Pyrex Journals 2035-6545

Pyrex Journal of Microbiology and Biotechnology Research
Vol 1 (3) pp. 038-048 September, 2015
http://www.pyrexjournals.org/pjmbr

Original Research Article

Whole Genome Sequencing: Bacterial Typing Revolutionized

CA Nsofor[1,2*], TE Ogbulie[1] and OC. Ugbogu[3]

[1] Department of Biotechnology, Federal Univeristy of Technology Owerri, Nigeria
[2] Key Laboratory of Medical Molecular Virology of Ministries of Education and Health, Institutes of Biomedical Sciences and Institute of Medical Microbiology, Shanghai Medical College, Fudan University, Shanghai, China
[3] Department of Microbiology, Federal University, Wukari, Nigeria

Accepted 28th May, 2015

Bacterial genotyping, or identifying bacteria at the genetic level, is particularly important for diagnosis, treatment, and epidemiological surveillance of bacterial infections. This is especially the case for bacteria exhibiting high levels of antibiotic resistance or virulence, and those involved in nosocomial or pandemic infections. Genotyping also has applications in studying bacterial population dynamics. Over the last two decades, molecular methods have progressively replaced phenotypic assays to type bacterial strains. Whole-genome sequencing of bacteria has recently emerged as a cost-effective and convenient approach for addressing many microbiological questions. Here, we briefly review the state of the art within this field and provide a step-by-step introduction to the workflow involved in genome sequencing, assembly and annotation. We also discussed the application of this technique in a clinical microbiology laboratory, focusing on three essential tasks: (1) identifying the species of an isolate, (2) testing its properties, such as resistance to antibiotics and virulence, and (3) monitoring the emergence and spread of bacterial pathogens. We predict that the application of whole generations sequencing will soon be sufficiently fast, accurate and cheap to be used in routine clinical microbiology practice, where it could replace many complex current techniques with a single, more efficient workflow.

Key words: Bacteria, genomes,genetics.

INTRODUCTION

Genetic diversity can ultimately explain most phenotypic variability in bacteria, such as geographic distribution, host specificity, pathogenicity, antibiotic resistance, and virulence. As bacterial strains pose ever greater challenges to human health, including increased virulence and transmissibility, resistance to multiple antibiotics, expanding host spectra, and the possibility of genetic manipulation of bioterrorism, identifying bacteria at the genetic level is increasingly important in modern microbiology (Fournier et al., 2009). The intraspecies diversity of bacteria mainly results from three genetic events: horizontal gene transfer, gene loss or acquisition, and recombination. The frequency of these three events makes the investigation of intraspecies diversity quite complicated (Fraser-Liggett, 2005). Bacterial genotyping, characterizing a number of strains in detail and ascertaining whether they are derived from a single parental organism, is a way to identify bacteria at the genetic level and to uncover the genetic diversity underlying important phenotypic characteristics.

Rapid advances in sequencing technology and Bioinformatics tools during the last decade have initiated a transition from classical molecular epidemiology to genomic epidemiology. This development has two major implications. First, by significantly scaling-up the numbers of genetic markers, genome wide approaches enhances the power and resolution for the above-mentioned applications and improves the reliability of conclusions (Steiner et al. 2013, Ekblom and Wolf 2014). Second, the application of genomic technologies opens novel axes of investigation (Allendorf et al. 2010, Ouborg et al. 2010). Genome-scale data provide information beyond neutral genetic variation or candidate gene approaches and thus enable screening for selectively important variation and assessing the adaptive potential of populations (Primmer, 2009). For example, approaches such as genome wide scans for selection, association mapping or quantitative trait loci (QTL) mapping can pinpoint loci of relevance for local adaptation of the target population (Steiner et al. 2013), with the potential to conserve evolutionary processes – a long sought after goal in conservation biology (Fraser and Bernatchez, 2001). Genomic approaches can also be applied to highly fragmented DNA from ancient material (e.g. from museum specimens; Pääboet al. 2004; Bi et al., 2013), to characterize environmental samples (Shokralla et al. 2012) and to understand how environmental perturbations affect

Corresponding Author: nsoforac@gmail.com

microbial communities (Mardis, 2008), representing largely unexplored terrain in molecular microbiology.

The above-mentioned applications do not necessarily require a reference genome sequence. Many analyses, including taxonomic delineation, characterization of demographic events, estimates of relatedness, can be successfully conducted in the absence of a genome reference. Instead, large-scale marker data, such as genotyping-by-sequencing (Elshire et al. 2011), RAD-Seq (Narum et al. 2013), reduced representation sequencing (Van Tassell et al. 2008), amplicon sequencing (Zavodna et al. 2013) or transcriptome sequencing (Ekblom and Galindo, 2011) can be effectively utilized without relying on a genome backbone. A complete and well-annotated genome sequence, however, provides the ultimate resource for genomic approaches. Whole-genome sequencing data with positional information along a genome sequence constitute the most complete account of bacterial genomic variation (e.g. structural rearrangements, copy number variation, insertion–deletion, single nucleotide polymorphisms (SNPs), sequence repeats). It also provides the basis for haplotype information and genomewide estimates of linkage disequilibrium which have great power to reveal recent population histories (Li and Durbin, 2011), timing of admixture events (Hellenthal et al. 2014) and to screen for signatures of selection (Hohenlohe et al. 2010). The study of selectively important variation strongly relies on annotated genome data to identify the functional genomic regions of interest. Reference sequences are further indispensable as a template for RNA-seq in detailed studies of (isoform-specific, allele-specific) gene expression (Vijay et al. 2013), epigenetic modifications (such as methylation; Herrera and Bazaga, 2011) and DNA–protein interactions (Auerbach et al. 2013). These approaches are only accessible to genome-enabled taxa (Kohn et al. 2006) that enjoy the added benefit of using the latest Bioinformatics tools developed in the biomedical sciences.

Here, we introduce the workflow of a typical whole-genome sequencing project conducted by different research groups. This review aims at introducing principles and concepts to beginners in the field and offers practical guidance for the many steps involved. We discussed sequencing, assembly and annotation, highlighting typical routines and analytical procedures and also discussed the application of whole genome sequencing in clinical microbiology.

1.1. Definition of Terms

Below is the definition of terms commonly used in whole genome sequencing practice. We have included this aspect for easy understanding by the reader.

