization tools	https://github.com/INNUENDOCON/INNUENDO_platform
	Nullarbor	A pipeline for generating public health microbiology reports from sequenced isolates including sequencing specifics, species ID, subtypes and core SNP	http://github.com/tseemann/nullarbor

MLST, multilocus sequence typing; MLVA, multiple-locus variable-number tandem repeat analysis; QA/QC, quality assurance/quality control; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; ST, sequence type.

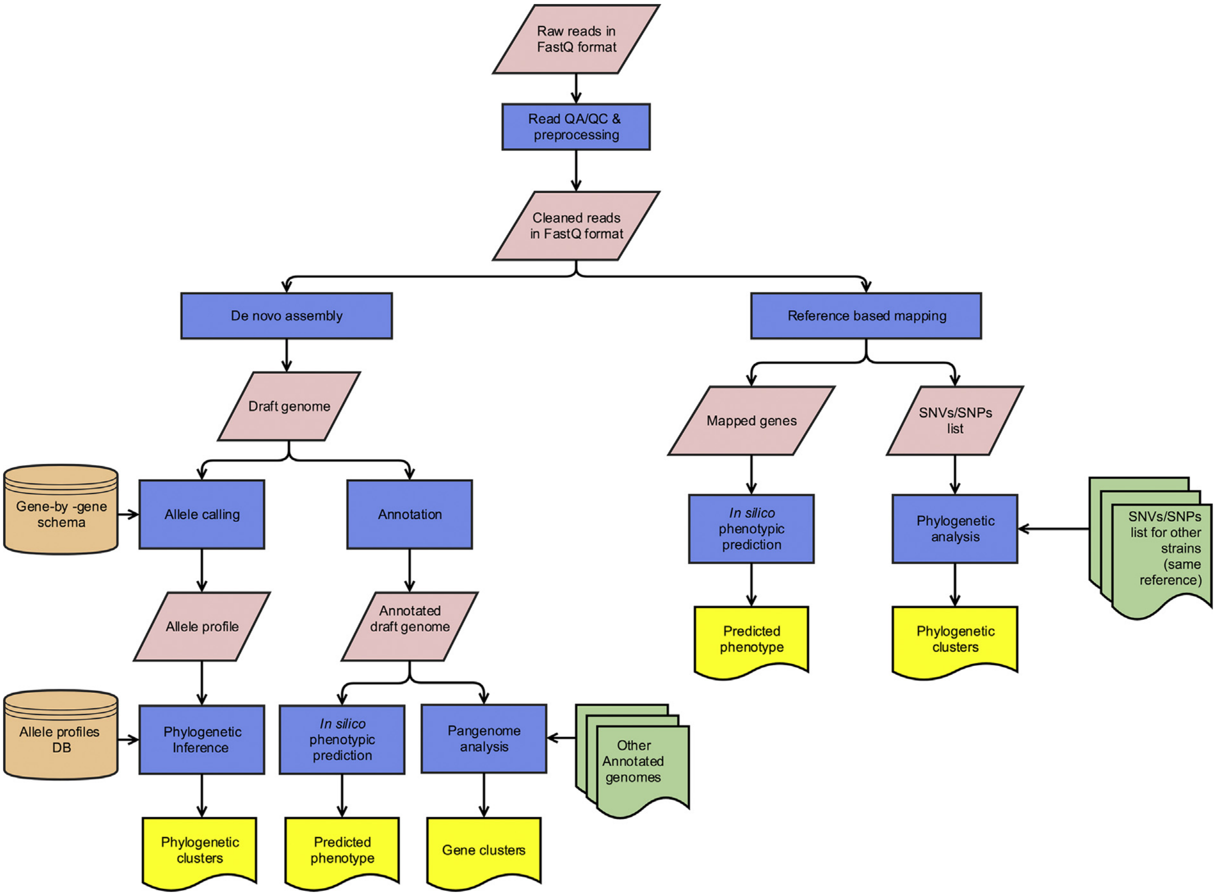

Fig. 1. Commonly used HTS analysis workflows.

nucleotide polymorphisms (SNPs), using software such as “SAMTOOLS”, “BCFTOOLS” [11] or “Freebayes” (https://arxiv.org/abs/1207.3907). The SNVs are usually reported in a Variant Calling Format (VCF) file, which contains information about SNVs and possible short insertions/deletions (indels) compared to the reference. It is important to note that the choice of reference sequence can have a major impact on the variation reported by the software. For example, if reads from a distantly related organism are compared against a reference, only the similar regions will be mapped; any unique or sufficiently diverged sequence data will not map and thus will be omitted from the variant report. The more distant the reference sequence used, the more regions will likely be excluded from the analysis. In the literature, when authors refer to SNP calling methods, they usually refer to the whole process, from read mapping and pileup generation, i.e. the generation of the consensus sequence from the aligned reads, to the final SNV call.

