Genome-wide Variant Discovery using Sequence Assembly, Mapping and Population-Wide Analysis

3

Simone Coughlan, Sofia Barreira, Cathal Seoighe and Tim Downing

Abstract

The transformation of DNA sequencing technologies has enabled more powerful and comprehensive genetic profiling of microbes. The sheer number of informative loci provided by genome-sequencing allows the investigation of structural variation and horizontal gene transfer as well as delivering novel insights into genetic origins, evolution and epidemiological history. Microbial genomes can be sequenced en masse at high coverage but have associated challenges of high mutation rates and low conservation of genome structure. Consequently, detecting changes in DNA sequences requires a nuanced approach specific to the organism, availability of similar genomes, and types of variation. Here, we outline the high power of genome-sequencing to detect a wide scope of polymorphism classes. Samples without related species on which to scaffold a genome sequence require specific assembly methods that can be enhanced by progressive procedures for improvement. Polymorphism identification depends on genome structure, and error rates in closely related specimens can be reduced by incorporating population-level information. The development of genome analysis platforms is hastening the optimization of variant discovery and has direct applications for pathogen surveillance. Robust variant screening facilitates more sensitive scrutiny of population history, including the origin and emergence of infectious agents, and a deeper understanding of the selective processes that shape microbial phenotypes.

Background

Microbial genomics is driven by the need to identify molecular markers and genetic switches associated with novel phenotypes. This is most extensively applied to address infectious disease and evolution but also is used for improving food, energy, water and biomolecule production (Suter *et al.*, 2006). The core aim is to link the trait of interest to a defined cellular signature, principally a metabolic or regulatory change. Distinguishing phenotypes at a genetic level provides an enhanced resolution of the molecular events associated with natural variation due to the density of markers.

Variability at the level of amino acids in peptides, repetitive DNA copy numbers and individual DNA nucleotides can provide sufficient power for discriminating traits: sample typing protocols have been developed on this basis (Wren, 2000). The most significant limitations for strain profiling are an adequate number of informative markers and the potential to overlook novel variation at other loci (Achtman, 2008). Although the first microbial genome sequence (RNA virus bacteriophage MS2) was completed in 1976 (Fiers *et al.*, 1976), and the era of microbial genomics was proposed to have begun in 1995 (Rasko and Mongodin, 2005) with the bacterial genome of *Haemophilus influenzae* (Fleischmann *et al.*, 1995), it was the recent development of DNA sequencing technologies more efficient than traditional capillary (Sanger) sequencing that made genome-sequencing accessible (Margulies *et al.*, 2005). Since a 100-fold improvement developed for the *Mycoplasma genitalium* genome (Margulies *et al.*, 2005), enhancements have continued and

now render the approach amenable for studying variation in any organism (Nowrousian, 2010).

The main challenges linked to genome sequencing include DNA isolation, genome assembly and mutation discovery through read-mapping (Nielsen *et al.*, 2011). In the context of all microbial life (viruses, Archaea, fungi, protozoa and algae), most work has been completed on bacteria, and consequently this chapter mainly concentrates on this class of microorganism. Although the major disease burden posed by microbial pathogens necessitates a focus on the genomics of infection and epidemiology (Walker *et al.*, 2012; World Health Organization, 2012), a significant asset of genomic technologies is their wide applicability to all organisms, including non-model ones.

A range of characteristics render microbial genomes ideal for assessment with genomic technology and for evolutionary studies. Firstly, their compact genome sizes makes genome-sequencing cheaper. Secondly, short generation times and fast mutation rates mean genetic changes can be observed in real time (Wren, 2000) and this can be performed *in vitro* and *in vivo* to test hypotheses (Blount *et al.*, 2012). Thirdly, microbial adaptability to extensive genome rearrangements, karyotype changes and gene transfer means that the range of mutation types is extensive (Frost *et al.*, 2005).

This section highlights the attributes of microbial variants in the context of genome-sequencing. The genome of a new microbial species can be assembled without the use of a related reference genome, and this draft sequence can be improved. This allows the discovery of all DNA-level mutations, which include single nucleotide polymorphisms (SNPs) where one DNA base is replaced by another, as well as larger structural variants (SVs): these are alterations of two or often many DNA bases. Genome sequences can be compared between species and within populations, for which unique methods have been developed to improve variant ascertainment. *De novo* polymorphism identification and the nuances of SV discovery are explored in particular. A range of tools and schematics help with examining microbial genomic variation: the most significant are highlighted here.

Microbial genetic variation and genome sequencing

The section outlines the main forms of diversity that can be assessed using whole-genome sequencing: SNPs and SVs, as well as other types of variation that cannot be resolved using current sequencing technologies. Greater inference power can be derived from investigating genome-wide SNP variation rather than with traditional approaches like multi-locus sequence typing (MLST).

MLST is a scalable DNA-based sequence analysis scheme for investigating local and global pathogen epidemiology that examines SNPs at multiple genes. It was first proposed for *Neisseria meningitidis* in 1998 (Maiden *et al.*, 1998): polymerase chain reaction (PCR) is used to amplify 400 to 600 bp DNA fragments of six to ten housekeeping genes (Enright and Spratt, 1999; Maiden, 2006), which are capillary sequenced (Sanger and Coulson, 1975; Sanger *et al.*, 1977). The MLST variation for each gene is combined to produce an allelic profile that defines the sequence type. Allele profiles are stored in MLST databases that can be accessed through PubMLST (www.pubmlst.org) (Enright and Spratt, 1999; Urwin and Maiden, 2003; Larsen *et al.*, 2012) and can be used for classification of bacterial types (Cheng *et al.*, 2011). MLST is a routine and powerful tool to compare isolates on a global scale that has been primarily used for bacterial specimens. It has been used to examine the evolutionary history of methicillin-resistant Staphylococcus aureus (Enright *et al.*, 2002), and the relationship between virulence and genotypes of Streptococcus pneumonia (Dicuonzo *et al.*, 2002; Brueggemann *et al.*, 2003). Methods are being developed to parallelise high-throughput genotyping with more efficient sample phenotyping using robots (Haase *et al.*, 2011).

MLST is a costly and time-consuming procedure (Larsen *et al.*, 2012) and lacks the discriminatory power of genome sequencing (Harris *et al.*, 2012). Some of the most pathogenic microbes exhibit low levels of DNA diversity and are genetically monomorphic – often a reflection of their recent rapid spread (Achtman, 2012). Examples include *Salmonella enterica* serovar Typhi (Holt *et al.*, 2008), *Yersinia pestis* (Achtman

et al., 1999), *Bacillus anthracis* (Van Ert *et al.*, 2007) and the *Mycobacterium tuberculosis* complex (Sreevatsan *et al.*, 1997). Techniques such as pulsed field gel electrophoresis (PGFE), multilocus variable number of tandem repeats analysis (MLVA) and spoligotyping (spacer oligonucleotide typing) are more suitable for the analysis of monomorphic species – as is genome sequencing (Achtman, 2008). Furthermore, incorporating high-throughput components to assess MLST data require comprehensive informatics (e.g. Laboratory Information Management Systems). And while many bacterial species have MLST profiles, few others have copious MLST data. Consequently, MLST requires much of the same genetics and informatics infrastructure associated with genomics but samples far fewer genes.

Recent advances in DNA sequencing can be defined as a revolution because of the order of magnitude increase in efficiency (Margulies *et al.*, 2005) beyond capillary and shotgun sequencing (Fleischmann *et al.*, 1995). This continued transformation enables the rapid sequencing of all DNA in a sample. This is being adopted as an extension of MLST analysis through the Bacterial Isolate Genome Sequence Database (Jolley and Maiden, 2010), where genome-based results can be integrated with MLST work. However, genome sequencing requires neither previous work nor prior information from related species. The cost of genome sequencing has dropped approximately ten-fold every 5 years (Service, 2006) to the point that it is below the cost of MLST (Larsen *et al.*, 2012). Despite potential informatics limitations of data processing, genome sequencing can give clinically informative results in a short time frame (Glenn, 2011). Consequently, the adoption of genome sequencing as the standard method of microbial profiling can provide additional information not obtainable through MLST.

One of the major advantages of whole genome sequencing is that it can be used to detect SVs, which moderate gene expression in a different and dose-dependent manner more frequently than SNPs (Medvedev *et al.*, 2009, 2010). Genome sequencing is conducted using DNA reads with a saturation of coverage, and so most SVs can be inferred from where the coverage changes significantly (Nielsen *et al.*, 2011). SVs are polymorphisms that affect the number of nucleotides in the genome, and are defined here to include insertions, deletions, translocations, inversions and copy number variation (CNVs) (Fig. 3.1). An insertion is a sequence in the sample genome that is absent in the reference sequence, and can be caused by errors in DNA replication due to DNA polymerase slippage at repetitive sequences. The opposite, when a sequence in the reference is not present in the sample, is called a deletion – insertions and deletions (indels) can range from single to many bases. CNVs are the repetition of a locus: tandem duplications are the simplest form of CNV; but if the duplicated element exists on another chromosome, it is a translocation. Inversions are the re-orienting of a locus in the reverse direction, often due to homologous recombination.

The comprehensive profiling of SNP and SV changes provided by genome sequencing enhances our understanding of drug resistance, virulence factors and enables more accurate pathogen tracing. Although DNA microarrays can capture a larger panel of variants than MLSTs and have been successfully applied to viral infections (Chiu *et al.*, 2008), they are inherently limited by not discovering novel mutations. The origin and spread of viral infections like influenza has been investigated more robustly using genomics (Smith *et al.*, 2009). For bacteria, this has been successfully applied to the study of adaptive evolution of *Staphylococcus aureus* during chronic cystic fibrosis infection (McAdam *et al.*, 2012). It has documented global transmission of *Vibrio cholerae* and its acquisition of antibiotic resistance elements (Mutreja *et al.*, 2011). The *E. coli* O104:H4 draft genome was completed just three days after DNA isolation during an outbreak, and the consensus genome sequence was reconstructed within two days (Rohde *et al.*, 2011). Hospital transmission and control of a methicillin-resistant *Staphylococcus aureus* (MRSA) outbreak was implemented using genome sequencing, which implicated a single carrier as the likely originator of the disease (Harris *et al.*, 2012). Given the low cost of bacterial genome sequencing, estimated at just £95 (~$160) per isolate for MRSA, affordable routine genome-based surveillance would have detected the outbreak earlier. These studies highlight

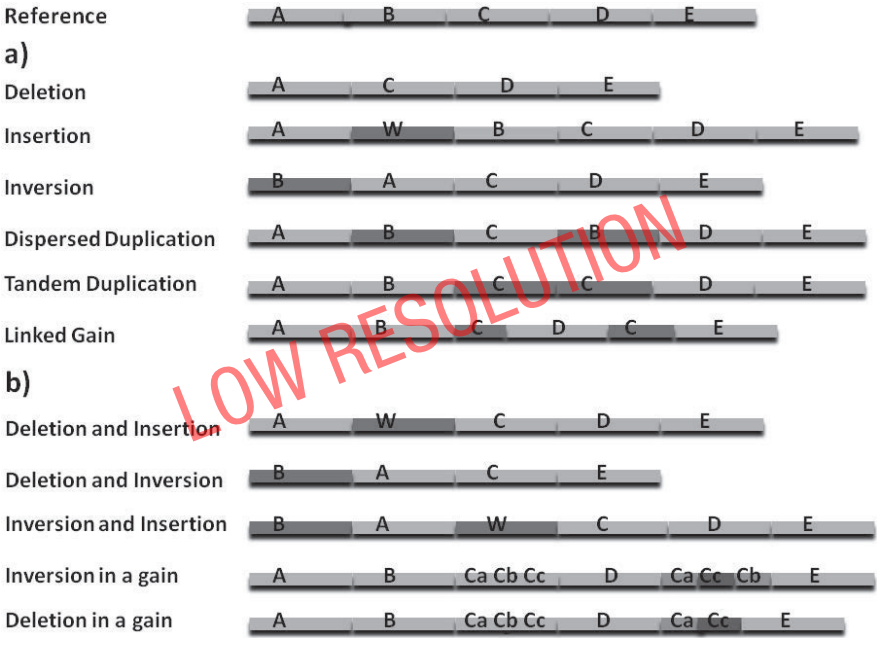

Figure 3.1 Types of (a) simple and (b) complex structural variant mutations. A reference sequence with five loci (A, B, C, D, E): (a) Simple SVs: a deletion (no B); an insertion (W); an inversion (A and B switched); a dispersed duplication (B repeated between C and D); a tandem duplication (C repeated between C and D); and a linked gain (small C element repeated between C and D). (b) Complex SVs: Shown in this diagram are co-occurrence of a deletion and an insertion (no B and W entered); co-occurrence of a deletion and an inversion (no C and A and B switched); co-occurrence of an inversion and an insertion (W entered and A and B switched); an inversion within a copy number gain (C duplicated and Cb and Cc switched); and a deletion within a copy number gain (C duplicated and Cb deleted).

another salient benefit of genome-sequencing of small genomes: many samples can be pooled on a single lane of a sequencing machine, each having its own unique set of DNA reads with a short DNA adaptor tag added during library preparation. These sample barcodes do lead to some loss of total output, but dramatically increase the total number of samples that can be processed per lane to 96 or more. Multiplexing can also be adopted to conduct large-scale pooling, a form of high-throughput MLST (O'Roak *et al.*, 2012).

Horizontally inherited DNA and genome sequencing

A major component of microbial genetic variation is horizontal gene transfer (HGT, also known as lateral gene transfer). This is the movement of genetic material in a way not facilitated by vertical transfer (sexual or asexual reproduction) – usually by transformation, transduction or conjugation (Ochman *et al.*, 2000). HGT can be considered in terms of a core genome – a set of genes encoding fundamental metabolic functions different to the non-essential accessory genome (Schmidt and Hensel, 2004). The accessory genes are significant because they frequently encode traits associated with drug resistance, virulence, and the ability to degrade xenobiotic compounds. This can contribute to the adaption of a microorganism to its environment and also to its diversification (Juhas *et al.*, 2009). Classical vectors of HGT are plasmids and bacteriophages, which can be detected from whole genome sequencing. A plasmid is a self-replicating double stranded DNA molecule (replicon) that is typically circular. However, linear plasmids with either a covalently closed hairpin loop or attached protein at each end have also been discovered in spirochaetes, Gram-positive and Gram-negative bacteria (Hinnebusch and Tilly, 1993; Frost *et al.*, 2005). Phage genomes can be up to several hundred kb in length and are

composed of single- or double- stranded DNA (Pedulla *et al.*, 2003). Temperate bacteriophages mediate HGT during lysogenization: the phage DNA integrates into the bacterial chromosome as a prophage and replicates with it, though in some cases the phage can replicate independently as a circular or linear plasmid (Canchaya *et al.*, 2003; Frost *et al.*, 2005).