Alignment: Similarity-based arrangement of DNA, RNA or protein sequences. In this context, the subject and query sequence should be orthologous and reflect evolutionary, not functional or structural relationships.
Annotation: A computational process of attaching biologically relevant information to genome sequence data.
Assembly: Computational reconstruction of a long sequence of smaller sequence reads.
Barcode: Short-sequence identifier for individual labelling (Barcoding) of sequencing libraries.
BAC: (Bacterial artificial chromosome) DNA constructs of various lengths (150–350 kb).
cDNA: Complementary DNA synthesized from an mRNA template.
Contig: A contiguous linear stretch of DNA or RNA consensus sequence. Constructed from a number of smaller, partially overlapping, sequence fragments (reads).
Coverage: Also known as ‘sequencing depth’. Sequence coverage refers to the average number of reads per locus and differs from physical coverage, a term often used in genome assembly, referring to the cumulative length of reads or read pairs expressed as a multiple of genome size.
De novo assembly: Refers to the reconstruction of contiguous sequences without making use of any reference sequence.
EST library: Expressed sequence tag library. A short subsequence of cDNA transcript sequence.
Fosmid: A vector for bacterial cloning of genomic DNA fragments that usually holds inserts of around 40 kb.
GC content: The proportion of guanine and cytosine bases in a DNA/RNA sequence.
Gene ontology (GO): Structured, controlled vocabularies and classifications of gene function across species and research areas.
InDel: Insertion/deletion polymorphism.
Insert size: Length of random sheared fragments (from the genome or transcriptome) sequenced from both ends.
K-mer: Short, unique element of the DNA sequence of length k, used by many assembly algorithms.
Library: The collection of DNA (or RNA) fragments modified in a way that is appropriate for downstream analyses, such as high-throughput sequencing in this case
Mapping: A term routinely used to describe alignment of short sequence reads to a longer reference sequence
Masking: Converting a DNA sequence [A,C,G,T] (usually repetitive or of low quality) to the uninformative character state N or to lower case characters [a,c,g,t] (soft masking)
Massively parallel (or next generation) sequencing: High-throughput sequencing nanotechnology used to determine the base-pair sequence of DNA/RNA molecules at much larger quantities than previous end-termination (e.g. Sanger sequencing) based sequencing techniques
Mate-pair: Sequence information from two ends of a DNA fragment, usually several thousand base-pairs long
N50: A statistic of a set of contigs (or scaffolds). It is defined as the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs
N90: Equivalent to the N50 statistic describing the length for which the collection of all contigs of that length or longer contains at least 90% of the total of the lengths of the contigs
Optical map: Genomewide, ordered, high-resolution restriction map derived from single, stained DNA molecules. It can be used to improve a genome assembly by matching it to the genomewide pattern of expected restriction sites, as inferred from the genome sequence
Paired-end sequencing: Sequence information from two ends of a short DNA fragment, usually a few hundred base pairs long
Read: Short base-pair sequence inferred from the DNA/RNA template by sequencing
RNA-Seq: High-throughput shotgun transcriptome (cDNA) sequencing. Usually not used synonymously to RNA-sequencing, which implies direct sequencing of RNA molecules skipping the CDNA generation step
Scaffold Two or more contigs joined together using read-pair information
Transcriptome: Set of all RNA molecules transcribed from a DNA template (Ekblom and Wolf 2014)

1.2. What does it mean to ‘sequence a genome’?

Ideally, a genome draft would represent the complete nucleotide base sequence for all chromosomes in the species of interest, a ‘physical map’ of its genetic content (as opposed to the ‘genetic or the linkage map’ which establishes the order and recombination distances among genetic markers), (Ekblom and Wolf 2014). However, in reality, there are a number of complications with the concept of a ‘genome sequence’. First, there is not one true sequence of a species because of individual genomic variation. In a single bacterium, such variation will manifest itself in the form of heterozygous positions, insertion/deletion (InDel) polymorphism, copy number variation or small-scale rearrangements. Even strains from the same bacteria can differ in genomic content due to mutations. The assembled genome sequence of the bacteria will also be only one representation of the total variation present in a species (paralleling the use of ‘type specimens’ for taxonomic classification). Generally, only a single individual is sequenced (Wheeler et al. 2008), but sometimes (like in the HUGO project) the genome represents a ‘consensus’ of a number of pooled samples (International Human Genome Sequencing Consortium, 2004). Note, however, that in diploid and polyploid organisms, the genome assembly already reflects a consensus sequence of several chromosome sets and fails to capture haplotypic variation (for most current short-read based methods). Second, it is essentially impossible to sequence and assemble all nucleotides in the genome (Ellegren 2014). Large parts of DNA sequence, especially the heterochromatic regions around centromeres and telomeres and other highly repetitive regions, are not well characterized even in mature genome assemblies like human or mice. Third, there will always be some degree of error in the characterized genome sequence, both on the level of individual nucleotides (stemming from sequencing errors) and in the ordering of sequence blocks (stemming from assembly errors). Forth, every genome assembly is the result of a series of assembly heuristics and should accordingly be treated as a working hypothesis.

1.3. The principle of genome sequencing and assembly

Currently, most genome projects use a shotgun sequencing strategy for genome sequencing. In a first step, genomic DNA is sheared into small random fragments. Depending on the technology, these are sequenced independently to a given length. Powerful computer algorithms are then utilized to piece the resulting sequence reads back together into long continuous stretches of sequence (contigs), a process known as de novo assembly. For correct assembly, it is important that there is sufficient overlap between the sequence reads at each position in the genome, which requires high sequencing coverage (or read depth). Naturally, for longer sequence reads, the more overlap can be expected, reducing the required raw read depth. Usually, longer fragments (several hundred base pairs) are sequenced from both ends (paired-end sequencing) to provide additional information on correct read placement in the assembly (Ekblom and Wolf 2014).

After the initial assembly, contigs are typically joined to form longer stretches of sequence (known as scaffolds). To achieve this, libraries from long DNA fragments spanning several kilobases (kb) of sequence in the genome are prepared and their endpoints sequenced. Depending on the technology and the specifics of the library preparation, these libraries are (somewhat confusingly) called, for example, paired-end, mate-pair or jump libraries. If the endpoint sequences of several independent fragments come to lie on two different contigs, they are joined into a scaffold. The expected fragment length of the library provides information on the physical distance between the two contigs, and the created gap is filled with the uninformative base-pair character ‘N’. Subsequent gap closing methods, ideally using long reads that read across repetitive sequences, help to fill in the missing base-pair information.

In a last step, the resulting scaffolds are often joined into linkage groups or placed on chromosomes (Ellegren 2014). Genetic maps constructed from pedigree data or crosses are arguably the best way for ordering and orienting scaffolds into longer sequence blocks (Ellegren et al. 2012). However, detailed genetic maps of species with conservation concern (usually not amendable to artificial crosses or half-sib breeding designs) require substantial genotyping effort, and deep pedigrees with a sufficient number of meioses are difficult to come by in most systems (Romanov et al. 2009). Given these difficulties, it is often not realistic to aim for a chromosome-level assembly, and this will also often not be necessary for most conservation biology applications. Most applications, including haplotype-based approaches that are powerful in revealing signatures of selection or depict recent demographic histories, generally work with high-quality contigs. As an alternative for placing and orienting the scaffolds onto putative chromosomes, synteny and gene order information from related species can be used. Note, however, that such information should be used with due caution as chromosomal rearrangements may have occurred even between very closely related species. There is also a risk that errors in the reference species assembly are transferred to the focal genome.

2.0. Genome sequencing

2.1. Sequencing technology and coverage considerations

Among the first decisions when starting, a genome sequencing project is the choice of sequencing platform, the type and amount of sequence data to generate. The latter is often limited by project funding, and the former may depend on which sequencing technology is promptly available. Judging from recently completed whole-genome sequencing projects, there is a clear trend moving away from traditional Sanger sequencing (~1 kb sequence reads) and Roche 454 sequencing (up to 800 bp) towards short read technologies such as IlluminaHiSeq (at present typically 150 bp) and SOLiD (typically 50 bp). Lately, there has been progress in producing longer reads at high throughput; several technologies offering this, such as Pacific Biosciences (up to 5 kb), IonTorrent (~500 bp) and IlluminaMoleculo (up to 10 kb), are entering the market, and we expect to see a broader spectrum of read lengths. While this development blurs the initial dichotomy of short reads (e.g. 35 bpIllumina reads) versus long reads (~1 kb Sanger reads), read length still has important Bioinformatics implications, as assembly algorithms optimized for long reads are fundamentally different from approaches targeting short reads. Recent studies have begun to combine data of different read length and from several different sequencing platforms (Korenet al. 2012). This strategy makes intuitive sense as the drawbacks of each method can be counterbalanced, although the jury is still out whether such hybrid assemblies always outperform single data type approaches (Bradnamet al. 2013). Here, we follow the principle of current common practice and base our considerations largely on sequencing of Illumina libraries of different lengths (we loosely refer to short reads at sequence lengths below 500 bp and long reads above this length). Many of the following reflections, however, more generally relate to the assembly problem and do not depend on the specific choice of sequencing library.