Common microbial bioinformatics analyses

The creation of a draft genome or mapping variant report are called secondary analysis steps [12] because at this point the researcher normally still lacks any actionable information on the sample analysed. The final analysis stage, called tertiary analysis, is the sense-making stage. The specific nature of the tertiary analysis is highly variable and depends on the goal of the study. We discuss here only the most commonly performed analysis workflows for microbial genome analysis; a general scheme depicting these workflows is provided in Fig. 2.

Draft genome analysis

Genome annotation is the process of identifying the location and biological role of genetic features present in a DNA sequence. It is typically one the first steps applied after assembly of a new draft genome. This process involves software pipelines that use multiple external feature prediction algorithms allowing the identification of genetic features such as protein coding sequences, transfer RNA genes (tRNAs), ribosomal RNA genes (rRNAs) and occasionally higher-order features such as operons, CRISPR elements and genomic islands. Several freely available online web services provide genome annotation: the National Center for Biotechnology Information (NCBI) prokaryotic genome annotation pipeline (PGAP) [13], a service integrated with the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/genome/annotation_prok/), can annotate any submitted assembly file and has a turnaround time of days; the “RAST Server” [14] is a well-known web service which can provide an annotated genome within 12 to 24 hours of submission. A faster option, given the availability of local computing power, is “PROKKA” [15], which is a downloadable software pipeline capable of fully annotating a draft genome in about 10 minutes on a typical desktop computer. The output of the annotation process is a GenBank file (GBK), which can be visually explored for genomic features of interest, compared to other genomes and further annotated with freely available software such as the “Artemis” genome browser [16].

Sequence-based microbial typing information can also be extracted from the draft genome. Multilocus sequence typing

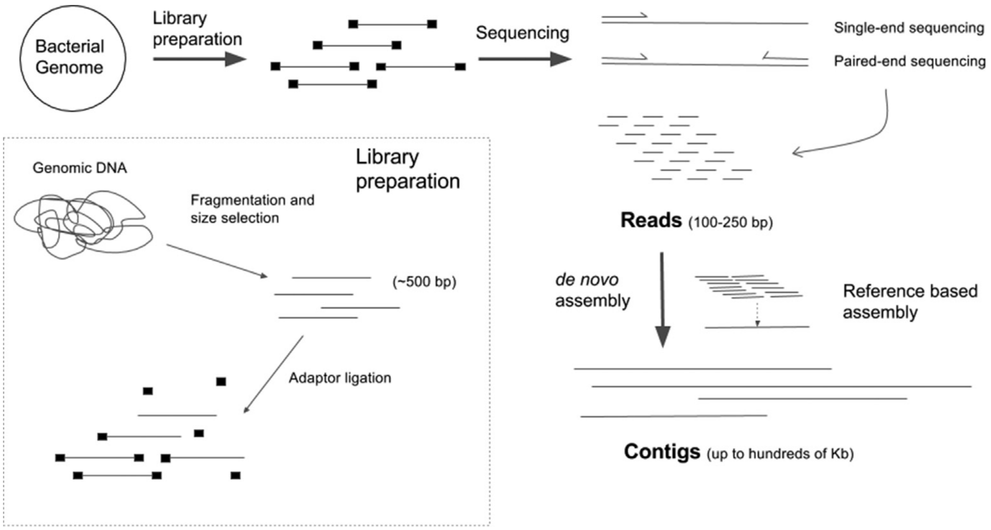

Fig. 2. Sequencing concepts from bacterial genomes to contigs.

(MLST) information can be easily and quickly assessed on a local computer or online web services such as "MLST1.8" [17]. This information can also be used to predict important phenotypic traits (*in silico* phenotypic analyses) such as predicting serotype and assessing the presence of antimicrobial resistance or virulence genes [18]. Despite its advantages, draft genome analysis does present challenges, including difficulty in distinguishing plasmid and chromosomal sequences, and misinterpretation arising from missing sequence data or misassembled contigs. The latter was recently reported for the MLST determination from *Legionella pneumophila* assemblies, and a computational solution was provided to ensure backwards compatibility [19].