In addition, an increasingly diverse array of mobile genetic elements (MGEs) such as genomic islands, mobilizable transposons and conjugative transposons have been discovered (Osborn and Boltner, 2002) that can encode genes enabling HGT within or between bacterial cells (Frost *et al.*, 2005). Genomic islands (GIs) are gene clusters of between 10 and 500 kb acquired by HGT (Osborn and Boltner, 2002). They were first discovered in uropathogenic *E. coli* as virulence genes with distinctive GC content and codon usage patterns compared to adjacent DNA (Hacker *et al.*, 1983). GIs can include a broad range of MGEs such as prophages, integrative conjugative elements, integrons, conjugative transposons and integrated plasmids (Langille *et al.*, 2010) and are usually chromosomally inserted near transfer RNA genes flanked by short direct repeat structures. GIs contain genes associated with mobility (Langille *et al.*, 2010) and include pathogenicity, fitness, symbiosis, metabolic or resistance islands depending on their functional gene composition (Hentschel and Hacker, 2001).

Although horizontal inheritance is common in prokaryotes, it is also prevalent in some eukaryote microbes as well. The protozoan parasites *Leishmania infantum* and *Leishmania major* can exchange eukaryotic episomes of 45+ kb (Coelho *et al.*, 2012): these are formed by homologous recombination between repeats and are small extra-chromosomal pieces of closed circular DNA that can replicate independently of the genome (Leprohon *et al.*, 2009; Downing *et al.*, 2011). They are analogous to plasmids in prokaryotes but differ from prokaryotic episomes that integrate into host chromosomes (Hinnebusch and Tilly, 1993).

Whether mediated by plasmids, bacteriophages, MGEs or GIs, HGT results in a mixed agglomeration of genes with different origins on a contiguous chromosome. By examining the phylogenetic distribution of mutations at these loci, genes resulting from HGT can be identified where they have sharply distinct genetic signatures (Lawrence and Ochman, 1998). Bacteria have characteristic patterns of substitutions (Hooper and Berg, 2002), codon bias (Sharp and Matassi, 1994), GC content (Muto and Osawa, 1987) and oligomer frequencies (Burge *et al.*, 1992) due to different environmental selection and mutational pressures (Sueoka, 1988; Rocha and Danchin, 2002). These characteristics can be used to identify foreign DNA where it has a composition that differs from those of the host genome. For microbes with extensive tolerance of HGT that resemble ecotypes more than species, the core genome will gradually diminish and the accessory genome continually expand as more strains are sequenced (Achtman, 2008). Consequently, only genome sequencing provides a complete picture of HGT signals, which can be tested using tools like Pyphy (Sicheritz-Ponten and Andersson, 2001), AMPHORA (Wu and Eisen, 2008) and PhyloNet (Than *et al.*, 2008).

Alternative genotyping approaches

Although genome-sequencing can provide a higher total density of molecular markers, it has low sensitivity for detecting variation in highly repetitive regions (Medvedev *et al.*, 2009). This section details the limitations of genome sequencing and discusses some alternative methods for distinguishing microbial strains based mainly on restriction enzyme digestion of DNA, repeat-counting and protein polarity. These methods have differing levels of resolution: pulsed field gel electrophoresis (PFGE) and clustered regularly interspaced palindromic repeat (CRISPR) analysis index variation at rapidly evolving loci and so are more applicable to population-level analysis. MLST as well as multi-locus enzyme electrophoresis (MLEE) and multi-locus microsatellite typing (MLMT) examine diversity at conserved housekeeping genes likely to be selectively neutral and so document deeper evolutionary history (Maiden, 1998; Singh *et al.*, 2006). The first widely used method of characterizing bacterial SNPs used the 16S rRNA gene: this contains

nine hypervariable regions (V1–V9) (see Fig. 9.1) flanked by highly conserved regions for PCR primer design (Van de Peer *et al.*, 1996) that can be used for species identification.

PFGE separates DNA fragments by size on agarose gels using alternately pulsed electric fields (Schwartz and Cantor, 1984). The DNA is digested with restriction enzymes to produce fragments (Schwartz and Cantor, 1984; Singh *et al.*, 2006): if the restriction enzymes cut the DNA at different positions, a different banding pattern for each isolate is obtained. PFGE can be modified by varying an electrical pulse applied to the gel (Singh *et al.*, 2006) – though contour clamped homogenous electric field is the most widely used approach (Goering, 2010), other techniques based on field inversion (Carle *et al.*, 1986), orthogonal fields (Carle and Olson, 1984) and transverse alternating fields are also used (Gardiner *et al.*, 1986). PFGE results can be shared (PulseNet, www.cdc.gov/pulsenet) and standardized protocols for pathogen surveillance, including *E. coli* O157:H7, non-typhoidal *Salmonella* serotypes, *Listeria monocytogenes* and *Shigella* have been developed (Swaminathan *et al.*, 2001). Although PFGE is a simple and inexpensive method, it can take days to complete and the reproducibility of results varies between labs (Noller *et al.*, 2003). Moreover, PFGE requires longer strain culturing than genomics (Lindstedt, 2005), which has moved from requiring 2 ng of material (Carter *et al.*, 2003) to only 10–20 cells (~100 pg in humans) (Peters *et al.*, 2012).

CRISPR elements are a family of 21 to 37 bp DNA repeats found in many prokaryotes and most Archaea (Haft *et al.*, 2005) and were first discovered in *E. coli* K12 (Ishino *et al.*, 1987). They are separated by highly variable non-repetitive spacers encoding small RNAs (Haft *et al.*, 2005) about the same size as the repeats (Jansen *et al.*, 2002). These spacers originated from MGEs and mutate rapidly (Haft *et al.*, 2005). CRISPR-associated genes encode conserved proteins (Haft *et al.*, 2005; Al-Attar *et al.*, 2011) that act as acquired immunity against viruses and conjugative elements by recognizing and inactivating foreign DNA (He and Deem, 2010). Spoligotyping amplifies the whole CRISPR region using the direct repeat region as a PCR target and is primarily used for genotyping *Mycobacterium tuberculosis*. The DNA product is hybridized to a membrane containing oligonucleotides homologous to the spacer sequences that differentiate samples based on spacer type presence and absence (van Soolingen *et al.*, 1993; Kamerbeek *et al.*, 1997; van der Zanden *et al.*, 2002; Al-Attar *et al.*, 2011). *M. tuberculosis* has CRISPR loci consisting of 36 bp repeats and unique spacers of 35–41 bp (Groenen *et al.*, 1993). *Mycobacterium* spoligotype variation originates from IS*6610* insertion element transposition, homologous recombination (Groenen *et al.*, 1993) and replication slippage that deletes spacers (Driscoll, 2009). However, spoligotyping has less discriminatory power than IS*1160* RFLP typing (Kamerbeek *et al.*, 1997; Kremer *et al.*, 1999).

Microsatellites are short variable number tandem repeats (VNTRs) of one to six bases that mutate at rates several orders of magnitude higher than SNPs (Ellegren, 2000). Polymerase slippage during DNA replication in the absence of DNA repair can result in changes in the number of repeat units (Strand *et al.*, 1993). MLMT is particularly useful for genetically monomorphic organisms and is also used for examining eukaryotic microbes such as *Trypanosoma* (Llewellyn *et al.*, 2009) and *Leishmania* (Bulle *et al.*, 2002). Conserved flanking regions provide a PCR template for MLMT profiling of the whole repeat region based on size – the same principle applies to VNTR and MLVA, which have characterized pathogens *Neisseria meningitidis* (Schouls *et al.*, 2006), *Legionella pneumophila* (Pourcel *et al.*, 2007) and *Leptospira interrogans* (Slack *et al.*, 2007). MLVA has equivalent sensitivity but higher specificity compared to PFGE for *E. coli* O157:H7 (Noller *et al.*, 2003).

Although MLEE was first used to analyse genetic variation in *Drosophila* (Hubby and Lewontin, 1966; Lewontin and Hubby, 1966), it is a standard method for exploring bacterial genetic diversity and epidemiology (Selander *et al.*, 1986). MLEE discriminates the electrophoretic mobilities of 20 or more intracellular housekeeping enzymes where amino acid diversity affecting the electrostatic charge is present (Enright and Spratt, 1998; Stanley and Wilson, 2003). MLEE has been used to examine *Yersinia* (Dolina and

Peduzzi, 1993), *E. coli* (Souza *et al.*, 1999), *Salmonella* (Beltran *et al.*, 1988), *Trypanosoma brucei* (Stevens and Tibayrenc, 1995), *Plasmodium falciparum* (Abderrazak *et al.*, 1999) and *Leishmania* (Hamarsheh, 2011). MLEE too has a number of limitations: it is time-consuming, largely ignores variation at non-charged amino acids, and can be difficult to replicate (Enright and Spratt, 1998; Gil-Lamaignere *et al.*, 2003).

Genome assembly: methods, tools and improvement

The construction of a draft genome sequence is a crucial step for understanding the biology of all species. The discovery of novel viral and bacterial samples means that *de novo* genome assembly can discover novel chromosomal architecture. Moreover, samples with known reference sequences that possess flexible plastic genomes (with high concentrations of repetitive sequence or extensive HGT) deserve additional scrutiny using the unbiased assessment provided by *de novo* genome assembly. Additional motivation for improved reference assemblies stems from the numbers of incomplete genomes compared to that of permanent drafts (14,743 bacterial genomes versus 1,781 drafts; 265 versus 40 for Archaea; 2,769 versus 34 for eukaryotes; taken from the Genomes OnLine Database in April 2013, www.genomesonline.org) (Pagani *et al.*, 2012). Genome sequencing typically produces overlapping redundant reads that are aligned against a related genome (Eisenstein, 2012), or are used to construct a minimal set of consensus sequences – this is genome assembly (Fig. 3.2). The genome assembly problem is where the DNA reads have to be aligned against one another: it typically arises in the context of non-model microbes lacking a closely related reference genome.

Genome assembly methods

Experimental design, read quality and read length are major determinants of assembly results (Salzberg *et al.*, 2012). Genome-sequencing platforms deliver short reads whose ends can be paired by sequencing both 5′ and 3′ parts of a single fragment of DNA. Although longer read lengths provide better scope for developing a map of the entire genome, paired reads are essential for developing an informative assembly because they link disparate elements (Miller *et al.*, 2010). Paired-end reads are oriented towards each other separated by an un-sequenced insert component that may be 200–600 bp with a low variance in size for a given mean insert length (Medvedev *et al.*, 2010) (Fig. 3.3). Less commonly, paired-end reads can also overlap each other such that each has a shared segment of 20 bases.

Mate-pair reads are oriented away from each other and because they have much longer insert sizes, often of the order of 4–6 kb, they provide additional resolution (Medvedev *et al.*, 2010). They are created by selecting long DNA fragments, circularising these sequences with an internal adaptor, and selecting the DNA fragments containing the internal adaptor after random shearing for amplification. If one of the mate pair reads can be uniquely mapped, untangling repetitive regions longer than paired-end read insert sizes is possible (Li and Homer, 2010). In prokaryotes, resolution may be improved by using shorter insert sizes for mate pairs (< 1 kb) determined by the genome's repetitiveness (Wetzel *et al.*, 2011). A broad range of genome-wide repeat sizes requires a more extensive array of insert sizes: consequently, assemblies created from multiple libraries with a wider range of insert sizes are more accurate (Flicek and Birney, 2009).

De novo genome reconstruction by assembly of DNA reads can be carried out using overlap-layout-consensus (OLC) (Flicek and Birney, 2009) or de Bruijn graph methods that use subsequences to reduce computational memory requirements through a smaller search space (Li *et al.*, 2012). A *k*-mer is short sequence of a defined length (k) that is odd value to avoid palindromes and is smaller than the DNA read length. DNA sequenced using capillary sequencing methods can be assembled by OLC-based methods where there are sufficiently long homologous regions between the sequences for unique overlaps. For this process, the potential overlap between the reads is computed by global or local nucleotide alignment: this is based on the seed-and-extend approach where the matching sequences are iteratively joined together (Altschul *et al.*, 1990) (Fig. 3.4). These distances are quantified as a graph or

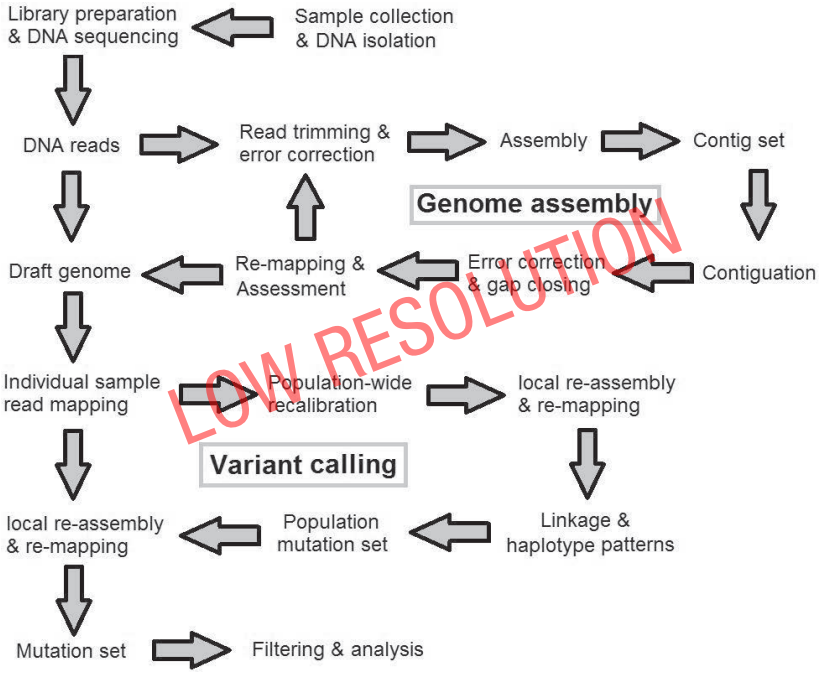

Figure 3.2 Genome assembly and variant discovery overview. An outline of genome assembly and variant discovery, which can be performed on individual specimens or on large sample sets. Genome assembly: DNA reads generated by genome sequencing are trimmed and corrected before assembly to produce an initial contig set (Fig. 3.4) that can be contiguated. Once these contigs are iteratively base-corrected and gaps are minimized, reads are re-mapped to assess the level of improvement. This is a repeatable process that can combine multiple DNA read libraries (Fig. 3.3) as well as other data sources like capillary reads, BACs, fosmids, optical maps and genetic maps to produce a draft genome sequence. Variant calling: DNA reads for each sample are mapped against the reference genome, with optional local re-assembly and re-mapping where required. This produces mutation sets for individual samples, and for population panels: population data can be iteratively recalibrated to improve the overall quality using genotype likelihoods. Linkage between mutations and the extended local haplotype can be used to further improve the population-wide mutation set. Finally, the individual and/or population variation should be filtered and repetitive regions should be masked out.