For most downstream applications, obtaining long contigs is essential. With long-read data, from traditional Sanger sequencing of individual BAC clones, this is feasible even with a rather limited sequencing depth. However, when using only short-read technologies, high total read coverage (>100×) is needed. Too little data will result in a highly fragmented assembly and severe problems with downstream applications such as annotation and variant calling. For initial contig assembly, one usually starts out with a high amount of paired-end short-read data. To subsequently merge contigs into scaffolds, it is necessary to generate additional libraries with long-insert sizes in the range of 3–40 KB. How much sequencing data are needed of each library type and insert size depends critically on a number of factors, including the size and repeat content of the genome, the degree of heterozygosity and the target quality of the assembly (Sims et al. 2014). As these parameters will differ between sequencing projects and organisms of interest, the optimal resource allocation will be unique to every project. It should be noted that coverage could sometimes also be too high, as the absolute number of sequencing errors increases as a function of reading number. According to past experience, down sampling from 100× to 50× coverage of a short-insert size library can significantly improve some steps in the assembly process.

To translate these recommendations into amount and type of sequencing needed for a specific project, basic knowledge on genome size, sequencing error rates, repeat content and the degree of genome duplications are needed. If no such information is available for the target species of interest at the start of the project, it is advisable to first perform a small pilot study using single-end or short-insert sequencing. The above-mentioned parameters can then be approximated using a k-mer counting approach (Marçais and Kingsford, 2011) http://josephryan.github.com/estimate_genome_size.pl.

Information on how to perform and interpret such k-mer counts can be found on web forums such as seqanswers. Generally, a larger amount of long-insert data are needed for correct assembly if the genome has a high repeat content or a high degree of duplications. Genome size estimates for a large number of species are also available in online databases.

2.2. Wet-lab procedures

The wet-lab part of the genome sequencing is often outsourced to sequencing centre's, and we will only very briefly touch upon the basic steps of library preparation that are important to consider at the planning stage of the project and that affect downstream analytical procedures.

2.2.1. DNA quality

Whole-genome sequencing, particularly of long-insert size libraries, requires high-quality, intact, nondegraded DNA at a sufficient amount (Wong et al. 2012). For sequencing, a full genome using a set of different libraries requires ~1 mg of DNA as starting material (~6 µg for short-insert libraries, ~40 µg for 2–10 kb libraries, ~60 µg for >20 kb libraries). Before engaging in genome sequencing, it is thus essential to obtain a large amount of high-quality DNA of the target species. This can be a major obstacle for many species with conservation concern. Prior to submitting a DNA sample, its integrity should be checked on a high-resolution gel (e.g. pulse-field electrophoresis; a sample should typically show fragments of >100 kb).

2.2.3. Library preparation

When choosing the necessary raw read depth, one should be aware that currently, most technologies include several PCR steps which can lead to a non negligible number of duplicated reads. While single read can occur in duplicate by chance if coverage is high enough, duplication is bound to be an artifact for identical read pairs which are very unlikely to occur by chance (as they follow a length distribution). As duplicated reads are of no added value and duplication artifacts can impair coverage-based quality validation, they should be removed prior to the assembly. Duplicates generally constitute a few percentages of short-insert size libraries (<500 bp), but can reach over 95% for long-insert libraries (>10 kb).

Another central question refers to what insert sizes to use. Generally, it is advisable to have a good mix of sizes in the range of 0.2–40 KB with the shorter libraries being sequenced to significantly higher depth (Gnerreet al. 2011). Insert sizes of >20 kb make a large difference to the final contiguity and the scaffold size of the assembly, but are not trivial to produce at high quality and currently constitute a limitation of many sequencing centre's. Library preparations differ in quality and in how well they represent different parts of the genome. Therefore, more than one library should ideally be generated per size class. Note that some assembly programs (such as ALLPATHS-LG) expect a predefined mix of sequencing libraries as input data. Another important issue for downstream analyses that comes with library preparation is read orientation. Depending on the technology used, reads can face inwards (; e.g. Illumina paired-end sequencing) or outward (; e.g. Illumina mate-pair sequencing) in relation to the original DNA fragment. Mis-oriented reads with unexpectedly short insert sizes can arise due to sequencing of pairs from within the original DNA fragment rather than at its ends. Also, mate-pairs with aberrant insert sizes and orientation often represent chimeric sequences from nonadjacent genomic regions. For most assembly methods, such artifacts need to be filtered out during the preassembly steps; often leaving only a small fraction of usable, unique read pairs for assembly after trimming. To correctly process the data, the bioinformatician handling the data always needs to be 'library aware'.

3.0 Genome assembly

3.1. Data management

The amount of data generated in a normal genome sequencing project is staggering. A genome with 100× coverage means data files in the order of several hundred gigabytes. During the assembly procedure, temporary files easily cross the terabyte boundary. An adequate data management and backup strategy is thus needed already at the start of the project. Many universities are connected to local or national computing grids, including data storage facilities, and it is highly recommended to utilize these whenever possible. Having Bioinformatics application experts working at the computing infrastructure provides a vital link between the biologist researchers and the computing grid system experts (Lampa et al. 2013). Such collaborations should already be established during the planning stage of the project. Sequencing centre's often also offer assistance with data analyses and assembly. It is thus vital to explicitly discuss what kind of support can be provided by the facility before the start of the project. More generally, it should be considered whether enough expertise exists in the core research group to perform the computational steps of an assembly. Most data processing and genomic analyses of large-scale sequencing data are conducted on high-performance computing clusters running a UNIX-based operating system. One does not need to be a Bioinformatics expert to handle whole-genome sequencing data, but is essential to have some familiarity with the UNIX environment and basic knowledge of command line software, writing shell scripts and applying scripts of commonly used languages for biological data analysis (such as Perl or Python).

3.2. Preassembly steps

Prior to the assembly, the quality of the sequencing data, overall GC content, repeat abundance or the proportion of duplicated reads should be assessed. Tools such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) providing summary statistics are a useful starting point. Trimming low-quality data and reads resulting from PCR duplications can be performed with a variety of different software and scripts (e.g. ConDeTri; Smeds and Künstner, 2011). Stand-alone error correcting, using a k-mer count approach (for example as applied in the SOAPdenovo pipeline), can also be a useful alternative for many data sets. Note, however, that the optimal stringency of quality filtering is specific to the individual project and the targeted assembly pipeline. Some assemblers, such as ALLPATHS-LG (Gnerreet al. 2011), where trimming and error correction are performed within the assembly pipeline, even require raw reads, without quality trimming as input.

Primer and vector sequences from the library preparation will most likely be present in the data (even if the sequencing facility claims to have removed them) and can be removed with simple scripts (like cutadapt; Martin 2011). Also, in Illumina sequencing, DNA from the PhiX phage is often added to the sequencing reaction, in order to calibrate sequence quality scores. Failure to remove such abundant contaminant sequences can disrupt the assembly process (due to the high read depth compared with the nuclear genome) and may result in the production of chimeric and contaminated contigs. The easiest way of removing known vector contamination from the raw data is to use a short read aligner (like BWA; Li and Durbin 2009) and delete all fragments mappings to the contamination sequence.

3.3. De novo assembly

Tools for genome assembly differ widely in their performance in terms of speed, scalability and the quality of the final genome sequence (Miller et al. 2010; Earl et al. 2011; Narzisi and Mishra 2011; Bradnam et al. 2013). While some assembly methods clearly outperform others, it is currently difficult to predict which of the tools might be most appropriate in a given situation. Every assembly project is unique in terms of generated data structure and the target genome differing, for instance, in size, base-composition, repeat content and polymorphism level. There is a number of software available for de novo assembly of shotgun whole-genome sequencing data, and new programs are constantly being added to the list. Some algorithms focus on minimizing mis-assemblies, while others mainly aim to improve contiguity (sometimes at the cost of accuracy). Most assembly algorithms perform optimally with a given distribution of library sizes, so it is important to consider the choice of assembly strategy already during the project planning and sequencing steps. Besides information from the primary literature and websites of assembly software, various web forums provide good entry points for up-to-date discussions and sharing the experiences of other researchers.