An important application of draft genomes is the use of gene-by-gene approaches [20], which extends the concept of classical MLST from incorporating a few discriminatory genes to a much larger number of targets comprising the core genome harboured by a given species (core genome (cg) MLST), or alternatively the whole (core and accessory) genome for a given species (whole genome (wg) MLST). These approaches greatly increase the discriminatory power over traditional MLST and are being adopted as one of the main methodologies for food-borne bacterial typing and molecular surveillance by PulseNet International [21]. There are three publicly available online databases that facilitate gene-by-gene analysis for an increasing number of bacterial species: PubMLST (https://pubmlst.org/), which hosts schemas (i.e. defined set of loci to be used in MLST or wg/cgMLST), among others, for *Campylobacter* spp [20,22]. and *Neisseria meningitidis* [23]; a *Listeria monocytogenes* schema [24] hosted at the Pasteur Institute (http://bigsdb.pasteur.fr/listeria/listeria.html); and Enterobase (https://enterobase.warwick.ac.uk/) which hosts wg/cgMLST schemas for *Salmonella, Escherichia/Shigella* and *Yersinia* species. Several other schemas are being developed by commercial vendors. This kind of analysis can also be performed with downloadable pipelines that can be run locally, such as "Genomic Profiler" [25] or "chewBBACA" [26]. Further information on gene-by-gene methods and SNP/SNV analysis can be found in other articles in this thematic issue.

It is often desirable to compare annotated draft genomes for gene presence or absence without requiring a predefined schema. The software "Roary" [27] offers a computationally efficient solution to compare the pan-genome of a large number of annotated draft genomes. "Roary" allows for the identification of specific genomic content in the draft genomes that can be further analysed for association with specific phenotypic traits using its companion software "Scoary" [28]. The "Roary" plus "Scoary" system provides an efficient solution for microbial genome-wide association analyses based on gene presence or absence. A related software is Neptune [29], which can identify differential abundance of genomic regions without requiring genome annotation information. Microbial genome-wide association methodologies can be used to help identify loci that are associated with strain virulence and host or niche adaptation.

Read mapping approaches

If multiple strains are mapped against a single reference genome, the common variants can be used to produce a phylogenetic tree, although not every detected variant may be suitable for this task. For example, SNVs in recombinant regions or in multiple-copy regions of the genome may mask the true phylogenetic signal. The "MUMmer" suite of sequence analysis tools [30] can be used to mask repeat regions, and software such as "Gubbins" [31] or "ClonalframeML" [32] identify regions with high SNV density, which correspond to regions with a high likelihood of having a recombinant origin. Several pipelines have been developed that automate the process of extracting high-quality, phylogenetically informative SNVs from read sequence data, such as "SNVPhyl" [33], "Lyve-SET" [34], "CFSAN SNP Pipeline" [35] and "Snippy". For the construction of phylogenetic trees from the final SNV table, several software options are available such as "MEGA" [36], "FastTree" [36,37] or "RaxML" [38].

Another useful read mapping approach relies on using a set of target genes or genomic regions instead of using a complete or draft genome as reference for mapping. This allows for a 'reads to type' approach, which allows for a faster assessment of gene presence or absence and allele identification compared to *de novo* assembly and annotation methods. Software such as "SRST2" [39] or "ReMatCh" [40] can be used to accurately determine MLST types or identify specific antimicrobial resistance genes of thousands of strains available in the SRA or the European Nucleotide Archive (ENA). Quick determination of the phylotype and/or inference of phenotypic traits (e.g. antimicrobial resistance) of a given pathogen might be advantageous in specific settings. However, the reliability of phenotype inference from genotypic data needs to be validated on a case-by-case basis. That said, decision making in real-life situations should preferably be based on more than one analytical approach, as time and resources allow.

One advantage of read mapping approaches over gene-by-gene approaches is the ability to analyse noncoding regions, which has the potential to reveal changes in gene regulatory regions that can be translated in phenotypic differences, although this is limited by the reference genome used [41].

Visualization software

The final output usually involves some annotated representation of the relationships between isolates, and to obtain actionable information, epidemiologic and genomic data need to be integrated in the visualization.

“Microreact” [42] offers a comprehensive web service where phylogenetic trees and associated geographic, genetic and epidemiologic data can be uploaded, visualized and dynamically explored, which promotes the sharing of large data sets in the platform. “GenGIS 2” [43] is another software providing geospatial data analysis that can use the trees and epidemiologic data provided by the user.

Focusing on the analysis of allelic profiles derived from gene-by-gene methods or SNV analysis, “PHYLOViZ 2.0” [44] offers multiple analysis methods for the analyses of wg/cgMLST, allowing the representation of minimum spanning trees or hierarchical clustering together with associated metadata. “PHYLOViZ Online” [45] extends PHYLOViZ capabilities in an online platform which allows data sharing and visualization of large trees and metadata directly in a web browser.