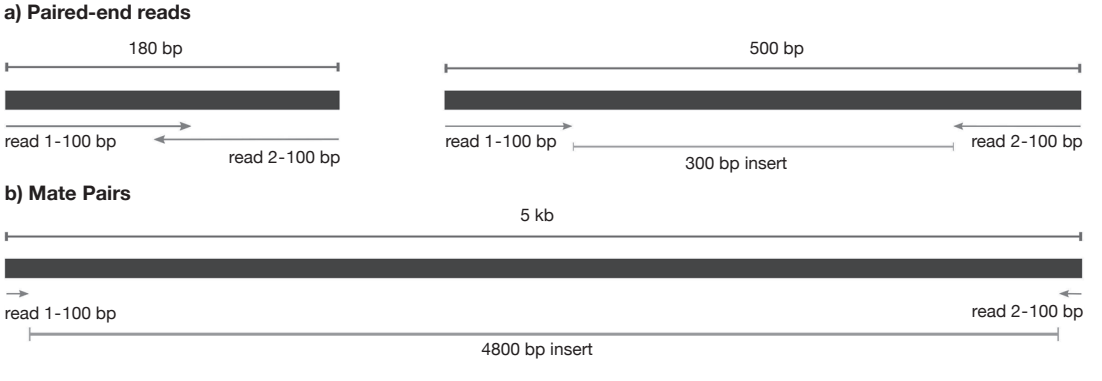

Figure 3.3 Paired-end and mate-pair read types. (a) Paired-end reads of 100 bases (grey) face each other (arrows) and are comprised of either overlapping pairs of reads (left) or have an insert element of 300–500 bases between the read sequences (right). (b) Mate pair reads have larger insert elements during sequencing (1–10kb).

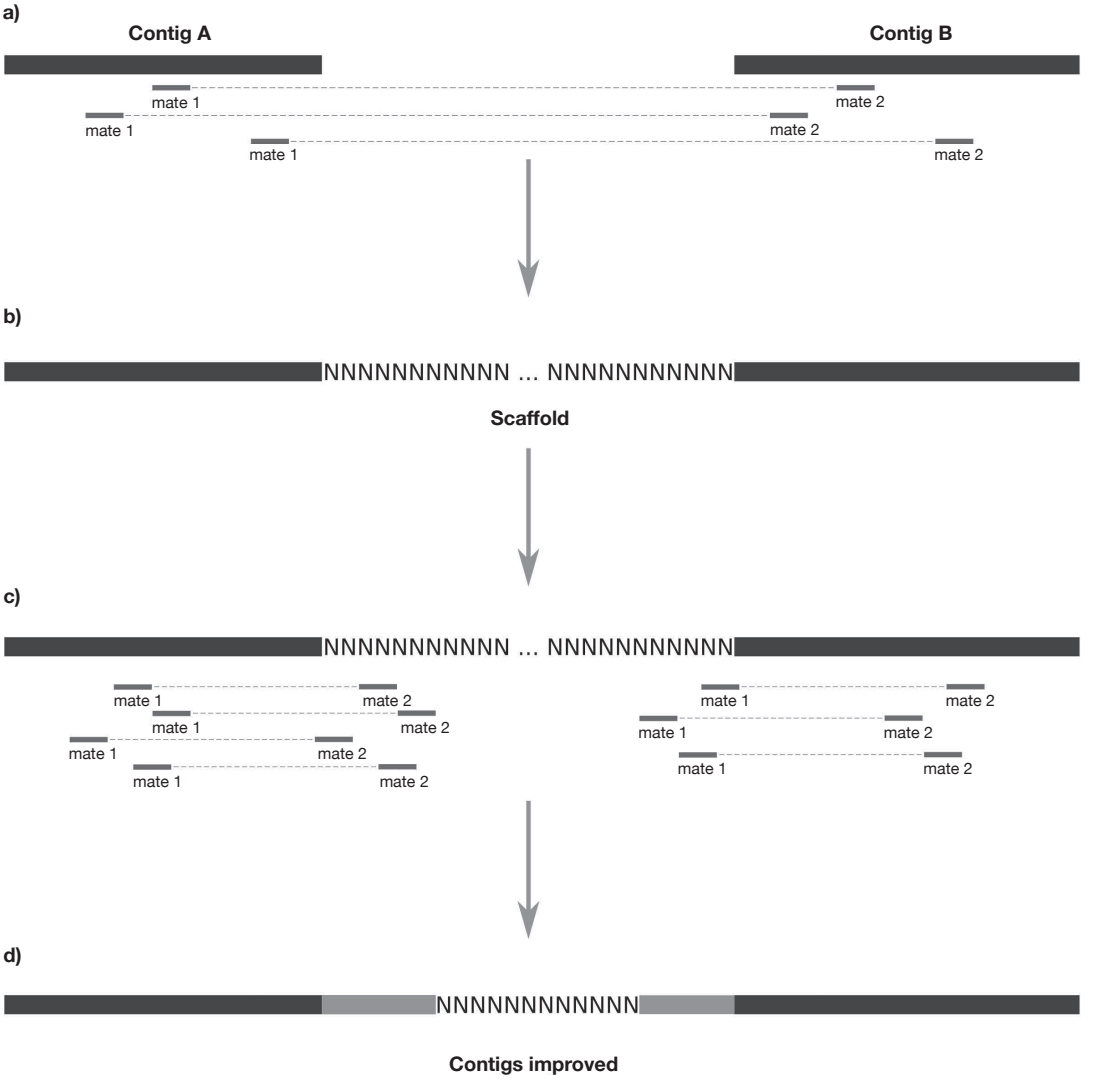

Figure 3.4 Scaffolding and contig extension. (a) Mate pair reads (grey) connect contigs A and B (black) with their long insert sizes (black) to (b) produce a consensus scaffold where gaps are denoted by unknown bases (N). (c) Paired-end reads (grey) are mapped locally to these gaps extend the contigs and shorten or close the gaps to reduce the gap (d) (gap closed interval shown in grey).

tree during the layout phase to produce initial sets of contigs of at least one read length. Then during the consensus step, the contigs are iteratively aligned to minimize redundancy.

De Bruijn graph methods align the sub-fragments (*k*-mers) rather than whole reads and are more effective than OLC for shorter reads (Zerbino and Birney, 2008). However, this non-exhaustive search depends on the user-defined *k* such that mapping accuracy is generally improved by increasing *k* but the total read coverage may be lower, reducing nucleotide-level accuracy. Consequently, the optimal *k*-mer depends on the uniqueness of the genome: this can be assessed by determining what fraction of the genome is unique for different *k* values. Multiple assemblies can also be combined from the same set of reads but for different values of *k*. Assembly methods using multiple *k* values are superior compared to those using only a single *k* (SOAPdenovo-MK, trans-ABySS and Oases-MK versus SOAPdenovo, ABySS, Oases and Trinity) (Zhao *et al.*, 2011).

Determining the unique sequence path in a set of reads using string graph algorithms represents a third assembly approach beyond OLC and de Bruijn graphs and may be faster (Myers, 2005).

Further stages may improve overall output: trimming the ends of reads can improve results – this applies particularly to Illumina reads where the polymerase-DNA synchrony declines with read length. As a result, sequence quality declines from 5′ to 3′ and so excluding low-quality portions may improve the assembly. In addition, both unprocessed reads and draft contigs should be aligned against contaminant and vector sequence databases, and orphan contigs not assigned to scaffolds can be compared against related species genome sequences for classification (Altschul *et al.*, 1990). Optical maps provide sequence data digested by a restriction enzyme chosen in accordance with the genome nucleotide composition (Valouev *et al.*, 2006). Optical maps can remove mis-assembled contigs because they are generated from individual ordered chromosomes. Processing optical map sequences can be integrated with short-read assembly (Lin *et al.*, 2012). For microbiome samples, iterative approaches exists for the genome assembly of multiple species (Sharon *et al.*, 2012)

The assembled contigs must be aligned, ordered and oriented with respect to one another: this is called contiguation. This transforms the consensus sequences into trans-contig genome units called scaffolds – these should approximate chromosomes for high-quality assemblies. Scaffolds are assigned to karyotyped chromosomes: genome size and thus mean read depth can be inferred from Feulgen image analysis densitometry (Hardie *et al.*, 2002). Assembly size is can be summarized as the minimum contig length in the set of contigs covering at least 50% of the genome (N50).

Genome assembly and improvement tools

Assembly accuracy varies widely and is only loosely linked to contiguity (Salzberg *et al.*, 2012). Many genome assembly and analysis tools have been designed for long mammalian genomes with low mean read coverage (5- to 15-fold). For short microbial genomes computational efficiency is not a significant limitation – similarly, coverage is generally saturated. Consequently, understanding the nuances of these tools can improve power by optimizing the assembly. A comprehensive list of assembly tools based on de Bruijn graph, OLC and string graph algorithms and their comparative performance has been collated (Fonseca *et al.*, 2012): see also http://assemblathon.org (Earl *et al.*, 2011) and www.jurgott.org/linkage/ListSoftware.pdf (Miller *et al.*, 2010). Newbler (http://contig.wordpress.com) operates on long reads from the Roche 454 platform and can split DNA reads between contigs (Table 3.1). Novoalign (www.novocraft.com) uses the Needleman–Wunsch OLC algorithm with gaps (<7 bases) to map long and paired-end reads with high levels of mismatches accounting for base quality parameters. For genome assemblies of *S. aureus* (2 Mb in length) and *Rhodobacter sphaeroides* (4 Mb), ALLPATHS-LG (Gnerre *et al.*, 2011) tended to have a higher N50, fewer errors and a lower total number of contigs than ABySS (Simpson *et al.*, 2009), String Graph Assembler (SGA) (Simpson and Durbin, 2012), SOAPdenovo (http://soap.genomics.org.cn) and Velvet (Salzberg *et al.*, 2012). ABySS has outperformed Velvet and SOAPdenovo for a small genome sequenced at high depth with paired-end reads (Lin *et al.*, 2011).

Several additional steps can be implemented to enhance draft assemblies. The first is contiguation: for large insert size assemblies (1+ kb), scaffold structure can be optimized by examining the assembly output for multiple libraries (Hunt *et al.*, in press; www.sanger.ac.uk/resources/software/reapr/) or data sources (Darling *et al.*, 2011). Likelihood-based assembly methods can evaluate assembly quality based on base error rates, the insert size distribution, and on the coverage uniformity (such as CGAL, http://bio.math.berkeley.edu/cgal/; Rahman and Pachter, 2013). Additionally, partially guided algorithms can optimize contiguation using read-pair information, such as ABACAS (http://abacas.sourceforge.net/; Assefa *et al.*, 2009) and SSPACE (Boetzer *et al.*, 2011). This has been extended to automate complete genome assembly using unguided assemblies without any laborious parameter optimization: A5 (Andrew And Aaron's Awesome

Table 3.1 Popular genome assembly tools

Algorithm type	Tool name	Reference
de Bruijn graph	ABySS	Simpson *et al*., 2009
	ALLPATHS-LG	Gnerre *et al*., 2011
	DNAbaser	www.dnabaser.com
	Euler-SR	Chaisson and Pevzner, 2008
	SOAPdenovo	http://soap.genomics.org.cn
	Velvet	Zerbino and Birney, 2008
OLC	Arachnae2	Jaffe *et al*., 2003
	Celera	Myers *et al*., 2000
	Maq	Li *et al*., 2008a
	Mira3	Chevreux *et al*., 2004
	Newbler	http://contig.wordpress.com
	Novoalign	www.novocraft.com
	PCAP	Huang *et al*., 2003
	Telescoper	Bresler *et al*., 2012
String graph	SGA	Simpson and Durbin, 2012

Tools are listed according to the type of algorithm used to compare reads: de Bruijn graph, OLC (overlap-layout-consensus) or string graph. A more comprehensive list of assembly tools (Earl *et al*., 2011) and their performance is available (Fonseca *et al*., 2012).

Assembly) has produced assemblies close in quality to SOAPdenovo and can assemble a typical bacterial genome on a generic computer within a day without supervision (Tritt *et al*., 2012).

The second improvement step is the local assembly of reads at contig edges to iteratively extend and join contigs: over-extension should be avoided. This can be carried out with IMAGE (Tsai *et al*., 2010) and GapFiller (Nadalin *et al*., 2012). The third component is to improve nucleotide-sequence level accuracy after initial draft assembly development either by iterative re-mapping of reads (ICORN, http://sourceforge.net/projects/icorn/; Otto *et al*., 2010), or by computing read quality values (e.g. Quake, www.cbcb.umd.edu/software/quake; Kelley *et al*., 2010). The fourth aspect is the automated annotation transfer between related species with optional manual improvement, which accelerates functional analysis of novel genome assemblies: such tools include RATT (http://ratt.sourceforge.net/; (Otto *et al*., 2011)), Glimmer (Aggarwal and Ramaswamy, 2002), xBASE (Chaudhuri and Pallen, 2006) and DIYA (Stewart *et al*., 2009). Finally, advances in genome visualization can render chromosome structure, composition, genes and function for manual checking and interpretation of assembly information with tools like GenomeVX (http://wolfe.gen.tcd.ie/GenomeVx/) and DNAPlotter (www.sanger.ac.uk/resources/software/dnaplotter/; Carver *et al*., 2009). These improvement tools are complementary and some have been packaged as a unit for nucleotide correction, gap-closing, contiguation, assessment and annotation mapping as a semi-ordered but iterative process (e.g. the Post Assembly Genome Improvement Toolkit; www.sanger.ac.uk/resources/software/pagit/; Swain *et al*., 2012): for example, this has been implemented on *Mycobacterium* (Ho *et al*., 2012).