Most software implementations designed for long-read technologies such as traditional Sanger sequencing (for example the Celera assembler, Arachne and PCAP; Batzoglou et al. 2002; Huang et al. 2003; Denisov et al. 2008) or Roche 454 sequencing (for example Newbler) use an assembly approach known as overlap-layout-consensus (OLC). These algorithms are generally considered too computationally intensive (mainly in terms of runtime) for Illumina or SOLiD data. Still, a few assemblers such as Edena (Hernandez et al. 2008), SGA (Simpson and Durbin 2012) and FERMI (Li 2012) pursue the OLC strategy for such short-read data (Miller et al. 2010). Most other strategies for de novo assembly of short sequence reads can be broadly divided into two classes: extension-based methods and De Bruijn (or Eulerian) graph algorithms (Nagarajan and Pop, 2013). Extension-based assemblers, such as SSAKE (Warren et al. 2007) and JR-Assembler (Chu et al. 2013) are usually computationally very efficient (in terms of both memory requirements and computational time), but are highly sensitive to sequencing errors, repeat regions and high levels of nucleotide polymorphism (Chu et al. 2013). The most commonly used approach for assembly of short-read data is therefore currently based on De Bruijn graphs, where reads are partitioned into k-mers (substrings of the read sequence of length k) that then form the nodes of the graph (network) and are linked when sharing a k-1 mer. Highly used assembly software, such as SOAPdenovo (Luoet al. 2012), ALLPATHS-LG (Gnerre et al. 2011), ABySS (Simpson et al. 2009) and Velvet (Zerbino and Birney 2008), all rely on De Bruijn graph algorithms. There are also ‘hybrid’ assembly approaches, for example Atlas (Havlak et al. 2004), Ray (Boisvert et al. 2010) and MaSuRCA (Zimin et al. 2013), combining features of different algorithms and utilizing data from multiple sequencing technologies. In general, it is advisable to test several different assembly methods and evaluate which is most appropriate for the specific data at hand. Draft genome building should be treated as an iterative process with several rounds of assembly, evaluation and parameter tweaking. For a more comprehensive review of different assembly algorithms and software; see for example (Miller et al. 2010; Nagarajan and Pop 2013).

After the initial contig building, it is common to use read-pair information from long-insert (mate-pair, fosmid-end or jump) libraries (Zhang et al. 2012) to combine contigs into scaffolds. Additional short-insert paired-end libraries are also often useful, for example to bridge, short low-complexity regions. The lengths of sequence gaps between contigs are estimated from the expected insert sizes and are usually filled with a stretch of Ns. The scaffolding step is already included in many commonly used assembly programs, but there are also some stand-alone applications, for example SSPACE (Boetzer et al. 2011) and BESST (Nystedt et al. 2013), to perform this step independently. Some of the gaps (N's) emerging from this process can be removed a posteriori using the original read-pair information with software such as GapCloser (Li et al. 2010), GapFiller (Boetzer and Pirovano 2012) and iMAGE (Tsai et al. 2010). Long-read data (for example from PacBio) has also recently emerged as a way of filling N regions in scaffolds (English et al. 2012).

When choosing assembly software, it is important to consider both the amount of sequencing data and which computational resources are available (Schatz et al. 2010a). De Bruijn graph methods, such as SOAPdenovo and ALLPATHS-LG, generally require large amounts of computing memory (RAM). If large computer clusters are not available locally, it will be necessary to consider collaborative equipment purchases, joint projects with Bioinformatics groups or utilization of commercially available computing clouds (Schatz et al. 2010b).

Another consideration to make is whether to use freely available programs (most programs mentioned above fall in this category) or to invest in commercial software (such as CLC workbench or Lasergene from DNASTAR). Commercial software is usually more user-friendly than freely available programs and thus readily used by researchers with limited bioinformatics skills. The downside of commercial software, apart from the (often substantial) cost involved in purchase and licensing, is that these act even more like 'black box' solutions, where it is often near impossible to inspect or alter details about the algorithms. Some commonly used software applications are also distributed together with the sequencing instruments and may be available through the sequencing facilities.

3.4. Quality assessment and validation

Once an assembly has been successfully performed, users will want to assess its quality or compare several assemblies using different methods. Yet, as discussed above, every draft genome assembly constitutes merely a hypothesis of the true underlying genome sequence, and in the absence of knowing the truth, assessing its quality remains a challenge. A variety of metrics reflecting different aspects of the assembly are available (Bradnam et al. 2013). They can be broadly divided into approaches that require additional information from external data and those solely based on information derived from the assembly itself. As external information is often not available in conservation genomics projects, intrinsic quality assessment of the assembly is a natural starting point. One basic metric is the proportion of the genome contained within the assembly. The expected genome size can either be inferred from C-value data or, alternatively, from k-mer frequency-based approaches. Another standard metric to evaluate assembly contiguity is the N50 statistic: by definition, 50% of the assembled nucleotides are found in contigs (contig N50) or scaffolds (scaffold N50) of at least this length. The N50 statistic thus describes a kind of median of assembled sequence lengths, giving greater weight to long sequences. Recently, variations of this metric, for example the NG50 and 'NG Graph' (Bradnam et al. 2013), incorporating the expected genome size was introduced and provides effective means of visualizing and evaluating differences in contiguity between assemblies.

However, the N50 statistic and variations thereof need to be interpreted with caution. They merely indicate contiguity and contain no information on assembly accuracy. To detect errors in the assembly, information from remapped paired-end or mate-pair data can be used (as, e.g., implemented in the software REAPR; Hunt et al. 2013). Low-coverage regions or mis-orientation of read pairs suggests mis-assemblies, while aberrant insert sizes indicate small insertions or deletions. Exceedingly high sequence coverage, high local SNP densities or correlated SNPs, where most of the reads carry one character state (but multiple others show another character state), can indicate the presence of collapsed, near-identical repeats. Software applications performing these steps are numerous, and examples can be found in the current literature (Earl et al. 2011; Bradnam et al. 2013). The amos validate pipeline (Phillippyet al. 2008; Schatz et al. 2013) encompasses several genome assembly diagnostics in one pipeline, but works best for small- or medium-sized genomes.

Independent experimental data sets from the target species arguably provide the best source of external information. Data from optical maps, for instance, allow validating short- and large- scale accuracy of scaffolds and expanding them further to approach chromosome level. Similarly, separately assembled sequences from BAC or fosmid libraries can help to assess sequence accuracy and repeat content. Both approaches, however, rely on correct assembly themselves and are not readily available for smaller laboratories at present. Independent de novo assemblies from shotgun transcriptome sequencing data (RNA-seq) are more easily generated, and expressed sequence tag (EST) libraries might already exist for species of conservation concern (although getting access to fresh tissues for RNA extraction may be a serious limitation if captive populations are not available). Sequence content and exon structure of transcriptome data thus constitute an important additional resource for validating sequence accuracy and for correcting scaffolding in cases where genes span across contigs.