Other online tools and databases

There are web resources that freely offer comprehensive analyses pipelines, such as the Center for Genomic Epidemiology (http://www.genomicepidemiology.org/), the NCBI Pathogen Detection platform (https://www.ncbi.nlm.nih.gov/pathogens/) or MicroScope [46], provided that there are no privacy or ethical issues in submitting data to a third-party online repository. In the NCBI Pathogen Detection platform, there are already several species-specific databases against which the isolates of interest can be compared. These websites offer an excellent starting point for a neophyte in the field, allowing for a quick appraisal of what can be done with HTS data. Other online tools can provide more specific analysis, such as Island Viewer [47], that allows the visualization and multistrain comparison of genomic islands directly from assemblies.

Of special note are bioinformatics resources such as the Comprehensive Antimicrobial Resistance Database (CARD) [48], which offers a curated resource of resistance genes and associated phenotypes; the Virulence Factors Database (VFDB) [49], which keeps an updated repository of virulence factors of various bacterial pathogens for more than 10 years; and ICEberg [50], an online resource for integrative and conjugative elements in bacterial species. These resources provide important information that can be used by other tools for annotation and identification of specific target genes.

Conclusions

In this review, we surveyed and discussed a small part of currently available bioinformatics software for HTS data analysis, reflecting the authors' personal experience and practice. In Table 1 we provide an extensive listing of available software. It is important to highlight that most of the described software runs in Linux environments and requires high-performance computing and storage resources when the number of analysed strains reaches a few dozen or higher. A recent review provides a more technical and thorough analysis of the technical requirements and project management challenges of HTS analyses [12].

Specific technical knowledge on how to install and maintain the software is necessary, making bioinformaticians a vital addition to any research group or microbiology unit in the hospitals or other public sectors that wishes to perform HTS data analysis. Still, nonspecialists should acquire the necessary understanding of the methods and basic bioinformatic skills, allowing them to effectively lead genomic studies, interpret their results and benefit from them in their practice.

Although we focused on freely available open-source software, Windows-based commercial software and web platforms can provide nonbioinformaticians with the tools needed for performing out-of-the-box analyses using user-friendly interfaces, without the need for Unix command line tools. Nevertheless, users should be aware of the software's limitations, which can be hard to assess achieve given the black box nature of the closed-source software of commercial solutions. The fast-moving nature of this field also creates difficulty for timely updates on commercial solutions.

The integration of HTS in applied microbiologic research, and even more so in routine clinical and public health microbiology, depends on a variety of factors, as reviewed elsewhere in this issue. With rapid improvements and resulting declining costs in DNA extraction and sequencing, bioinformatics analysis has become the rate-limiting step in the widespread use of HTS. It is thus crucial that nonspecialists in bioinformatics master a basic understanding of the concepts of HTS analysis, as we have tried to summarize here, to facilitate communication within the multidisciplinary teams necessary for the implementation of such innovative and powerful technology in routine laboratory work.

Transparency Declaration

JAC reports funding from BacGenTrack (research grant TUBITAK/0004/2014; FCT/Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Arasrrma Kurumu, Tübitak)). MR reports personal fees from Pfizer, personal fees from GlaxoSmithKline and personal fees from Merck Sharp & Dohme outside the submitted work; and funding on the review subject from the following grants: ONEIDA project (LISBOA-01-0145-FEDER-016417), cofunded by Fundos Europeus Estruturais e de Investimento (FEEI) from Programa Operational Regional Lisboa 2020; and by national funds from Fundação para a Ciência e a Tecnologia (FCT) and BacGenTrack (TUBITAK/0004/2014; FCT/Scientific and Technological Research Council of Turkey (Türkiye Bilimsel ve Teknolojik Arasrrma Kurumu, Tübitak)). The other authors report no conflicts of interest relevant to this article.

References

[1] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114–20.
[2] Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;15. R46.
[3] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85:2444–8.
[4] Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2013;2:10.
[5] Koren S, Treangen TJ, Hill CM, Pop M, Phillippy AM. Automated ensemble assembly and validation of microbial genomes. BMC Bioinform 2014;15:126.
[6] Jünemann S, Prior K, Albersmeier A, Albaum S, Kalinowski J, Goesmann A, et al. GABenchToB: a genome assembly benchmark tuned on bacteria and benchtop sequencers. PLoS One 2014;9:e107014.
[7] Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 2013;29:1072–5.
[8] Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics 2012;28:3169–77.