Mapping and filtering DNA reads

A high-quality draft genome sequence represents a reference point against which the sequence data of other strains can be compared. DNA read bases are aligned against the genome sequence to compute the probability of a match and thus their best mapping location. This is a nucleotide-level

comparison, distinct from mapping that relates the reference and sample regions covered by the reads (Li, 2011). Longer reads are not required if the reference and comparison strain are closely related, so short reads are sufficient to discover most chromosomal, SNP and structural mutations (Fig. 3.2).

Mapping DNA reads to a reference genome

To increase computational efficiency, read-mapping tools first create an index of the reference genome sequence, colloquially termed a hash index. This is a set of short subsequences whose size (k) is defined (generally < 21) and sampled across the genome using the seed-and-extend algorithm (Altschul *et al.*, 1990). Read-mapping tools identify matches between the hash index sequences and the reads – the accuracy of mapping can be improved by increasing the hash index k and its sampling density at the cost of more memory usage. For short genomes, computational speed is not an issue, but for short read libraries (< 50 bases) a higher hash sample density is advised. These tools map the sequence reads from a fastq file to an indexed reference hash of the genome fasta file to produce SAM (sequence alignment/map) and BAM (binary alignment/map) format output files. These tools generally expect that a known the distribution of read insert sizes: this can be determined with tools like Picard (http://picard.sourceforge.net/).

The choice of read mapping tool depends on the genome's architecture (Fonseca *et al.*, 2012). A subset of read mapping tools adopt the more efficient Burrows-Wheeler transform of the hash index for alignment (Graf *et al.*, 2007) to increase in processing speed, these include SOAP3 (Li *et al.*, 2008b; Liu *et al.*, 2012), BWA (Li and Durbin, 2009) and Bowtie2 (Langmead *et al.*, 2009b; Langmead and Salzberg, 2012). BWA was initially designed for short (< 200 bp) reads but was extended for longer (but not paired-end) ones (BWA-SW) (Li and Durbin, 2010). BWA includes indels in the genome-read alignment, and computes mapping quality scores (the chances of the read mapping uniquely) across the genome, and so is computationally slower than other tools. BWA-SW can tolerate higher rates of mismatch as the read length increases, making it useful for error-prone reads, detecting structural variants, and also improving assemblies *de novo*. Bowtie2 can map 50–1000 base reads including gaps and can be used to perform local mapping (http://bowtie-bio.sourceforge.net/bowtie2). Bowtie2 has advantages compared to BWA for the incorporation of base quality values into mapping accuracy and can require less 3′ DNA read trimming. SOAP3 maps across gaps, and for high-identity data (0–3 mismatches per read) performs marginally better than BWA and Bowtie in addition to being more computationally rapid (Li *et al.*, 2008b; Liu *et al.*, 2012).

Smalt uses a banded Smith–Waterman algorithm to locally align reads, and has a three to five-fold lower error rate and maps a greater total number of reads compared to BWA for 100 bp reads with 0.5–1.0% mismatches (www.sanger.ac.uk/resources/software/smalt; Ponstingl and Ning, 2010). CLC Assembly Cell may be more computationally efficient than BWA, Bowtie or Smalt, but this may not reflect contig quality (www.clcbio.com/wp-content/uploads/2012/10/whitepaper-on-CLC-read-mapper.pdf). For mapping reads from a divergent species (> 1% mismatches) to a reference genome, this may not only assist assembly contiguation (e.g. Stampy, www.well.ox.ac.uk/project-stampy; Lunter and Goodson, 2011) but also SV discovery (e.g. CUSHAW2, http://cushaw2.sourceforge.net; Liu and Schmidt, 2012).

Filtering out platform, amplification and coverage sequencing errors

The stochastic nature of sequence amplification, alignment and mutation detection means that read output and mapping results vary (Malhis and Jones, 2010). Significant sources of error for microbial mutation discovery are the DNA amplification steps during library preparation and cluster generation, the sequencing machine version used and the variance in coverage. These can be reduced by examining sequence quality metrics produced by read mapping tools: these screens compute for each site the number and quality of reference and variant bases at the site, proximity of other mutations and the number of erroneous reads. Metrics for examining individual samples

are outlined below (those exploiting population-wide information are detailed later):

1 The proximity of SNPs to SVs: SNPs and indels are more prevalent at larger SVs, which reduces power to accurately infer any mutation type.
2 The minimum and maximum coverage (4–1000 reads) (Li *et al.*, 2008a).
3 The insert size range where reads have pairs.
4 The base quality – based on the log-transformed fluorescent wavelength pattern unique to each nucleotide given the expected distribution (Brockman *et al.*, 2008). This can be extended to examining base quality variation across the read.
5 The mutation (SNP) quality score: this is effectively an analogue of the base quality score – the probability the genotype is the major non-reference allele (Ning *et al.*, 2001).
6 Mapping quality: the probability of a significantly unique alignment of a read with the candidate mutation to a single locus derived from the observed base quality across the set of mapped reads. A substantial minority of reads at a site may have mapping qualities significantly lower than the other reads (DePristo *et al.*, 2011). The difference in mapping quality for the reference and variant alleles may also differ.
7 An excess of errors may be present at contig edges and chromosome ends.
8 The inclusion of contaminant DNA as an additional set of contigs during read-mapping can exclude foreign reads more similar to the contaminant sequence.
9 Masking repetitive, homopolymeric and low-complexity sequence by inferring the local distribution of uniqueness (e.g. with Tantan; Frith, 2011).
10 Sites with extreme coverage assuming a quasi-normal distribution of chromosome-wide coverage, this can also include a depth-adjusted quality score (DePristo *et al.*, 2011).
11 PCR duplicate reads can be identified from their quality uniformity.
12 Allele (strand) bias reflects errors induced during prior PCR amplification of the reads and can be identified where the errors are present on one (forward or reverse) strand only: true mutations will be present on multiple reads for both strands (Schmitt *et al.*, 2012). Tackling this problem requires higher coverage and this should be considered when inferring heterozygous alleles.
13 Adjustments to library preparation and cluster generation chemistry may yield errors specific to individual sequencing runs present in all lanes but absent in other runs.
14 Sequence artefacts unique to the sequencing machine used – caution is advised even for analysis including different versions of the same platform type.
15 The mean read position of the variant bases.
16 The variance in read position of variant and reference alleles: a low variance may reflect local sequence artefacts.
17 The frequency of mismatches on reads, particularly if there is a difference between those with the variant base compared to ones carrying reference bases.
18 The chromosome or whole-genome copy number may vary between samples (Garrison and G, 2012).
19 The local GC content (Quail *et al.*, 2012).

Other genotyping approaches such as the capillary sequencing of PCR amplicons and allelic-specific SNP genotyping are used for validation of new SNPs discovered during genome-sequencing (Manske *et al.*, 2012). However, this neglects the differences between these systems in SNP detection: as a result, the capacity for genome sequence data to confirm known variants provides a more quantifiable measure of precision. Verification can be completed by the dose-dependent amplification of specific loci by quantitative PCR for small SVs; by fluorescent *in situ* hybridization of probes to targets for large SVs; or by comparative genomic hybridization using arrays for medium to large SVs.

Variant discovery using mapped reads

A significant asset for identifying mutations is the scope to optimize this process by calling variants with multiple tools. Different tools have

differing sensitivities and specificities for detecting mutations, and so accuracy can be improved by examining the properties of shared valid mutations. However, this may become overly conservative and omit true variants if sensitivity is limited. More powerful SNP callers infer likely polymorphisms based on the posterior probability of a non-reference genotype given the observed read data (Nielsen *et al.*, 2011). Given the observed genotype (O) for each read i at a site with n reads, the likelihood of a putative genotype (G) can be expressed as

$$\prod_{i=1}^{n} P(O_i|G)$$

This assumes each read is independent, conditional on O_i. For instance, SOAPsnp computes the observed genotypes depending on base quality, mismatches and errors (Liu *et al.*, 2012). Certain tools account for observed genotype dependencies, prior probabilities of variants, and call SNPs across a population rather individual samples. A further consideration for microbial genomes is the assumption of diploidy implicit in some (e.g. Samtools mpileup; Li *et al.*, 2009b) but not all tools (Garrison and Marth, 2012). This likelihood-based genotype inference can be used to examine the site frequency spectrum in diploid organisms (ANGSD, http://popgen.dk/software/angsd.html; Nielsen *et al.*, 2012)

There are four established and complementary methods of detecting SVs (Table 3.2): depth of coverage (DOC), paired-end mapping (PEM) using paired reads, split read mapping (SR), and assembly based (AS) (Alkan *et al.*, 2011). Assuming coverage saturation to minimize bias (Xie and Tammi, 2009), DOC assumes a normal or Poisson distribution of read depth values to detect large genomic rearrangements (> 1 kb). PEM detects smaller events like SNPs or small indels (< 10 bp), and reflects the interpretation of gaps by the mapping algorithm (Medvedev *et al.*, 2009). SVs of 10–50 bp are challenging to discover because their length is the same as the insert size variance between paired-end reads.

SR and AS are effective for medium-sized SVs (> 50 bp) and rely on the pattern of read pairs (mates or paired-end) in terms of their depth, orientation, insert sizes and uniqueness (Ameur *et al.*, 2010). SR mapping is a significant advantage in tackling complex SVs, though short split reads may cause spurious mapping (Medvedev *et al.*, 2011). If one mate of a pair can be uniquely mapped to the reference, then the other unmapped mate can be efficiently mapped using the insert size (Ye *et al.*, 2009). A single read may map to two sections of the genome with a long gap between the 5′ and 3′ parts of the read, which indicates that these sections are next to each other in the sample genome but not the reference. If the distance between the reads is significantly smaller than the expected insert size, then a segment may be inserted in the sample (Fig. 3.5). Correspondingly, a deletion in the sample would have pairs mapping further away than expected (Xi *et al.*, 2010). Thus, insertions occupy the low end of the insert size distribution and deletions the upper end. If the insertion is bigger than the insert size of the reads, the insertion signature may not be detected (Medvedev and Brudno, 2009). For such large insertions the read pairs or mates do not map to any locus, because they are contained in the inserted region in the sample that is absent in the reference. These can be confirmed by locally re-assembling the insert element (Rausch *et al.*, 2009). Translocations can be distinguished where the read pairs map to different chromosomes. Inversions can be

Table 3.2 Structural variant discovery methods and detection power

Type	PEM	DOC	SR	AS
Deletion	Y	Y	Y	Y
Tandem duplication	Y	Y	Y	Y
Inversion	Y	–	Y	Y
Translocation	Y	–	Y	Y
Small insertion	Y	–	–	Y
Large insertion	–	–	–	Y

Established methods of detecting SVs: depth of coverage (DOC), paired-end mapping (PEM), split read (SR), and assembly based (AS) (Alkan *et al.*, 2011). Assuming coverage saturation to minimize bias (Xie and Tammi, 2009), depth assumes a normal or Poisson distribution. PEM interprets gaps with a mapping algorithm (Medvedev *et al.*, 2009). SR and AS rely on the pattern of read pairs (mates or paired-end) in terms of their depth, orientation, insert sizes and uniqueness (Ameur *et al.*, 2010).

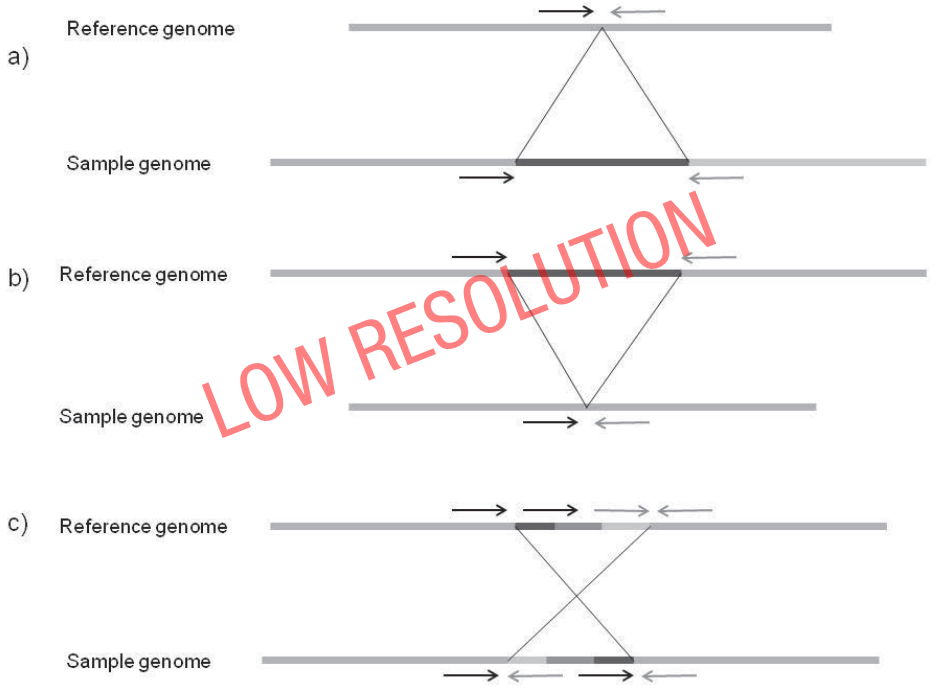

Figure 3.5 Detecting indels and inversions by mapping paired end reads. The black and grey arrows represent DNA paired-end reads from a sample genome mapped to a reference genome following the trajectory of the lines connecting the genomes. The distance between the sample read pairs reflects their expected insert size. (A) Insertions are denoted where the distance between the reads is less than the insert size (Medvedev and Brudno, 2009). (B) Deletions are denoted where the distance between the reads is greater than the insert (Xi *et al.*, 2010). (C) Inversions can be detected from changed ordering between read pairs, even though the insert size may be the same (Korbel *et al.*, 2007).

detected where the read orientation but not the insert size has changed (Korbel *et al.*, 2007). For tandem duplications, the orientation is reversed but some of these mis-oriented reads will overlap at the duplication point.

Different types of algorithms often specialize in finding distinct sizes and types of SVs, and so several different types of algorithms can be integrated: for example, SVMerge refines breakpoints using local *de novo* assembly (Wong *et al.*, 2010). Complex SVs can also be reconstructed manually with the aid of data visualization software where they cannot be interpreted properly (Reese *et al.*, 2010). Erroneous chimaeric reads mapping to homologous disparate loci, or accidentally joined during cluster generation, may produce false SVs. As a result, the quality of the reference genome assembly has a major impact on SV detection power.