Comparative genomic approaches provide another avenue, which does not require the generation of additional data. For example, quantifying the presence and completeness of orthologous core eukaryotic protein sequences (Parra et al. 2007) provides first intuition on the comprehensiveness of the assembly. In cases where high-quality reference genomes of sister taxa exist, genome comparisons might be of guidance in detecting mis-assemblies and chimeric contigs under the assumption of broad-scale synteny and gene order conservation. Small-scale rearrangements, however, might be real and require in depth investigation. DNA from other organisms are likely to have contaminated the genomic samples at various stages (during both sampling and laboratory procedures) and will be present in the sequencing data. Although mainly being a nuisance, contaminations at the sampling stage may actually be interesting from a conservation point of view, as they can carry information about parasites or other microorganisms related to the study species. External genomic resources aid in finding such contaminations that might have been assembled as separate contigs or are interspersed with target sequence in the same contig. Positive hits from a BLAST search or similar local alignment routines are often employed to find such traces of contamination, but results need to be interpreted with caution. Even correctly assembled sequences can lead to best hits from distantly related species with well-annotated genomes, particularly if taxon sampling within the target group of organisms is scarce. Likewise, small stretches of contamination in a large contig or scaffold may be missed entirely if other parts of the sequence yield significant hits on the target clade.

3.5. Genome annotation

To harness the full potential of a genome sequence, it needs to be annotated with biologically relevant information that can range from gene models and functional information, such as gene ontology (GO) terms (Gene Ontology Consortium 2004; Primmer et al. 2013) or 'Kyoto encyclopedia of genes and genomes' (KEGG) pathways (Kanehisa and Goto 2000), to micro RNA and epigenetic modifications (The ENCODE Project Consortium 2012). In the context of genetic nonmodel organisms, annotation is often confined to protein-coding sequence (CDS) or transcripts more generally. Despite the considerable challenge to annotate genes in newly sequenced species where preexisting gene models are mostly lacking, automated gene annotation has in principle become possible for individual research groups (Yandell and Ence 2012). Still, a complete genome annotation constitutes a considerable effort and requires bioinformatic proficiency. Before starting, it should be noted that successful annotation strongly depends on the quality of the genome assembly. Only contiguous near-complete (~90%) genomes interrupted only by small gaps will yield satisfying results. As a rule of thumb, large genomes have longer genes and thus need more contiguous assemblies for successful annotation. The annotation process can be conceptually divided into two phases: a ‘computational phase’ where several lines of evidence from other genomes or from species-specific transcriptome data are used in parallel to create initial gene and transcript predictions. In a second ‘annotation phase’, all (sometimes contradicting) information is then synthesized into a gene annotation, following a set of rules determined by the annotation pipeline.

Prior to gene prediction, it is of vital importance to mask repetitive sequences including low-complexity regions and transposable elements. As repeats are often poorly conserved across species, it is advisable to create a species-specific repeat library using tools like Repeat Modeler or Repeat Explorer (Novák et al. 2013). Once repeats are masked (e.g. with Repeat Masker; http://www.repeatmasker.org), ab initio algorithms trained on gene models from related species can be used for baseline prediction of coding sequence (CDS) (e.g. AUGUSTUS; Stanke et al.2006). Protein alignments (using e.g. tblastx) and syntenic protein lift-overs from a variety of other species provide a valuable resource to complement the predicted gene models. Arguably, the best evidence comes from detailed EST or RNA-seq data, which in addition to CDS, provides gene models with information on splice sites, transcription start sites and untranslated regions (UTRs). If possible, mRNA should be sequenced strand-specifically, as this helps resolve gene models, facilitates transcriptome assembly and eventually aids in the evaluation of the genome assembly.

In a next step, all the evidence from ab initio prediction and protein-, EST- or RNA-alignments need to be synthesized into a final set of gene annotations. As the evidence is mostly incomplete and sometimes contradicting, this is a difficult task that often benefits from manual curation. Still, several automated annotation tools like MAKER (Cantarelet al. 2008) or PASA (Haas et al. 2003) exist that incorporate, and weigh the evidence from, several sources. Although these tools generally provide good results, qualitative validation is important (e.g. by assessing the length of open-reading frames). Visual inspection of the annotation is another vital component to detect systematic issues such as intron leakage (introns being annotated as exons due to the presence of pre-mRNA) or gene fusion. Tools like WebApollo (Lee et al. 2013) from the GMOD project are particularly useful, as they allow the user to edit the annotation directly through the visual interface.

3.6. Publishing the genome

Draft genome sequences are now being produced at an ever-increasing rate. Traditional databases such as ENSEMBL from the European Molecular Biology Labs (EMBL) and the Wellcome Trust Sanger Institute, or genomic databases from the National Center for Biotechnology Information (NCBI) providing access to genomes and meta-information can no longer annotate and curate all incoming genomes. NCBI therefore already provides the possibility to upload draft genome sequences and user-generated annotation. To allow other users to improve the assembly and its annotation, all available raw data should be uploaded, together with the assembled genome and all relevant meta-data, for example as a BioProject on NCBI.

4.0. Applications of Whole Genome Sequencing in Microbiology

The major advantage of whole-genome sequencing is to yield all of the available DNA information content on isolates in a single rapid step following culture (sequencing without culture is discussed in the 'Future directions' section). In principle, the result contains all of the data that are currently used for diagnostic and typing needs, even though it is not always yet known how to interpret these data. However, the genome also includes vast amounts of additional data that are currently unavailable for routine processing, thus opening the prospect for large-scale research into pathogen genotype–phenotype associations of routinely collected data. The hurdles to implementing whole-genome sequencing in clinical and public health laboratories are substantial, as widespread adoption would require incorporating the knowledge from more than a century of characterizing pathogens — currently delivered by a skilled workforce — into an entirely new framework of mainly computer-driven genome processing.

This would require a radical shift towards a new operational paradigm for routine laboratories. In addition, a new understanding of genotype-to-phenotype relationships needs to be established, evaluated and deployed in parallel with current routine methods, which will require a major effort leading to gradual replacement of present-day methodologies over many years. Crucially, the translation of sequence technology into new practices in clinical microbiology is facilitated by genetic features of bacteria. Compared with eukaryotic genomes, bacterial genomes are much smaller (2–6 Mb), and bacteria usually possess a single haploid chromosome (although a few possess two haploid chromosomes). However, they are much more diverse than eukaryotic species, partly because ~10–40% of the genome may consist of dispensable sequences that are not shared with all members of the same species (Didelotet al., 2012, Medinietal 2005). Many of these dispensable elements are also mobile: for example, episomal structures such as plasmids. The plasmids and other mobile elements often encode antibiotic resistance and even virulence determinants, and as such they are highly relevant to clinical microbiology.

4.1. Species identification

As highlighted above, the identification of species is a crucial initial step in managing infectious diseases and tracking pathogens. Currently, taxonomic approaches are based on keeping a type strain collection as a gold standard (with the exception of MALDI–TOF, which can use a set of references for each species). Using whole-genome sequencing, this could be replaced by a 'type sequence'. That is, species would be taxonomically defined by their sequence, and the 'type sequence' would constitute a reference point against which to compare sequence data from other isolates. The relationship of the species to all previously sequenced organisms can be determined using phylogenetic analysis. A ribosomal MLST (rMLST) scheme has recently been proposed Jolley, (2012)that rely on the sequences of 53 genes encoding ribosomal proteins, which are present in all bacteria. Acquiring the sequences of such a large number of genes is best done by first sequencing the whole genome and then extracting individual genes using, for example, BLAST (Altschul, 1997). The BIGSdb database system is an integrated platform that enables users to find many genes in many genomes using BLAST and to record the results for future use (Jolley et al 2010). More than 1,900 bacterial genomes from 452 bacterial genera have been analyzed using the rMLST scheme. Any newly sequenced genome can easily be added to the database, can have its ribosomal genes extracted and can have its phylogenetic relationships with other genomes assessed. In a separate effort, a new method has recently been developed that allows the automatic in silico application to any genome sequence of the MLST schemes of 66 distinct species based on hundreds of genes, thus potentially revealing both the species to which the genome belongs and its sequence type within the relevant MLST scheme (Larsen, 2012).