[9] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357–9.

[10] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589–95.

[11] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25:2078–9.

[12] Lynch T, Petkau A, Knox N, Graham M, Van Domselaar G. A primer on infectious disease bacterial genomics. Clin Microbiol Rev 2016;29:881–913.

[13] Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 2016;44:6614–24.

[14] Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res 2014;42:D206–14.

[15] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30:2068–9.

[16] Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics 2000;16:944–5.

[17] Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, et al. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol 2012;50:1355–61.

[18] Friães A, Machado MP, Pato C, Carriço J, Melo-Cristino J, Ramirez M. Emergence of the same successful clade among distinct populations of emm89 *Streptococcus pyogenes* in multiple geographic regions. MBio 2015;6. e01780–15.

[19] Gordon M, Yakunin E, Valinsky L, Chalifa-Caspi V, Moran-Gilad J, ESCMID Study Group for Legionella Infections. A bioinformatics tool for ensuring the backwards compatibility of *Legionella pneumophila* typing in the genomic era. Clin Microbiol Infect 2017;23:306–10.

[20] Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol 2013;11:728–36.

[21] Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, Gilpin B, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill 2017;22:30544.

[22] Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. A core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. J Clin Microbiol 2017;55:2086–97.

[23] Bratcher HB, Corton C, Jolley KA, Parkhill J, Maiden MCJ. A gene-by-gene population genomics platform: *de novo* assembly, annotation and genealogical analysis of 108 representative *Neisseria meningitidis* genomes. BMC Genomics 2014;15:1138.

[24] Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome–based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2016;2:16185.

[25] Zhang J, Halkilahti J, Hänninen ML, Rossi M. Refinement of whole-genome multilocus sequence typing analysis by addressing gene paralogy. J Clin Microbiol 2015;53:1765–7.

[26] Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, et al. chewBBACA: a complete suite for gene-by-gene schema creation and strain identification. bioRxiv 2018. https://doi.org/10.1101/173146.

[27] Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31:3691–3.

[28] Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome–wide association studies with Scoary. Genome Biol 2016;17:238.

[29] Marinier E, Zaheer R, Berry C, Weedmark KA, Domaratzki M, Mabon P, et al. Neptune: a bioinformatics tool for rapid discovery of genomic variation in bacterial populations. Nucleic Acids Res 2017;45:e159.

[30] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5.

[31] Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 2015;43:e15.

[32] Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 2015;11:e1004041.

[33] Petkau A, Mabon P, Sieffert C, Knox NC, Cabral J, Iskander M, et al. SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology. Microb Genom 2017;3:e000116.

[34] Katz LS, Griswold T, Williams-Newkirk AJ, Wagner D, Petkau A, Sieffert C, et al. A comparative analysis of the Lyve-SET phylogenomics pipeline for genomic epidemiology of foodborne pathogens. Front Microbiol 2017;8:375.

[35] Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Comput Sci 2015;1:e20.

[36] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 2016;33:1870–4.

[37] Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.

[38] Stamatakis A. RAxML version 8:a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–3.

[39] Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 2014;6:90.

[40] Machado MP, Ribeiro-Gonçalves B, Silva M, Ramirez M, Carriço JA. Epidemiological surveillance and typing methods to track antibiotic resistant strains using high throughput sequencing. Methods Mol Biol 2017;1520:331–56.

[41] Thorpe HA, Bayliss SC, Hurst LD, Feil EJ. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. Genetics 2017;206:363–76.

[42] Argimón S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2016;2:e000093.

[43] Parks DH, Mankowski T, Zangooei S, Porter MS, Armanini DG, Baird DJ, et al. GenGIS 2: geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. PLoS One 2013;8:e69885.

[44] Nascimento M, Sousa A, Ramirez M, Francisco AP, Carriço JA, Vaz C, et al. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics 2017;33:128–9.

[45] Ribeiro-Gonçalves B, Francisco AP, Vaz C, Ramirez M, Carriço JA. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res 2016;44:W246–51.

[46] Vallenet D, Calteau A, Cruveiller S, Gachet M, Lajus A, Josso A, et al. MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res 2017;45:D517–28.

[47] Bertelli C, Laird MR, Williams KP, Simon Fraser University Research Computing Group, Lau BY, Hoad G, et al. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res 2018;45(Web Server issue):W30–5.

[48] Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 2017;45:D566–73.

[49] Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: Hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res 2016;44. D694–7.

[50] Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, et al. ICEberg: a web-based resource for integrative and conjugative elements found in bacteria. Nucleic Acids Res 2012;40:D621–6.