Indels of less than 10 bp can be detected using the same alignment approach used for SNPs and a wide range of callers can call indels and SVs (Table 3.3), though caution should be advised for interpreting indels at homopolymers. Some tools perform local assembly on regions with unmapped reads to resolve SVs by exhaustively exploring the de Bruijn graph network (SOAPindel and SOAPsv; Li *et al.*, 2009b). While SOAPindel performs better at long indel detection (> 10 bp) than Dindel (Albers *et al.*, 2011), Pindel (Ye *et al.*, 2009) and the Genome Analysis Tool Kit (GATK; DePristo *et al.*, 2011), it retains a high false positive rate of 10% for indels longer than five bases (Liu *et al.*, 2012). Other callers find SVs in shotgun sequences (CNV-seq, http://tiger.dbs.nus.edu.sg/cnv-seq/; Xie and Tammi, 2009) or target medium size SVs (MoDIL; Lee *et al.*, 2009). For more complex SVs involving two large gaps – like tandem duplications or inversions – separately mapping the distal 5′ and 3′

Table 3.3 Popular software available for detecting structural variants

Name	Indel	Inversions	Reference
SOAPindel	No	No	Li *et al*. (2009b)
MAQ	<10bp	No	Li *et al*. (2008a)
MoDiL	10–50bp	No	Lee *et al*. (2009)
Dindel	<50bp	No	Albers *et al*. (2011)
SOAPsv			Liu *et al*. (2012)
Pindel			Ye *et al*. (2009)
GATK	Yes	No	DePristo *et al*. (2011)
Samtools			Li *et al*. (2009a)
CNVseq			Xie and Tammi (2009)
CNV-seq	No	Yes	Korbel *et al*. (2007), Xie and Tammi (2009)
PEMer	Yes	Yes	Korbel *et al*. (2007)
BreakDancer			Chen *et al*. (2009)
VariationHunter			Hormozdiari *et al*. (2010)
Cnvnator			Abyzov *et al*. (2011), Mills *et al*. (2011)

A wide range of callers can all be used to call SVs that vary in their power to detect different SV classes. All listed tools compute PE (paired-end) but not SE (single-end) reads except MAQ, which does both, and CNVseq, which only does SE reads.

read-pair ends can improve alignment (Alignment with Gap Excision; Abyzov and Gerstein, 2011). Notably, the substantial effects of the reference genome index format and read-mapping approach on variant calling has led to the development of integrated pipelines incorporating these key steps: for example, Crossbow combines Bowtie for read-mapping with SOAPsnp for genotype inference (Langmead *et al*., 2009a; Gurtowski *et al*., 2012).

Detecting population-wide variation from mapped DNA reads

Genotype inference needs to mitigate biases from heterogeneous sequencing technologies and coverage levels to remove low-quality false positive SNPs (Gronau *et al*., 2011). These extend from the parameters outlined above for individual samples that can be computed across the population: base, SNP and mapping qualities as well as read coverage distributions such as the fraction of samples with sufficient depth (Manske *et al*., 2012). Population-wide variant calling can be improved using a number of approaches in addition to estimating the individual read base error rate. The first is likelihood-based, which estimates the prior probability of a variant at a site based on the population-wide frequency of the variant allele (Nielsen *et al*., 2011). This can be improved by testing genotype confidence using a likelihood ratio test, by computing site-specific uncertainty due to other non-reference alleles, and testing for Hardy–Weinberg equilibrium (Kim *et al*., 2011). This will reduce the prior probability of observing a homozygous SNP for rare polymorphisms, and may ignore the clonal microbes (Tibayrenc and Ayala, 2012). Likelihood-based approaches can be applied to SVs as well as SNPs (Genome STRiP, www.broadinstitute.org/software/genomestrip/; Handsaker *et al*., 2011) and without a reference dataset (FreeBayes, http://bioinformatics.bc.edu/marthlab/FreeBayes; Garrison and Marth, 2012).

Known SNPs can be used to infer the prior probability of a variant (McKenna *et al*., 2010) and as a training set for recalibrating the expected quality properties of new SNPs (DePristo *et al*., 2011). For example, the GATK (www.broadinstitute.org/gatk) recalibration protocol incorporates read position, base quality, cluster cycle, population-level frequency and parental

information by sampling a binomial distribution to infer the relative chances of the number of reads with non-reference allele. GATK also uses the transition–transversion ratio (Ti/Tv) to calibrate substitution patterns: inter-species comparisons suggest a Ti/Tv of ~2.0–2.1 for genomes (Ebersberger *et al.*, 2002) and 3.0–3.3 for coding SNPs in humans (Freudenberg-Hua *et al.*, 2003). However, microbial substitution patterns may differ significantly (Tennessen *et al.*, 2012). Ti/Tv is important for evolutionary analysis of substitution rates (Yoder and Yang, 2000): a neutral Ti/Tv of 0.5 (n) should be observed because there are eight possible options for transversions and four for transitions (DePristo *et al.*, 2011). Transversions are rarer than transitions because they cause a bigger change in the nucleotide molecular shape, and so DNA repair mechanisms cannot compensate as easily (Kristina Strandberg and Salter, 2004), but Ti/Tv may be reduced by sequencing errors, alignment artefacts and data processing variability. Thus the fraction of false positive variants (FDR) can be estimated fro

$$FDR = \frac{o-n}{e-n}$$

where o is the observed Ti/Tv and e the expectation (DePristo *et al.*, 2011). Recalibration reflects the accuracy of the known SNPs so structurally variable regions may bias novel SNP detection power without an extensive reference mutation database: the lack of variation naturally present in certain microbes may also limit inference power. Additionally, the requirement of a draft reference template may exclude non-model microorganisms with lower quality assemblies.

The second approach is to infer population-wide parameters: this most important of these is to infer allele frequencies using read depth values across all samples (Jiang *et al.*, 2009; Lynch, 2009; Futschik and Schlotterer, 2010). By comparing the allele frequencies across all reads relative to the allele frequencies adjusted for each genotype call, sites with significant departures between the mean genotype and population allele frequency can be discovered. This switch of relating population frequencies straight from read information removes biases associated with SNP calling in individual samples (Kim *et al.*, 2011). Singleton SNPs are more likely to be errors than abundant ones, and so comparing their frequency during screening can be informative. This can be extended to patterns of homozygous and heterozygous SNPs. Computing the observed allele frequency spectrum in relation to the prior and posterior genotype probabilities can inform on the power of the approach to detect new SNPs and SVs (Nielsen *et al.*, 2011).

The third is based on the linkage patterns between mutations after initial population-wide genotype inference (e.g. Beagle; Browning and Browning, 2007). For variants on the same chromosome, their tendency to co-occur reflects their linkage disequilibrium (LD): the probability of co-inheritance. For mutations in high LD, the probability of observing one polymorphism affects the likelihood of observing the other (Abecasis *et al.*, 2012). Extending this to compute haplotype scores also allows imputation (phasing) of genotypes not examined using genome-sequencing (Browning and Browning, 2011). For low (< 15-fold) coverage data, likelihood-based methods estimate the population-wide frequencies more accurately and improve trait mapping (Kim *et al.*, 2011).

A fourth approach for SNP detection adopts the local reference-free assembly of reads to improve resolution at regions with extensive variation and avoid the inherent bias associated with SNPs unique to the reference (Iqbal *et al.*, 2012a). For monomorphic species, *de novo* local assembly can detect more variants (Iqbal *et al.*, 2012b). For instance, Cortex compares assemblies with low and high k to determine if the patterns of homozygous and heterozygous SNPs detected with Stampy (Lunter and Goodson, 2011) are significantly different (http://cortexassembler.sourceforge.net; Iqbal *et al.*, 2012b). If the less conservative (low k) data present a significant change in the ratio of heterozygous to homozygous SNPs, then this may reflect repetitive regions that can be resolved better with the lower k. A jump in coverage for the reference-free assembly can also indicate structural anomalies specific to the reference. Inevitably, these may still not resolve highly repetitive regions sufficiently, and manual interrogation of regions with high concentrations of heterozygous SNPs is advised.

Genome-wide data analysis and future trends

Microbial genome analysis platforms and databases

A variety of genome analysis platforms are available, some of which incorporate a community-based element for collaboration. For example, Galaxy (http://usegalaxy.org) allows users to upload, analyse, track and develop workflows for genome analysis: this covers monitoring sequencing, validating sequence output, read assembly, read mapping, variant filtering, genotype calling, population-wide recalibration, analysis tools and integration with other datasets (Goecks *et al.*, 2010). This has been extended with more advanced visual analysis tools like Trackser (Goecks *et al.*, 2012) and Peas v1.0 (Xu *et al.*, 2010). Visualization of reads, mutations, coverage and annotation for sets of samples as a single group is possible with Artemis and Bamview (Carver *et al.*, 2010, 2012). The Artemis Comparison Tool presents homology alignment scores between multiple species' genomes (Carver *et al.*, 2008), and viewing read mapping alignment to the reference sequence for individual samples is possible with Samtools (Li *et al.*, 2009a).

Although developed primarily for human genetics, genotype-phenotype genome–wide association tests can be applied, visualized and shared using Plink and associated tools (http://pngu.mgh.harvard.edu/~purcell/plink/; Purcell *et al.*, 2007). Similarly, examining LD, recombination and haplotypes is possible by co-opting techniques developed for vertebrate genomes like Haploview (www.broadinstitute.org/haploview; Barrett *et al.*, 2005). There are a wide variety of microbial genomics resources, including the DOE Joint Genome Institute (http://genome.jgi-psf.org), Microbes Online (www.microbesonline.org/), the Genome Encyclopaedia of Microbes (www.gem.re.kr), and the Comprehensive Microbial Resource (Peterson *et al.*, 2001). Integrated Microbial Genomes (http://img.jgi.doe.gov) is a community resource for comparative genomics and also analysis of genomes related to the Human Microbiome Project (www.hmpdacc-resources.org/img_hmp; Markowitz *et al.*, 2012a,b). GeneDB is an interactive functional database of pathogens and comparative genome datasets (www.genedb.org; Logan-Klumpler *et al.*, 2012).

Evolution, population structure and recombination

Genome sequencing provides a sufficient density of markers to provide adequate resolution of evolutionary and population-level variation (Twyford and Ennos, 2012). To construct phylogenies, an accurate substitution rate is required, which can be estimated by calculating the correlation of sample isolation time with the root-tip genetic distance. If the isolation date varies sufficiently, the dates of the ancestral nodes in the phylogeny can be estimated. Bayesian sampling of phylogenies from multiple loci may also provide a means to infer adaptation and historical population variation (Drummond and Rambaut, 2007; Drummond and Suchard, 2010; Drummond *et al.*, 2012), and can include recombination (Didelot *et al.*, 2010). Exploring genetic history in diploid genomes can be attempted with both multilocus and genomic data (Anderson *et al.*, 2005). Unlinked 1 kb genome segments can yield genealogical information by applying a coalescent-based scheme (Burgess and Yang, 2008) to estimate ancient population sizes, divergence times and gene flow from single samples representative of entire populations (Gronau *et al.*, 2011). Recent admixture can be inferred without estimating LD by exploiting recombination and haplotype switch points (Wegmann *et al.*, 2011) and can be extended to the origin, dispersal and spread of infectious microbes over space and time (Lemey *et al.*, 2010). An extensive variety of tools for scrutinizing bacterial population structure and recombination have been developed (http://pubmlst.org/software/). These extend from studying population structure (BAPS; Corander *et al.*, 2008) to examining copy number data (BASTA; Marttinen *et al.*, 2009a), estimating bacterial community composition (for 454 data with BEBaC; Cheng *et al.*, 2012), and phylogenetics (BANANAS; Siren *et al.*, 2011). For species with low recombination observed as the absence of LD decay with distance, specific tools applicable to MLST or genomic data are available (e.g. Clonalframe; Didelot and Falush, 2007). There are variety of tools that test for recombination, such as PHI (Bruen *et al.*, 2006),

BRAT (on individual genes; Marttinen *et al.*, 2009b), BRATNextGen (for genomes; Marttinen *et al.*, 2012) and others that integrate multiple methods (RDP3; Martin *et al.*, 2010).

More stringent filtering of genotype errors is required for the discovery of mutations functionally relevant to phenotypically distinct groups (Huang *et al.*, 2009). For binary comparisons of bacterial strains (such as drug-resistant and -sensitive sets), tools have been developed that only identify the SNPs differentiating the groups (e.g. VAAL, ftp://ftp.broadinstitute.org/pub/crd/VAAL/; Nusbaum *et al.*, 2009). Comparing divergent sequences can be completed through alignment, and so a variety of tools have been developed to compare genome consensus sequences beyond well-established local alignment tools like BLAST (Altschul *et al.*, 1990): the Robusta version of T-Coffee (Notredame, 2010), Mauve (Darling *et al.*, 2004) and Mercator (Dewey, 2007). Eukaryotic microbiologists can exploit powerful genotype inference tools developed for human data, which can examine clonal populations (e.g. MuTect; Banerji *et al.*, 2012), samples taken from the same patient over time-course (e.g. Vcf2diploid; Rozowsky *et al.*, 2011), or infer evolutionary history (Reuveni and Giuliani, 2012) to determine SNPs that improve assemblies (Catchen *et al.*, 2011) or compare pooled samples (Boitard *et al.*, 2012).

The effects of evolutionary selection pressures on allele frequencies can be estimated from longitudinal high-coverage genome samples (Tsibris *et al.*, 2009; Jabara *et al.*, 2011; Henn *et al.*, 2012). This has been applied extensively to HIV-1, whose high mutation rate enables the effects of selection on allele frequencies to be observed over short time periods. HIV-1 genomic data have provided information on the timing of host immune responses evasion by the virus, and also the duration for immune escape mutations to revert to the wild-type state in the absence of the immune response. The rate at which a mutant allele increases in frequency over time from an initial low level in the viral population can provide a means to quantify the strength of selection acting on the allele. This is achieved using a method based on diffusion models to estimate the selective coefficient from longitudinal allele frequency data (Bollback *et al.*, 2008). This can identify regions evolving under the action of positive selection by exploiting more accurate estimates of allele frequencies from sample sequence barcoding (Jabara *et al.*, 2011) and avoids biases in PCR amplification that can distort allele frequency estimates (Kanagawa, 2003).