With further development, these comparative approaches could reach the level required to replicate current species identification procedures with high precision. As this is progressively being achieved, our definitions of bacterial species will probably need refining to reflect new accumulated knowledge based on sequence comparison. Indeed, it has already been shown that sequence data, even at the level of fractional sequencing (for example, MLST), are robust at differentiating Streptococcus pneumoniae or Campylobacter jejuni from closely related species. However, it has also revealed that some named species do not represent monophyletic units of diversity: for example, in the case of Bacillus cereus and Bacillus thuringiensis. Although the increased statistical power of having the whole genomic sequence data considerably improves the precision of such analysis for differentiating all species, it will probably also reveal more ambiguity at the boundaries of currently defined species than have already been recognized from fractional sequencing. Such findings are likely to give impetus to a reconsideration of the notion of bacterial species, eventually leading to great simplification and clarity to the early steps in diagnostic clinical microbiology. For example, a genomic criterion for species definition has been proposed whereby two isolates belong to the same species if their average nucleotide identity is at least 95%, and this was shown closely to replicate current definitions based on DNA–DNA hybridization tests (Goris, 2007).

Several challenges remain to be overcome before routine species identification by whole-genome sequencing can become a reality for most pathogens. This includes achieving a turnaround time approaching hours for sequencing and analyzing the isolate data. This will depend on new rapid sequencing, new assembly techniques, new phylogenetic techniques and developing software and databases that are able to store large numbers of genomes. Software packages will need to be user-friendly and will need to yield clinically meaningful results. Quality-control procedures will need to be developed as well as criteria for run success, software validation and proficiency testing for laboratories. Before its deployment as a diagnostic system, a detailed clinical evaluation will be needed, including a comparison with currently used methods.

4.2. Testing for Antibiotic Resistance

In principle, it should be possible to predict resistance phenotypes by identifying genetic determinants of antimicrobial resistance and thus to permit rapid antibiotic treatment decision making. Currently, there are a few examples (including from S. aureus, Vibrio cholerae and Burkholderiadolosa McAdam et al., 2011) in which genetic determinants of antimicrobial resistance identified from whole-genome data are consistent with recorded variation in phenotype. These early data suggest that a sequence-based approach holds substantial promise. Indeed, a few methods for predicting antibiotic resistance from genetic rather than phenotypic data are already widely used: for example, the detection by PCR of mecA, which confers methicillin resistance in S. aureus and sequences that are known to encode resistance to isoniazid, rifampicin, ethambutol, aminoglycosides, capreomycin and fluoroqinolones in M. tuberculosis (known as the Genotype MTBDR assay) (Hilleman, 2005). In principle, whole-genome data could improve these tests, as the computational querying of the sequence may be more sensitive than using PCR primers, and it would be easier to search for more determinants.

Several challenges need to be overcome to achieve clinical adoption of whole-genome sequencing in resistance prediction. First, a comprehensive set of genetic determinants of antimicrobial resistance would need to be identified for each species. Such genetic determinants include: the presence of genes that confer resistance (such as TEM β-lactamase); point mutations in essential genes (such as in rpoB, which confers rifampicin resistance); and changes in the expression of genes (for example, reversion in the mutant operator sequences of E. coli ampC, leading to an increase in β-lactamase expression). Importantly, even where resistance determinants are well characterized, others may be revealed by further research. Furthermore, new mechanisms of antimicrobial resistance arise all too frequently: recent examples include quinolone resistance in Salmonella typhi, New Delhi metallo-β-lactamase-1 in Enterobacteriaciaeand multi-resistance in Neisseria gonorrheae (Bolan et al., 2012) Therefore, compiling a list of genetic determinants of resistance would be an ongoing task.

The sequence details of these determinants would need to be incorporated into a database that is kept up-to-date (to include novel resistance determinants) and that allows international data exchange via, for example, the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, USA, and the European Centre for Disease Prevention and Control (ECDC) in Stockholm, Sweden. Such a database would also facilitate identifying and reporting trends in resistance and new acquisition of resistance genes from other species. Predictions about the resistance and susceptibility from sequence data need to be accurate: falsely inferring susceptibility where the isolate is resistant represents a substantial risk to the patient. Therefore, performance needs to be established to high degrees of confidence in robust and well-powered clinical studies before deployment in a regulated environment. For example, in the United Kingdom, this would require Clinical Pathology Accreditation, and in the United States this would require approval from the Federal Food and Drug Administration. Therefore, although sequence data have the potential to support fast and cheap identification of resistance, we envisage a two-pronged approach that combines ongoing comparison of clinical outcome data with genetic data and phenotypic resistance screening. For example, ongoing phenotypic testing will be needed to identify new resistance and to keep the proposed database up-to-date.

4.3. Detecting Virulence Determinants

The genetic basis of many recognized virulence phenotypes is known, and yet our understanding of virulence factors is incomplete. The genome sequence of an isolate could yield information on all of the known virulence factors in one step and could create the opportunity for the discovery of new virulence factors through association studies that link the isolate genomic data with patient disease manifestation and outcome data. One early example of a finding from such an association study was the discovery of a prophage associated with whether Neisseria meningitides causes meningitis. Another example is the finding that non-synonymous mutations in specific genes in S. aureus occurred just before the development of invasive disease (Young et al., 2012).

More recently, whole-genome sequencing of isolates from major outbreaks has demonstrated the potential for identifying recognized virulence genes and pathogenicity gene clusters and for providing new understanding of virulence factors. For example, the recent analysis of whole-genome data from E. coli O104 showed the speed and precision of whole-genome sequencing. Draft sequencing took three days using the IonTorrent PGM and the first assembly was released two days later. Within a week of data becoming available, the strain was shown to be a novel E. coli O104:H4 variant that had acquired a prophage encoding Shiga toxin 2 and additional virulence and antibiotic resistance determinants. Similarly, sequencing of isolates from the 2010 Haitian Vibrio cholerae outbreak was claimed to be achieved in less than a day using the PacBio system, and sequence analysis allowed the detection and characterization of a toxin encoded by the CTX phage (Chin, 2011).

Similarly to the situation for antimicrobial resistance, identifying virulence determinants from analysis of whole genomic sequences is at an early stage, and substantial challenges need to be overcome before implementing this approach in a routine service environment. In particular, it requires the development of a database that includes all known virulence determinants and can incorporate new determinants. New software is needed to analyze genome sequences for the presence and absence of known virulence determinants as well as conducting ongoing association studies as described for antimicrobial resistance. The requirement for high sensitivity is generally lower for identifying virulence factors than for antimicrobial resistance, as identifying virulence has major clinical consequences in only a few cases.

4.4. Outbreak Detection and Surveillance

Genome sequences potentially provide a high-resolution, accurate and reproducible means for relating organisms. For example, sequencing the genomes of a diverse collection of Chlamydia trachomatis isolates has demonstrated the limitations of current clinical typing techniques for identifying phylogenetic relationships (Golubchik, et al. (2012). Compelling examples of the effectiveness of whole-genome analyses for unravelling the origins and dispersal of pathogens at regional and global scales have recently been published. This approach was used to investigate the emergence and global dispersal of ST239 isolates of methicillin-resistant S. aureus. In another example, the emergence of serotype 19A pneumococcal capsular variants, following the introduction in the United States of a pneumococcal vaccine, was documented and its spread tracked across the country. A comparative study of 154 whole genomes of Vibrio cholerae enabled the history of pandemic cholera over the past 50 years to be compiled, revealing that the seventh and current cholera pandemic has comprised three successive, partially overlapping waves with strong geographical and temporal structure. In Mycobacterium leprae, genome sequencing of isolates from 50 patients and 33 wild armadillos showed that these animals represent a major source of zoonotic transmission of leprosy in the southern United States. In a previous study, the spread of M. leprae was shown to follow human migration and historical trade routes. Finally, a comparison of 17 whole genomes and SNP typing in 286 globally representative isolates established strong geographical clustering in Yersinia pest is that is compatible with a Chinese origin for the Black Death pandemic (Morelli, 2010)