Future trends

There are a number of emergent trends that may affect future microbial genomic analysis. The most significant is the increasing sequencing accuracy and speed through improved library preparation (Fitzsimons *et al.*, 2013) and sequencing chemistry (Chaisson and Pevzner, 2008). Using 1000 base reads delivers assemblies six times more continuous than 100 base reads (Kingsford *et al.*, 2010). Allele bias during PCR that produces errors present on one (forward or reverse) strand only can be circumvented using duplex sequencing (Schmitt *et al.*, 2012). This reduces sequencing error rates from 10^{-3} to 10^{-9} per base and permits investigation of DNA damage, DNA repair and mutation rates. New short run-time platforms like the Illumina Miseq and the Ion PGM mean short genomes (< 10 Mb) can be sequenced within a day, though with less efficiency (Didelot *et al.*, 2012). This increase in speed may be continued by nanopore-based machines, which may be able to deliver genomes in hours.

Single-molecule real-time (SMRT) sequencing can obviate library preparation completely, though with ten-fold less output (Travers *et al.*, 2010). It can sequence genome-wide information from as little as 1 ng within 8 h – this was tested on both viral and bacterial (MRSA) genomes (Coupland *et al.*, 2012). SMRT sequencing uses hairpin adaptors to sequence double-stranded DNA and can circumvent both PCR bias and produce long reads, but has a high error rate (Eid *et al.*, 2009).

Strobe sequencing is an application of the SMRT approach in which polymerases read circularised rather than linear DNA strands: it produces a portion of long reads despite an overall lower mean length (Lo *et al.*, 2011). A strobe read consists of multiple subreads from a single contiguous DNA fragment: if the strobe read has two subreads then it is a paired-end reads (with a more variable insert size; Raphael, 2012). Those with at

least three subreads are akin to paired-end reads with multiple partners, and the information from multiple subreads can assist with detecting SVs as well as performing *de novo* assembly of complex variants in highly repetitive regions or with high breakpoint densities (Ritz *et al.*, 2010).

Long SMRT reads can also present a template that can be corrected by alignment and assembly with short read data to produce a hybrid consensus sequence (PBcR in Celera Assembler11; Koren *et al.*, 2012). This has been applied to *V. cholerae* isolates from the Haitian cholera outbreak in October 2010 for which it resolved new virulence-related variants in repetitive regions, further supporting the Nepalese origin of this epidemic: this may be a general schematic for the automated finishing of bacterial genomes (Bashir *et al.*, 2012).

An alternative amplification approach called multiple annealing and looping-based amplification cycles (MALBAC) improves coverage uniformity such that it can be applied to individual cells lysed to produce picograms of 10–100 kb DNA fragments to detect SVs (Zong *et al.*, 2012). It works by a two-stage amplification protocol causing loops in the amplicons to prevent cross-hybridization and further amplification. This can produce contiguous phased maps of whole chromosomes, resolving linkage and recombination patterns between cell replication cycles (Lu *et al.*, 2012).

Contiguation and genome assembly can be improved by utilizing Strand-seq, a strategy that uses bromodeoxyuridine (BrdU) bases at the cell cycle stage of sister chromatid exchange during DNA replication of the temporally single-strand DNA fragments (Falconer *et al.*, 2012). Parental BrdU-free and progeny BrdU-exposed sequencing can predict local region orientation, order and alignment in comparison to a reference genome to refine the assembly. Finally, restriction site-associated DNA (RAD) sequencing may be useful where the total sequence read output may provide insufficient coverage for accurate *de novo* assembly for long or polyploid microbial genomes and where there is no relevant relate reference against which to map reads or contiguate the contigs.

A second major development is the combining of haplotype-based mutation detection in populations (Salem *et al.*, 2005; Howie *et al.*, 2012) with local de Bruijn graph assemblies, which could provide increased accuracy for genotype imputation and detection (Iqbal *et al.*, 2012a). This incorporates the variation naturally present in a diploid reference in terms of the de Bruijn graph topology in the population for each allele in a subsequence. Long repeats and variants unique to the reference are evident in the reconstructed contig alignments. An additional asset of this approach is that homozygosity can be examined as a linked block in a population and not individual mutations in different sequence libraries.

A third trend is the increased incorporation of gene expression and regulation data into genome analysis (Shendure and Aiden, 2012). *De novo* assembly of genomic data presents opportunities for amalgamating disparate results from different organisms by mapping experimental data from one genome to another (Mortazavi *et al.*, 2008). Regulatory information in chromatin immunoprecipitation DNA mapping peaks (ChIP-seq) can be locally re-assembled for alignment and motif discovery between genetically distinct specimens (Pinball, ftp://ftp.ebi.ac.uk/pub/databases/ensembl/avilella/pinball; Vilella *et al.* in prep.). Combining this with longer reads also allows clearer resolution of methylation and thus regulatory signatures in pathogenic samples (Fang *et al.*, 2012) and also between species (Murray *et al.*, 2012) and metagenomic patterns relevant to human disease (Jostins *et al.*, 2012).

Conclusions

Continued improvements in genome sequencing chemistry and computational tools enable the application of these methods to any microbe. This chapter explored the scope of microbial variation, and how assembly is the process of taking a large number of short DNA sequencing reads to develop a representation of the original chromosomes. Local as well as global genome-wide assembly is a powerful tool for inferring variation in species and ecotypes. Robust and accurate assemblies provide a platform against which SNPs and SVs can be inferred, though sensitive quality control measures are essential. Bridging variant recalibration with population genetic analysis and imputation methods (Browning and Browning,

2011) are protocols to infer population properties allowing for missing data (Ferretti *et al.*, 2012): this area has a swift rate of technological advances (Baker, 2011; Iqbal *et al.*, 2012a).

Genome sequencing is accelerating the treatment of infectious disease: initial genome-based approaches for studying infection progression and dynamics focused on short viral sequences (Lemey *et al.*, 2007), but have been more widely applied for general disease monitoring and surveillance (Walker and Beatson, 2012). These have tracked the emergence, transmission (McAdam *et al.*, 2012) and population structure (Everett *et al.*, 2012) of bacteria in clinical settings where clonal epidemics are not sufficiently resolved with traditional methods (Harris *et al.*, 2010). This increased resolution can be applied to document both prospective ongoing outbreaks and to retrospective historical evolution and spread (Monot *et al.*, 2009). In addition, genomic approaches can be more broadly applied to enhancing diagnostics and vaccine development (Seib *et al.*, 2012). New approaches to public health through the rapid real-time analysis of microbes using benchtop platforms will change how microbiology research is performed in hospitals as well as labs.

Acknowledgements

We thank the Irish Research Council, NUI Galway, and Science Foundation Ireland for funding. We also thank Hideo Imamura (Institute of Tropical Medicine, Antwerp, Belgium) for discussions.

References

Abderrazak, S.B., Oury, B., Lal, A.A., Bosseno, M.F., Force-Barge, P., Dujardin, J.P., Fandeur, T., Molez, J.F., Kjellberg, F., Ayala, F.J., *et al.* (1999). *Plasmodium falciparum*: population genetic analysis by multilocus enzyme electrophoresis and other molecular markers. Exp. Parasitol. *92*, 232–238.

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abyzov, A., and Gerstein, M. (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics *27*, 595–603.

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

Achtman, M. (2008). Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. *62*, 53–70.

Achtman, M. (2012). Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Philos. Trans. R. Soc. Lond. B Biol. Sci. *367*, 860–867.

Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A., and Carniel, E. (1999). *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. Proc. Natl. Acad. Sci. U.S.A. *96*, 14043–14048.

Aggarwal, G., and Ramaswamy, R. (2002). *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. J. Biosci. *27*, 7–14.

Al-Attar, S., Westra, E.R., van der Oost, J., and Brouns, S.J. (2011). Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol. Chem. *392*, 277–289.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Res. *21*, 961–973.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Ameur, A., Wetterbom, A., Feuk, L., and Gyllensten, U. (2010). Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol. *11*, R34.

Anderson, C.N., Ramakrishnan, U., Chan, Y.L., and Hadly, E.A. (2005). Serial SimCoal: a population genetics model for data from multiple populations and points in time. Bioinformatics *21*, 1733–1734.

Assefa, S., Keane, T.M., Otto, T.D., Newbold, C., and Berriman, M. (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics *25*, 1968–1969.

Baker, M. (2011). Sorting out sequencing data. Nat. Methods *8*, 799–803.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., *et al.* (2012). Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405–409.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263–265.

Bashir, A., Klammer, A.A., Robins, W.P., Chin, C.S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., *et al.* (2012). A hybrid approach for the automated finishing of bacterial genomes. Nat. Biotechnol. *30*, 701–707.

Beltran, P., Musser, J.M., Helmuth, R., Farmer, J.J., 3rd, Frerichs, W.M., Wachsmuth, I.K., Ferris, K., McWhorter, A.C., Wells, J.G., Cravioto, A., *et al.* (1988). Toward a population genetic analysis of Salmonella: genetic

diversity and relationships among strains of serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. Proc. Natl. Acad. Sci. U.S.A. *85*, 7753–7757.

Blount, Z.D., Barrick, J.E., Davidson, C.J., and Lenski, R.E. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature *489*, 513–518.

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

Boitard, S., Schlotterer, C., Nolte, V., Pandey, R.V., and Futschik, A. (2012). Detecting selective sweeps from pooled next-generation sequencing samples. Mol. Biol. Evol. *29*, 2177–2186.

Bollback, J.P., York, T.L., and Nielsen, R. (2008). Estimation of 2Nes from temporal allele frequency data. Genetics *179*, 497–502.

Bresler, M., Sheehan, S., Chan, A.H., and Song, Y.S. (2012). Telescoper: *de novo* assembly of highly repetitive regions. Bioinformatics *28*, i311–317.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., and Jaffe, D.B. (2008). Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. *18*, 763–770.

Browning, S.R., and Browning, B.L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. *81*, 1084–1097.

Browning, S.R., and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. *12*, 703–714.

Brueggemann, A.B., Griffiths, D.T., Meats, E., Peto, T., Crook, D.W., and Spratt, B.G. (2003). Clonal relationships between invasive and carriage *Streptococcus pneumoniae* and serotype- and clone-specific differences in invasive disease potential. J. Infect. Dis. *187*, 1424–1432.

Bruen, T.C., Philippe, H., and Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. Genetics *172*, 2665–2681.

Bulle, B., Millon, L., Bart, J.M., Gallego, M., Gambarelli, F., Portus, M., Schnur, L., Jaffe, C.L., Fernandez-Barredo, S., Alunda, J.M., *et al.* (2002). Practical approach for typing strains of *Leishmania infantum* by microsatellite analysis. J. Clin. Microbiol. *40*, 3391–3397.

Burge, C., Campbell, A.M., and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. U.S.A. *89*, 1358–1362.

Burgess, R., and Yang, Z. (2008). Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. *25*, 1979–1994.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.L., and Brussow, H. (2003). Phage as agents of lateral gene transfer. Curr. Opin. Microbiol. *6*, 417–424.

Carle, G.F., and Olson, M.V. (1984). Separation of chromosomal DNA molecules from yeast by orthogonal-field-alternation gel electrophoresis. Nucleic Acids Res. *12*, 5647–5664.

Carle, G.F., Frank, M., and Olson, M.V. (1986). Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science *232*, 65–68.

Carter, M.G., Hamatani, T., Sharov, A.A., Carmack, C.E., Qian, Y., Aiba, K., Ko, N.T., Dudekula, D.B., Brzoska, P.M., Hwang, S.S., *et al.* (2003). *In situ*-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling. Genome Res. *13*, 1011–1021.

Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J., and Rajandream, M.A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics *24*, 2672–2676.

Carver, T., Thomson, N., Bleasby, A., Berriman, M., and Parkhill, J. (2009). DNAPlotter: circular and linear interactive genome visualization. Bioinformatics *25*, 119–120.

Carver, T., Bohme, U., Otto, T.D., Parkhill, J., and Berriman, M. (2010). BamView: viewing mapped read alignment data in the context of the reference sequence. Bioinformatics *26*, 676–677.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J., and McQuillan, J.A. (2012). Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics *28*, 464–469.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping Loci *de novo* from short-read sequences. G3 (Bethesda) *1*, 171–182.

Chaisson, M.J., and Pevzner, P.A. (2008). Short read fragment assembly of bacterial genomes. Genome Res. *18*, 324–330.

Chaudhuri, R.R., and Pallen, M.J. (2006). xBASE, a collection of online databases for bacterial comparative genomics. Nucleic Acids Res. *34*, D335–337.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Cheng, L., Connor, T.R., Aanensen, D.M., Spratt, B.G., and Corander, J. (2011). Bayesian semi-supervised classification of bacterial samples using MLST databases. BMC Bioinformat. *12*, 302.

Cheng, L., Walker, A.W., and Corander, J. (2012). Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Res. *40*, 5240–5249.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E., Wetter, T., and Suhai, S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. *14*, 1147–1159.

Chiu, C.Y., Urisman, A., Greenhow, T.L., Rouskin, S., Yagi, S., Schnurr, D., Wright, C., Drew, W.L., Wang, D., Weintrub, P.S., *et al.* (2008). Utility of DNA microarrays for detection of viruses in acute respiratory tract infections in children. J. Pediatr. *153*, 76–83.

Coelho, A.C., Leprohon, P., and Ouellette, M. (2012). Generation of leishmania hybrids by whole genomic DNA transformation. PLoS Negl. Trop. Dis. *6*, e1817.

Corander, J., Marttinen, P., Siren, J., and Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformat. *9*, 539.

Coupland, P., Chandra, T., Quail, M., Reik, W., and Swerdlow, H. (2012). Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation. Biotechniques *53*, 365–372.

Darling, A.C., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. *14*, 1394–1403.

Darling, A.E., Tritt, A., Eisen, J.A., and Facciotti, M.T. (2011). Mauve assembly metrics. Bioinformatics *27*, 2756–2757.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dewey, C.N. (2007). Aligning multiple whole genomes with Mercator and MAVID. Methods Mol. Biol. *395*, 221–236.

Dicuonzo, G., Gherardi, G., Gertz, R.E., D'Ambrosio, F., Goglio, A., Lorino, G., Recchia, S., Pantosti, A., and Beall, B. (2002). Genotypes of invasive pneumococcal isolates recently recovered from Italian patients. J. Clin. Microbiol. *40*, 3660–3665.

Didelot, X., and Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. Genetics *175*, 1251–1266.

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

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E., and Crook, D.W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. *13*, 601–612.

Dolina, M., and Peduzzi, R. (1993). Population genetics of human, animal, and environmental Yersinia strains. Appl. Environ. Microbiol. *59*, 442–450.