Early reports also strongly suggest that using sequencing to detect outbreaks that include person-to-person transmission within communities and hospitals is a major benefit to health care; this has been recently shown for S. aureus and C. difficile using rapid bench-top sequencing A report on using whole-genome sequencing to study a tuberculosis outbreak on Vancouver Island suggested that genealogical analysis of whole genomic sequences could be a major advance for tuberculosis contact tracing compared with the current cumbersome approaches. The current approaches depend heavily on identifying transmission networks through interviews, supplemented by M. tuberculosis-specific MIRU–VNTR typing, which is less discriminatory than whole-genome sequencing. Similar observations have been reported in a subset of MRSA isolates cultured from a hospital in Thailand, suggesting that phylogenetic analysis could be used to infer local hospital transmission Harris et al., (2010). The previously discussed studies of V. cholera and shigatoxin-producing E. coli O104 indicate that sequencing can also rapidly provide a clear understanding of the origins of a local outbreak.

Whole-genome sequencing is becoming the method of choice in research settings for monitoring pathogens over long time courses and wide geographical scales, as well as for identifying outbreaks. Sequence data gathered for diagnostic purposes can be accumulated for pathogen surveillance, outbreak detection and evolutionary studies. In principle, detection of an outbreak could occur as early as the first secondary case. Consequently, the deployment of sequencing technology for diagnostic purposes in local laboratories would also meet the needs for surveillance, as long as the genome sequences can be linked with the epidemiological information. To be fully useful, the data would have to be shared locally, nationally and internationally: new integrated approaches to store epidemiological and genomic data jointly are under development (Aanensen et al., 2009). It can be expected that national reference laboratories will adopt whole-genome sequencing as a single technology for typing all pathogens — replacing many species-specific typing methods — even if this is not done in the near future in routine diagnostic laboratories.

Future directions

Clinical microbiology is on the threshold of incorporating genome sequencing into routine practice. Although this Review focuses on the promise of this technology for bacterial pathogens, there is also rapid progress towards its adoption for viral, fungal and parasitic pathogen diagnostics and surveillance. It is likely that commercial developments based on sequencing technologies will focus on steps in current processing of cultured isolates that are discrete, high-cost and high-value. An example in which adoption may occur soon is in the analysis of mycobacterial cultures. Whole-genome sequencing is likely soon to provide, at a lower cost, all of the information that is currently provided by the MTBDR assay and also more details about species identification and resistance determinants. Similarly, sequencing could yield, at little additional cost, more definitive typing information than MIRU–VNTR testing. As discussed above, another setting in which adoption of whole-genome sequencing has already started is the investigation of putative outbreaks of major pathogens.

In this Review, we have focused on cases in which the pathogen has been cultured, but there is also potential for sequencing without culturing: that is, to sequence the entire DNA in a sample (for example, pus, cerebro-spinal fluid or sputum). Such a metagenomics approach has been used to define the microbiomes of diverse samples and environments (Cho and Blaser, 2012) Approaches such as bioinformatically masking the human sequences, then assembling pathogen genomes de novo or mapping reads to a reference genome from the hypothesized pathogen are likely to be useful, subject to the availability of sufficient data to overcome the low proportion of pathogen DNA in a clinical sample. In samples in which pathogen cell counts are low (such as M. tuberculosis that is present among many other organisms in sputum or the blood of a bacteraemic patient with 1–100 bacterial-colony-forming units per millitre), recovering complete bacterial genome sequences may depend on very cheap, fast sequencing or enhanced methods to deplete background material. New, very fast single-molecule long-read sequencing approaches should make it possible to sequence at great depth and low cost.

Adopting whole-pathogen sequencing would require major changes in the organization, skill mix and infrastructure of diagnostic laboratories and would therefore be disruptive, even if the main use of sequencing were after culture of the pathogen. Areas of focus will be strengthening competence in bioinformatics and software development. Advances are required in databases, efficient software and algorithms for analysis, software that automatically updates knowledge bases and sophisticated links between pathogen genomics databases and patient clinical record systems. To ensure that the benefits are accessible to the wider community, especially where a number of providers (commercial or otherwise) are developing systems, information needs to be shared in line with agreed standards. The opportunities for global surveillance of infectious diseases are vast, but political resolve is required to enable the sharing of sequence and meta-data on a global scale.

References

Aanensen, D. M., Huntley, D. M., Feil, E. J., al-Own, F. & Spratt, B. G. (2009) EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. PLoS ONE 4, e6968

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart (2010). Genomics and the future of conservation genetics. Nature Reviews Genetics11:697–709.

Altschul, S. F. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402

Auerbach, R. K., B. Chen, and A. J. Butte (2013). Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. Bioinformatics 29:1922–1924.

Batzoglou, S., D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger (2002). ARACHNE: a whole-genome shotgun assembler. Genome Research 12:177–189.

Bi, K., T. Linderoth, D. Vanderpool, J. M. Good, R. Nielsen, and C. Moritz (2013). Unlocking the vault: next-generation museum population genomics. Molecular Ecology 22:6018–6032.

Boetzer, M., and W. Pirovano (2012). Toward almost closed genomes with GapFiller. Genome Biology 13:R56.

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler, and W.Pirovano 2011. Scaffolding pre-assembled contigs using SSPACE.Bioinformatics 27:578–579.

Boisvert, S., F. Laviolette, and J. Corbeil (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of Computational Biology 17:1519–1533.

Bolan, G. A., Sparling, P. F. &Wasserheit, J. N. (2012).The emerging threat of untreatable gonococcal infection. N. Engl. J. Med. 366, 485–487

Bradnam, K., J. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:10

Bradnam, K., J. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:10.

Bradnam, K., J. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. GigaScience 2:10.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 18:188–196.

Chin, C. S. (2011).The origin of the Haitian cholera outbreak strain. N. Engl. J. Med.364, 33–42

Cho, I. &Blaser, M. J. (2012).The human microbiome: at the interface of health and disease.Nature Rev. Genet. 13, 260–270

Chu, T. -C., C.-H. Lu, T. Liu, G. C. Lee, W.-H. Li, and A. C.-C. Shih (2013). Assembler for de novo assembly of large genomes.Proceedings of the National Academy of Sciences 110:E3417–E3424.

Denisov, G., B. Walenz, A. L. Halpern, J. Miller, N. Axelrod, S. Levy, and G. Sutton (2008). Consensus generation and variant detection by Celera Assembler. Bioinformatics 24:1035–1040.

Earl, D., K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Research 21:2224–2241.

Earl, D., K. Bradnam, J. St. John, A. Darling, D. Lin, J. Fass, H. O. K. Yu et al. (2011. Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Research 21:2224–2241.

Ekblom, R. and Wolf, J. B. W. (2014), A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications, 7: 1026–1042.doi: 10.1111/eva.12178

Ekblom, R., and J. Galindo (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity107:1–15.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends in Ecology & Evolution 29:51–63.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backstrom, T. Kawakami, A. Kunstner (2012). The genomic landscape of species divergence in Ficedula flycatchers. Nature 491:756–760.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379.

English, A. C., S. Richards, Y. Han, M. Wang, V. Vee, J. Qu, X. Qin (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One 7:e47768.

Fraser-Liggett CM (2005) Insights on biology and evolution from microbial genome sequencing.Genome Res 15: 1603–1610.

Fraser, D. J., and L. Bernatchez (2001). Adaptive evolutionary conservation: towards a unified concept for defining conservation units.Molecular Ecology 10:2741–2752.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32(Suppl 1):D258–D261.