Downing, T., Imamura, H., Decuypere, S., Clark, T.G., Coombs, G.H., Cotton, J.A., Hilley, J.D., de Doncker, S., Maes, I., Mottram, J.C., *et al.* (2011). Whole genome sequencing of multiple Leishmania donovani clinical isolates provides insights into population structure and mechanisms of drug resistance. Genome Res. *21*, 2143–2156.

Driscoll, J.R. (2009). Spoligotyping for molecular epidemiology of the Mycobacterium tuberculosis complex. Methods Mol. Biol. *551*, 117–128.

Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.

Drummond, A.J., and Suchard, M.A. (2010). Bayesian random local clocks, or one rate to rule them all. BMC Biol. *8*, 114.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H.O., Buffalo, V., Zerbino, D.R., Diekhans, M., *et al.* (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Res. *21*, 2224–2241.

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. *70*, 1490–1497.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. Science *323*, 133–138.

Eisenstein, M. (2012). The battle for sequencing supremacy. Nat. Biotechnol. *30*, 1023–1026.

Ellegren, H. (2000). Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. *16*, 551–558.

Enright, M.C., and Spratt, B.G. (1998). A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology *144(Pt. 11)*, 3049–3060.

Enright, M.C., and Spratt, B.G. (1999). Multilocus sequence typing. Trends Microbiol. *7*, 482–487.

Enright, M.C., Robinson, D.A., Randle, G., Feil, E.J., Grundmann, H., and Spratt, B.G. (2002). The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). Proc. Natl. Acad. Sci. U.S.A. *99*, 7687–7692.

Everett, D.B., Cornick, J., Denis, B., Chewapreecha, C., Croucher, N., Harris, S., Parkhill, J., Gordon, S., Carrol, E.D., French, N., *et al.* (2012). Genetic characterisation of Malawian pneumococci prior to the roll-out of the PCV13 vaccine using a high-throughput whole genome sequencing approach. PLoS One *7*, e44250.

Falconer, E., Hills, M., Naumann, U., Poon, S.S., Chavez, E.A., Sanders, A.D., Zhao, Y., Hirst, M., and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods *9*, 1107–1112.

Fang, G., Munera, D., Friedman, D.I., Mandlik, A., Chao, M.C., Banerjee, O., Feng, Z., Losic, B., Mahajan, M.C., Jabado, O.J., *et al.* (2012). Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. Nat. Biotechnol. *30*, 1232–1239.

Ferretti, L., Raineri, E., and Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. Genetics *191*, 1397–1401.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., *et al.* (1976).

Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature *260*, 500–507.

Fitzsimons, M.S., Novotny, M., Lo, C.C., Dichosa, A.E., Yee-Greenbaum, J.L., Snook, J.P., Gu, W., Chertkov, O., Davenport, K.W., McMurry, K., *et al.* (2013). Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. Genome Res. *23*, 878–888.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496–512.

Flicek, P., and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. Nat. Methods *6*, S6-S12.

Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. Bioinformatics *28*, 3169–3177.

Freudenberg-Hua, Y., Freudenberg, J., Kluck, N., Cichon, S., Propping, P., and Nothen, M.M. (2003). Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. Genome Res. *13*, 2271–2276.

Frith, M.C. (2011). A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. *39*, e23.

Frost, L.S., Leplae, R., Summers, A.O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. *3*, 722–732.

Futschik, A., and Schlotterer, C. (2010). The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics *186*, 207–218.

Gardiner, K., Laas, W., and Patterson, D. (1986). Fractionation of large mammalian DNA restriction fragments using vertical pulsed-field gradient gel electrophoresis. Somat. Cell Mol. Genet. *12*, 185–195.

Gardy, J.L., Johnston, J.C., Ho Sui, S.J., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., *et al.* (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N. Engl. J. Med. *364*, 730–739.

Garrison, E., and Marth G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv *1207*, 3907.

Gil-Lamaignere, C., Roilides, E., Hacker, J., and Muller, F.M. (2003). Molecular typing for fungi--a critical review of the possibilities and limitations of currently and future methods. Clin. Microbiol. Infect. *9*, 172–185.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. *11*, 759–769.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. U.S.A. *108*, 1513–1518.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. *11*, R86.

Goecks, J., Coraor, N., Team, T.G., Nekrutenko, A., and Taylor, J. (2012). NGS analyses by visualization with Trackster. Nat. Biotechnol. *30*, 1036–1039.

Goering, R.V. (2010). Pulsed field gel electrophoresis: a review of application and interpretation in the molecular epidemiology of infectious disease. Infect. Genet. Evol. *10*, 866–875.

Graf, S., Nielsen, F.G., Kurtz, S., Huynen, M.A., Birney, E., Stunnenberg, H., and Flicek, P. (2007). Optimized design and assessment of whole genome tiling arrays. Bioinformatics *23*, i195–204.

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993). Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method. Mol. Microbiol. *10*, 1057–1065.

Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. *43*, 1031–1034.

Gurtowski, J., Schatz, M.C., and Langmead, B. (2012). Genotyping in the cloud with Crossbow. Curr. Protoc. Bioinformatics *Chapter 15*, Unit 15.3.

Haase, J.K., Murphy, R.A., Choudhury, K.R., and Achtman, M. (2011). Revival of Seeliger's historical 'Special Listeria Culture Collection'. Environ. Microbiol. *13*, 3163–3171.

Hacker, J., Knapp, S., and Goebel, W. (1983). Spontaneous deletions and flanking regions of the chromosomally inherited hemolysin determinant of an *Escherichia coli* O6 strain. J. Bacteriol. *154*, 1145–1152.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput. Biol. *1*, e60.

Hamarsheh, O. (2011). Distribution of Leishmania major zymodemes in relation to populations of *Phlebotomus papatasi* sand flies. Parasit. Vectors *4*, 9.

Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. *43*, 269–276.

Hardie, D.C., Gregory, T.R., and Hebert, P.D. (2002). From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. J. Histochem. Cytochem. *50*, 735–749.

Harris, S.R., Feil, E.J., Holden, M.T., Quail, M.A., Nickerson, E.K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J.A., *et al.* (2010). Evolution of MRSA during hospital transmission and intercontinental spread. Science *327*, 469–474.

Harris, S.R., Cartwright, E.J., Torok, M.E., Holden, M.T., Brown, N.M., Ogilvy-Stuart, A.L., Ellington, M.J., Quail, M.A., Bentley, S.D., Parkhill, J., Peacock, S.J. (2012). Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. Lancet Infect. Dis. *13*, 130–136.

He, J., and Deem, M.W. (2010). Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). Phys. Rev. Lett. *105*, 128102.

Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin, A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., *et al.* (2012). Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. *8*, e1002529.

Hentschel, U., and Hacker, J. (2001). Pathogenicity islands: the tip of the iceberg. Microbes Infect. *3*, 545–548.

Hinnebusch, J., and Tilly, K. (1993). Linear plasmids and chromosomes in bacteria. Mol. Microbiol. *10*, 917–922.

Ho, Y.S., Adroub, S.A., Abadi, M., Al Alwan, B., Alkhateeb, R., Gao, G., Ragab, A., Ali, S., van Soolingen, D., Bitter, W., *et al.* (2012). Complete Genome Sequence of *Mycobacterium vaccae* Type Strain ATCC 25954. J. Bacteriol. *194*, 6339–6340.

Holt, K.E., Parkhill, J., Mazzoni, C.J., Roumagnac, P., Weill, F.X., Goodhead, I., Rance, R., Baker, S., Maskell, D.J., Wain, J., *et al.* (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. Nat. Genet. *40*, 987–993.

Hooper, S.D., and Berg, O.G. (2002). Detection of genes with atypical nucleotide sequence in microbial genomes. J. Mol. Evol. *54*, 365–375.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E.E., and Sahinalp, S.C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics *26*, i350–357.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. *44*, 955–959.

Huang, L., Wang, C., and Rosenberg, N.A. (2009). The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am. J. Hum. Genet. *85*, 692–698.

Huang, X., Wang, J., Aluru, S., Yang, S.P., and Hillier, L. (2003). PCAP: a whole-genome assembly program. Genome Res. *13*, 2164–2170.

Hubby, J.L., and Lewontin, R.C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. Genetics *54*, 577–594.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012a). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. Nat. Genet. *44*, 226–232.

Iqbal, Z., Turner, I., and McVean, G. (2012b). High-throughput microbial population genomics using the Cortex variation assembler. Bioinformatics, *29*, 275–276.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. J. Bacteriol. *169*, 5429–5433.

Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A., and Swanstrom, R. (2011). Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc. Natl. Acad. Sci. U.S.A. *108*, 20166–20171.

Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. *13*, 91–96.

Jansen, R., van Embden, J.D., Gaastra, W., and Schouls, L.M. (2002). Identification of a novel family of sequence repeats among prokaryotes. OMICS *6*, 23–33.

Jiang, R., Tavare, S., and Marjoram, P. (2009). Population genetic inference from resequencing data. Genetics *181*, 187–197.

Jolley, K.A., and Maiden, M.C. (2010). BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformat. *11*, 595.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., *et al.* (2012). Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W., and Crook, D.W. (2009). Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. *33*, 376–393.

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., *et al.* (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. J. Clin. Microbiol. *35*, 907–914.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (PCR). J. Biosci. Bioeng. *96*, 317–323.

Kelley, D.R., Schatz, M.C., and Salzberg, S.L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. *11*, R116.

Kim, S.Y., Lohmueller, K.E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., *et al.* (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformat. *12*, 231.

Kingsford, C., Schatz, M.C., and Pop, M. (2010). Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformat. *11*, 21.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science *318*, 420–426.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., *et al.* (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. *30*, 693–700.

Kremer, K., van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W., Martin, C., Palittapongarnpim, P.,

Plikaytis, B.B., Riley, L.W., Yakrus, M.A., *et al.* (1999). Comparison of methods based on different molecular epidemiological markers for typing of Mycobacterium tuberculosis complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. *37*, 2607–2618.

Kristina Strandberg, A.K., and Salter, L.A. (2004). A comparison of methods for estimating the transition:transversion ratio from DNA sequences. Mol. Phylogenet. Evol. *32*, 495–503.

Langille, M.G., Hsiao, W.W., and Brinkman, F.S. (2010). Detecting genomic islands using bioinformatics approaches. Nat. Rev. Microbiol. *8*, 373–382.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Schatz, M.C., Lin, J., Pop, M., and Salzberg, S.L. (2009a). Searching for SNPs with cloud computing. Genome Biol. *10*, R134.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009b). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

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

Lawrence, J.G., and Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. U.S.A. *95*, 9413–9417.

Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. (2009). MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat. Methods *6*, 473–474.

Lemey, P., Kosakovsky Pond, S.L., Drummond, A.J., Pybus, O.G., Shapiro, B., Barroso, H., Taveira, N., and Rambaut, A. (2007). Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput. Biol. *3*, e29.

Lemey, P., Rambaut, A., Welch, J.J., and Suchard, M.A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. *27*, 1877–1885.

Leprohon, P., Legare, D., Raymond, F., Madore, E., Hardiman, G., Corbeil, J., and Ouellette, M. (2009). Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*. Nucleic Acids Res. *37*, 1387–1399.

Lewontin, R.C., and Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics *54*, 595–609.

Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics *27*, 1157–1158.

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

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. *11*, 473–483.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. *18*, 1851–1858.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., and Kristiansen, K. (2009b). SNP detection for massively parallel whole-genome resequencing. Genome Res. *19*, 1124–1132.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713–714.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., *et al.* (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct. Genomics *11*, 25–37.

Lin, H.C., Goldstein, S., Mendelowitz, L., Zhou, S., Wetzel, J., Schwartz, D.C., and Pop, M. (2012). AGORA: Assembly Guided by Optical Restriction Alignment. BMC Bioinformat. *13*, 189.

Lin, Y., Li, J., Shen, H., Zhang, L., Papasian, C.J., and Deng, H.W. (2011). Comparative studies of *de novo* assembly tools for next-generation sequencing technologies. Bioinformatics *27*, 2031–2037.

Lindstedt, B.A. (2005). Multiple-locus variable number tandem repeats analysis for genetic fingerprinting of pathogenic bacteria. Electrophoresis *26*, 2567–2582.

Liu, C.M., Wong, T., Wu, E., Luo, R., Yiu, S.M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., *et al.* (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics *28*, 878–879.

Liu, Y., and Schmidt, B. (2012). Long read alignment based on maximal exact match seeds. Bioinformatics *28*, i318-i324.

Llewellyn, M.S., Miles, M.A., Carrasco, H.J., Lewis, M.D., Yeo, M., Vargas, J., Torrico, F., Diosque, P., Valente, V., Valente, S.A., *et al.* (2009). Genome-scale multilocus microsatellite typing of *Trypanosoma cruzi* discrete typing unit I reveals phylogeographic structure and specific genotypes linked to human infection. PLoS Pathog. *5*, e1000410.

Lo, C., Bashir, A., Bansal, V., and Bafna, V. (2011). Strobe sequence design for haplotype assembly. BMC Bioinformat. *12(Suppl. 1)*, S24.

Logan-Klumpler, F.J., De Silva, N., Boehme, U., Rogers, M.B., Velarde, G., McQuillan, J.A., Carver, T., Aslett, M., Olsen, C., Subramanian, S., *et al.* (2012). GeneDB--an annotation database for pathogens. Nucleic Acids Res. *40*, D98–108.

Lu, S., Zong, C., Fan, W., Yang, M., Li, J., Chapman, A.R., Zhu, P., Hu, X., Xu, L., Yan, L., *et al.* (2012). Probing meiotic recombination and aneuploidy of single sperm

cells by whole-genome sequencing. Science *338*, 1627–1630.

Lunter, G., and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. *21*, 936–939.

Lynch, M. (2009). Estimation of allele frequencies from high-coverage genome-sequencing projects. Genetics *182*, 295–301.

McAdam, P.R., Templeton, K.E., Edwards, G.F., Holden, M.T., Feil, E.J., Aanensen, D.M., Bargawi, H.J., Spratt, B.G., Bentley, S.D., Parkhill, J., *et al.* (2012). Molecular tracing of the emergence, adaptation, and transmission of hospital-associated methicillin-resistant *Staphylococcus aureus*. Proc. Natl. Acad. Sci. U.S.A. *109*, 9107–9112.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Maiden, M.C. (1998). Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. Clin. Infect. Dis. *27(Suppl. 1)*, S12–20.

Maiden, M.C. (2006). Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. *60*, 561–588.

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

Malhis, N., and Jones, S.J. (2010). High quality SNP calling using Illumina data at shallow coverage. Bioinformatics *26*, 1029–1035.

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., O'Brien, J., Djimde, A., Doumbo, O., Zongo, I., *et al.* (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature *487*, 375–379.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376–380.

Markowitz, V.M., Chen, I.M., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., *et al.* (2012a). IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. *40*, D123–129.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., *et al.* (2012b). IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res. *40*, D115–122.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefeuvre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics *26*, 2462–2463.

Marttinen, P., Myllykangas, S., and Corander, J. (2009a). Bayesian clustering and feature selection for cancer tissue samples. BMC Bioinformat. *10*, 90.

Marttinen, P., Tang, J., De Baets, B., Dawyndt, P., and Corander, J. (2009b). Bayesian clustering of fuzzy feature vectors using a quasi-likelihood approach. IEEE Trans. Pattern Anal. Mach. Intell. *31*, 74–85.

Marttinen, P., Hanage, W.P., Croucher, N.J., Connor, T.R., Harris, S.R., Bentley, S.D., and Corander, J. (2012). Detection of recombination events in bacterial genomes from large population samples. Nucleic Acids Res. *40*, e6.

Medvedev, P., and Brudno, M. (2009). Maximum likelihood genome assembly. J. Comput. Biol. *16*, 1101–1116.

Medvedev, P., Stanciu, M., and Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods *6*, S13–20.

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., and Brudno, M. (2010). Detecting copy number variation with mated short reads. Genome Res. *20*, 1613–1622.

Medvedev, P., Pham, S., Chaisson, M., Tesler, G., and Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. J. Comput. Biol. *18*, 1625–1634.

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

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., *et al.* (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G.M., Donoghue, H.D., Bouwman, A., *et al.* (2009). Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat. Genet. *41*, 1282–1289.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Murray, I.A., Clark, T.A., Morgan, R.D., Boitano, M., Anton, B.P., Luong, K., Fomenkov, A., Turner, S.W., Korlach, J., and Roberts, R.J. (2012). The methylomes of six bacteria. Nucleic Acids Res. *40*, 11450–11462.

Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. U.S.A. *84*, 166–169.

Mutreja, A., Kim, D.W., Thomson, N.R., Connor, T.R., Lee, J.H., Kariuki, S., Croucher, N.J., Choi, S.Y., Harris, S.R., Lebens, M., *et al.* (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. Nature *477*, 462–465.

Myers, E.W. (2005). The fragment assembly string graph. Bioinformatics *21(Suppl. 2)*, ii79–85.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., *et al.* (2000). A

whole-genome assembly of Drosophila. Science *287*, 2196–2204.

Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformat. *13(Suppl. 14)*, S8.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. *12*, 443–451.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., and Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One *7*, e37558.

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001). SSAHA: a fast search method for large DNA databases. Genome Res. *11*, 1725–1729.

Noller, A.C., McEllistrem, M.C., Stine, O.C., Morris, J.G., Jr., Boxrud, D.J., Dixon, B., and Harrison, L.H. (2003). Multilocus sequence typing reveals a lack of diversity among Escherichia coli O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. J. Clin. Microbiol. *41*, 675–679.

Notredame, C. (2010). Computing multiple sequence/structure alignments with the T-coffee package. Curr. Protoc. Bioinformatics *Chapter 3*, Unit 3.8, 1–25.

Nowrousian, M. (2010). Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot. Cell *9*, 1300–1310.

Nusbaum, C., Ohsumi, T.K., Gomez, J., Aquadro, J., Victor, T.C., Warren, R.M., Hung, D.T., Birren, B.W., Lander, E.S., and Jaffe, D.B. (2009). Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing. Nat. Methods *6*, 67–69.

O'Roak, B.J., Vives, L., Fu, W., Egertson, J.D., Stanaway, I.B., Phelps, I.G., Carvill, G., Kumar, A., Lee, C., Ankenman, K., *et al.* (2012). Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. Science *21*, 1619–1622.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304.

Osborn, A.M., and Boltner, D. (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid *48*, 202–212.

Otto, T.D., Sanders, M., Berriman, M., and Newbold, C. (2010). Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology. Bioinformatics *26*, 1704–1707.

Otto, T.D., Dillon, G.P., Degrave, W.S., and Berriman, M. (2011). RATT: Rapid Annotation Transfer Tool. Nucleic Acids Res. *39*, e57.

Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M., and Kyrpides, N.C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. *40*, D571–579.

Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. Cell *113*, 171–182.

Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., *et al.* (2012). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature *487*, 190–195.

Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K., and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Res. *29*, 123–125.

Ponstingl, H., and Ning, Z. (2010). SMALT – A new mapper for DNA sequencing reads. F1000 Posters *1*, 313.

Pourcel, C., Visca, P., Afshar, B., D'Arezzo, S., Vergnaud, G., and Fry, N.K. (2007). Identification of variable-number tandem-repeat (VNTR) sequences in Legionella pneumophila and development of an optimized multiple-locus VNTR analysis typing scheme. J. Clin. Microbiol. *45*, 1190–1199.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genom. *13*, 341.

Rahman, A., and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. Genome Biol. *14*, R8.

Raphael, B.J. (2012). Chapter 6: Structural variation and medical genomics. PLoS Comput. Biol. *8*, e1002821.

Rasko, D.A., and Mongodin, E.F. (2005). The first decade of microbial genomics: what have we learned and where are we going next? Genome Biol. *6*, 341.

Rausch, T., Koren, S., Denisov, G., Weese, D., Emde, A.K., Doring, A., and Reinert, K. (2009). A consistency-based consensus algorithm for de novo and reference-guided sequence assembly of short reads. Bioinformatics *25*, 1118–1124.

Reese, M.G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G.T., Stein, L., Flicek, P., Yandell, M., and Eilbeck, K. (2010). A standard variation file format for human genome sequences. Genome Biol. *11*, R88.

Reuveni, E., and Giuliani, A. (2012). A novel multi-scale modeling approach to infer whole genome divergence. Evol. Bioinform. Online *8*, 611–622.

Ritz, A., Bashir, A., and Raphael, B.J. (2010). Structural variation analysis with strobe reads. Bioinformatics *26*, 1291–1298.

Rocha, E.P., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. Trends Genet. *18*, 291–294.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J., *et al.* (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N. Engl. J. Med. *365*, 718–724.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., *et al.* (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol. *7*, 522.

Salem, R.M., Wessel, J., and Schork, N.J. (2005). A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum. Genom. *2*, 39–66.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T.J., Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. *94*, 441–448.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. *74*, 5463–5467.

Schatz, M.C., Delcher, A.L., Roberts, M., *et al.* (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. *22*, 557–567.

Schmidt, H., and Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. *17*, 14–56.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. U.S.A. *109*, 14508–14513.

Schouls, L.M., van der Ende, A., Damen, M., and van de Pol, I. (2006). Multiple-locus variable-number tandem repeat analysis of *Neisseria meningitidis* yields groupings similar to those obtained by multilocus sequence typing. J. Clin. Microbiol. *44*, 1509–1518.

Schwartz, D.C., and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell *37*, 67–75.

Seib, K.L., Zhao, X., and Rappuoli, R. (2012). Developing vaccines in the era of genomics: a decade of reverse vaccinology. Clin. Microbiol. Infect. *18(Suppl. 5)*, 109–116.

Selander, R.K., Caugant, D.A., Ochman, H., Musser, J.M., Gilmour, M.N., and Whittam, T.S. (1986). Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl. Environ. Microbiol. *51*, 873–884.

Service, R.F. (2006). Gene sequencing. The race for the $1000 genome. Science *311*, 1544–1546.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2012). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. *23*, 111–120.

Sharp, P.M., and Matassi, G. (1994). Codon usage and genome evolution. Curr. Opin. Genet. Dev. *4*, 851–860.

Shendure, J., and Aiden, E.L. (2012). The expanding scope of DNA sequencing. Nat. Biotechnol. *30*, 1084–1094.

Sicheritz-Ponten, T., and Andersson, S.G. (2001). A phylogenomic approach to microbial evolution. Nucleic Acids Res. *29*, 545–552.

Simpson, J.T., and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. Genome Res. *22*, 549–556.

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

Singh, A., Goering, R.V., Simjee, S., Foley, S.L., and Zervos, M.J. (2006). Application of molecular techniques to the study of hospital infection. Clin. Microbiol. Rev. *19*, 512–530.

Siren, J., Marttinen, P., and Corander, J. (2011). Reconstructing population histories from single nucleotide polymorphism data. Mol. Biol. Evol. *28*, 673–683.

Slack, A., Symonds, M., Dohnt, M., Harris, C., Brookes, D., and Smythe, L. (2007). Evaluation of a modified Taqman assay detecting pathogenic Leptospira spp. against culture and Leptospira-specific IgM enzyme-linked immunosorbent assay in a clinical environment. Diagn. Microbiol. Infect. Dis. *57*, 361–366.

Smith, G.J., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S., *et al.* (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature *459*, 1122–1125.

Souza, V., Rocha, M., Valera, A., and Eguiarte, L.E. (1999). Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. Appl. Environ. Microbiol. *65*, 3373–3385.

Sreevatsan, S., Pan, X., Stockbauer, K.E., Connell, N.D., Kreiswirth, B.N., Whittam, T.S., and Musser, J.M. (1997). Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. Proc. Natl. Acad. Sci. U.S.A. *94*, 9869–9874.

Stanley, T.G., and Wilson, I. (2003). Multilocus enzyme electrophoresis: a practical guide. Mol. Biotechnol. *24*, 203–220.

Stevens, J.R., and Tibayrenc, M. (1995). Detection of linkage disequilibrium in *Trypanosoma brucei* isolated from tsetse flies and characterized by RAPD analysis and isoenzymes. Parasitology *110(Pt 2)*, 181–186.

Stewart, A.C., Osborne, B., and Read, T.D. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics *25*, 962–963.

Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature *365*, 274–276.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. Proc. Natl. Acad. Sci. U.S.A. *85*, 2653–2657.

Suter, B., Auerbach, D., and Stagljar, I. (2006). Yeast-based functional genomics and proteomics technologies: the first 15 years and beyond. Biotechniques *40*, 625–644.

Swain, M.T., Tsai, I.J., Assefa, S.A., Newbold, C., Berriman, M., and Otto, T.D. (2012). A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. Nat. Protocols *7*, 1260–1284.

Swaminathan, B., Barrett, T.J., Hunter, S.B., and Tauxe, R.V. (2001). PulseNet: the molecular subtyping network

for foodborne bacterial disease surveillance, United States. Emerg. Infect. Dis. *7*, 382–389.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., *et al.* (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Than, C., Ruths, D., and Nakhleh, L. (2008). PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformat. *9*, 322.

Tibayrenc, M., and Ayala, F.J. (2012). Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. Proc. Natl. Acad. Sci. U.S.A. *109*, E3305–3313.

Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. *38*, e159.

Tritt, A., Eisen, J.A., Facciotti, M.T., and Darling, A.E. (2012). An integrated pipeline for *de novo* assembly of microbial genomes. PLoS One *7*, e42304.

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

Tsibris, A.M., Korber, B., Arnaout, R., Russ, C., Lo, C.C., Leitner, T., Gaschen, B., Theiler, J., Paredes, R., Su, Z., *et al.* (2009). Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy *in vivo*. PLoS One *4*, e5683.

Twyford, A.D., and Ennos, R.A. (2012). Next-generation hybridization and introgression. Heredity (Edinburgh) *108*, 179–189.

Urwin, R., and Maiden, M.C. (2003). Multi-locus sequence typing: a tool for global epidemiology. Trends Microbiol. *11*, 479–487.

Valouev, A., Zhang, Y., Schwartz, D.C., and Waterman, M.S. (2006). Refinement of optical map assemblies. Bioinformatics *22*, 1217–1224.

Van de Peer, Y., Chapelle, S., and De Wachter, R. (1996). A quantitative map of nucleotide substitution rates in bacterial rRNA. Nucleic Acids Res. *24*, 3381–3391.

van der Zanden, A.G., Kremer, K., Schouls, L.M., Caimi, K., Cataldi, A., Hulleman, A., Nagelkerke, N.J., and van Soolingen, D. (2002). Improvement of differentiation and interpretability of spoligotyping for Mycobacterium tuberculosis complex isolates by introduction of new spacer oligonucleotides. J. Clin. Microbiol. *40*, 4628–4639.

Van Ert, M.N., Easterday, W.R., Huynh, L.Y., Okinaka, R.T., Hugh-Jones, M.E., Ravel, J., Zanecki, S.R., Pearson, T., Simonson, T.S., U'Ren, J.M., *et al.* (2007). Global genetic population structure of *Bacillus anthracis*. PLoS One *2*, e461.

van Soolingen, D., de Haas, P.E., Hermans, P.W., Groenen, P.M., and van Embden, J.D. (1993). Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. J. Clin. Microbiol. *31*, 1987–1995.

Walker, M.J., and Beatson, S.A. (2012). Epidemiology. Outsmarting outbreaks. Science *338*, 1161–1162.

Walker, T.M., Ip, C.L., Harrell, R.H., Evans, J.T., Kapatai, G., Dedicoat, M.J., Eyre, D.W., Wilson, D.J., Hawkey, P.M., Crook, D.W., *et al.* (2012). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect. Dis. *13*, 137–146.

Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., *et al.* (2011). Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. *43*, 847–853.

Wetzel, J., Kingsford, C., and Pop, M. (2011). Assessing the benefits of using mate-pairs to resolve repeats in *de novo* short-read prokaryotic assemblies. BMC Bioinformat. *12*, 95.

Wong, K., Keane, T.M., Stalker, J., and Adams, D.J. (2010). Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biol. *11*, R128.

World Health Organization (2012). Global invasive bacterial vaccine preventable diseases (IB-VPD) information and surveillance bulletin.

Wren, B.W. (2000). Microbial genome analysis: insights into virulence, host adaptation and evolution. Nat. Rev. Genet. *1*, 30–39.

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol. *9*, R151.

Xi, R., Kim, T.M., and Park, P.J. (2010). Detecting structural variations in the human genome using next generation sequencing. Brief Funct. Genomics *9*, 405–415.

Xie, C., and Tammi, M.T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformat. *10*, 80.

Xu, S., Gupta, S., and Jin, L. (2010). PEAS V1.0: a package for elementary analysis of SNP data. Mol. Ecol. Resour. *10*, 1085–1088.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Yoder, A.D., and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. *17*, 1081–1090.

Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. *18*, 821–829.

Zhao, Q.Y., Wang, Y., Kong, Y.M., Luo, D., Li, X., and Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics *12(Suppl. 14)*, S2.

Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science *338*, 1622–1626.