Gnerre, S., I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences 108:1513–1518.

Golubchik, T. et al. (2012). Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. Nature Genet. 44, 352–355

Goris, J. (2007).DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91

Harris, S. R. et al. (2010) Evolution of MRSA during hospital transmission and intercontinental spread. Science 327, 469–474

Havlak, P., R. Chen, K. J. Durbin, A. Egan, Y. Ren, X.-Z. Song, G. M. Weinstock (2004). The Atlas genome assembly system.Genome Research 14:721–732.

Hernandez, D., P. François, L. Farinelli, M. Østerås, and J. Schrenzel (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Research 18:802–809.

Herrera, C. M., and P. Bazaga (2011). Untangling individual variation in natural populations: ecological, genetic and epigenetic correlates of long-term inequality in herbivory. Molecular Ecology 20:1675–1688.

Hilleman, D. (2005) Use of the genotype MTBDR assay for rapid detection of rifampin and isoniazid resistance in Mycobacterium tuberculosis complex isolates. J. Clin. Microbiol. 43, 3699–3703

Huang, X., J. Wang, S. Aluru, S.-P. Yang, and L. Hillier (2003). PCAP: a whole-genome assembly program. Genome Research13:2164–2170.

Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. Otto (2013).

REAPR: a universal tool for genome assembly evaluation. Genome Biology 14:R47.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature431:931–945.

Jolley, K. A. (2012).Ribosomal multi-locus sequence typing: universal characterisation of bacteria from domain to strain. Microbiology 158, 1005–1015 This is a database system for whole genomes that provides a smooth transition for users from working with MLST to working with genomes.

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

Kanehisa, M., and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28:27–30.

Kohn, M. H., W. J. Murphy, E. A. Ostrander, and R. K. Wayne (2006). Genomics and conservation genetics. Trends in Ecology & Evolution 21:629–637.

Koren, S., M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30:693–700.

Lampa, S., M. Dahlo, P. Olason, J. Hagberg, and O. Spjuth (2013). Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. GigaScience 2:9.

Larsen, M. V. (2012).Multilocus sequence typing of total-genome-sequenced bacteria. J. Clin. Microbiol. 50, 1355–1361

Lee, E., G. Helt, J. Reese, M. C. Munoz-Torres, C. Childers, R. M. Buels, L. Stein (2013). Web Apollo: a web-based genomic annotation editing platform. Genome Biology 14:R93.

Li, H. (2012). Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28:1838–1844.

Li, H., and R. Durbin (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Li, H., and R. Durbin (2011). Inference of human population history from individual whole-genome sequences. Nature 475:493–496.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18.

Marçais, G., and C. Kingsford (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics27:764–770.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9:387–402.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal 17:10–12.

McAdam, P. R., Holmes, A., Templeton, K. E. & Fitzgerald, J. R. (2011).Adaptive evolution ofStaphylococcusaureus during chronic endobronchial infection of a cystic fibrosis patient. PLoS ONE 6, e24301.

Medini, D., Donati, C., Tettelin, H., Masignani, V. &Rappuoli, R. (2005) The microbial pan-genome. Curr. Opin. Genet. Dev. 15, 589–594

Miller, J. R., S. Koren, and G. Sutton (2010). Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.

Miller, J. R., S. Koren, and G. Sutton (2010). Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.

Morelli, G. (2010).Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nature Genet. 42, 1140–1143

Nagarajan, N., and M. Pop (2013). Sequence assembly demystified. Nature Reviews Genetics 14:157–167.

Narum, S. R., C. A. Buerkle, J. W. Davey, M. R. Miller, and P. A. Hohenlohe (2013). Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology 22:2841–2847.

Narzisi, G., and B. Mishra (2011). Comparing de novo genome assembly: the long and short of it. PLoS One 6:e19175.

Novák, P., P. Neumann, J. Pech, J. Steinhaisl, and J. Macas (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–793.

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin, D. G. Scofield, F. Vezzi (2013). The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584.

Ouborg, N. J., C. Pertoldi, V. Loeschcke, R. Bijlsma, and P. W. Hedrick (2010). Conservation genetics in transition to conservation genomics. Trends in Genetics 26:177–187.

Pääbo, S., H. Poinar, D. Serre, V. Jaenicke-Despres, J. Hebler, N. Rohland, M. Kuch (2004). Genetic analyses from ancient DNA.Annual review of genetics 38:645–679.

Parra, G., K. Bradnam, and I. Korf (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics23:1061–1067.

Phillippy, A., M. Schatz, and M. Pop (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biology 9:R55.

Primmer, C. R. (2009). From conservation genetics to conservation genomics. Annals of the New York Academy of Sciences1162:357–368.

Romanov, M. N., E. M. Tuttle, M. L. Houck, W. S. Modi, L. G. Chemnick, M. L. Korody, E. M. S. Mork (2009). The value of avian genomics to the conservation of wildlife. BMC Genomics 10(Suppl 2):S10.

Schatz, M. C., A. L. Delcher, and S. L. Salzberg (2010a). Assembly of large genomes using second-generation sequencing. Genome Research 20:1165–1173.

Schatz, M. C., B. Langmead, and S. L. Salzberg (2010b). Cloud computing and the DNA data race. Nature Biotechnology 28:691.

Shokralla, S., J. L. Spall, J. F. Gibson, and M. Hajibabaei (2012). Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21:1794–1805.

Simpson, J. T., and R. Durbin (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Research22:549–556.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and İ. Birol (2009). ABySS: a parallel assembler for short read sequence data. Genome Research 19:1117–1123.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15:121–132.

Smeds, L., and A. Künstner (2011). ConDeTri – a content dependent read trimmer for Illumina data. PLoS One 6:e26314.

Steiner, C. C., A. S. Putnam, P. E. A. Hoeck, and O. A. Ryder (2013). Conservation genomics of threatened animal species. Annual Review of Animal Biosciences 1:261–281.

Tsai, I. J., T. D. Otto, and M. Berriman (2010). Method improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology 11:R41.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5:247–252.

Vijay, N., J. W. Poelstra, A. Künstner, and J. B. W. Wolf (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology 22:620–634.

Warren, R. L., G. G. Sutton, S. J. M. Jones, and R. A. Holt (2007). Assembling millions of short DNA sequences using SSAKE.Bioinformatics 23:500–501.

Wenjun Li, Didier Raoult& Pierre-Edouard Fournier (2009). Bacterial strain typing in the genomic era.FEMS Microbiol Rev 33 892–916.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.

Wong, P., E. Wiley, W. Johnson, O. Ryder, S. O'Brien, D. Haussler, K.-P. Koepfli (2012). Tissue sampling methods and standards for vertebrate genomics. GigaScience 1:8.

Xavier Didelot, Rory Bowden Daniel J., Wilson Tim E. A. Petoand Derrick W. Crook (2012). Transforming clinical microbiology with bacterial genome sequencing. Nature Reviews Genetics 13, 601-612

Yandell, M., and D. Ence (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13:329–342.

Young, B. C. et al. (2012). Evolutionary dynamics of Staphylococcus aureus during progression from carriage to disease. Proc. Natl Acad. Sci. USA 109, 4550–4555

Zavodna, M., C. E. Grueber, and N. J. Gemmell (2013). Parallel tagged next-generation sequencing on pooled samples – a new approach for population genetics in ecology and conservation. PLoS One 8:e61471.

Zerbino, D. R., and E. Birney 2008. Velvet: algorithms for de novo short read 7 assembly using de Bruijn graphs. Genome Research18:821–829.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang (2012). The oyster genome reveals stress adaptation and complexity of shell formation. Nature 490:49–54.

Zimin, A., G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke (2013). The MaSuRCA genome Assembler. Bioinformatics29:2669–2677.