BIOINFORMATICS REVIEW

Vol. 31 no. 1 2015, pages 1–9
doi:10.1093/bioinformatics/btu600

Genome analysis

Advance Access publication September 4, 2014

Insights into structural variations and genome rearrangements in prokaryotic genomes

Vinita Periwal[1,2] and Vinod Scaria[1,2,*]

[1]GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi 110007 and [2]Academy of Scientific & Innovative Research (AcSIR), Anusandhan Bhawan, New Delhi 110001, India

Associate Editor: Jonathan Wren

ABSTRACT

Structural variations (SVs) are genomic rearrangements that affect fairly large fragments of DNA. Most of the SVs such as inversions, deletions and translocations have been largely studied in context of genetic diseases in eukaryotes. However, recent studies demonstrate that genome rearrangements can also have profound impact on prokaryotic genomes, leading to altered cell phenotype. In contrast to single-nucleotide variations, SVs provide a much deeper insight into organization of bacterial genomes at a much better resolution. SVs can confer change in gene copy number, creation of new genes, altered gene expression and many other functional consequences. High-throughput technologies have now made it possible to explore SVs at a much refined resolution in bacterial genomes. Through this review, we aim to highlight the importance of the less explored field of SVs in prokaryotic genomes and their impact. We also discuss its potential applicability in the emerging fields of synthetic biology and genome engineering where targeted SVs could serve to create sophisticated and accurate genome editing.

Contact: vinods@igib.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 20, 2014; revised on August 11, 2014; accepted on August 31, 2014

1 INTRODUCTION

The sequencing of a large number of prokaryotic genomes and sequence comparison of closely related species have unraveled a large repertoire of genomic variations (Skovgaard *et al*., 2011; Srivatsan *et al*., 2008). A large number of studies focusing on variations in prokaryotic genomes have been majorly concentrated on single-nucleotide variations and small insertion–deletion events (Sun *et al*., 2012). It is now being increasingly recognized that in addition to single-nucleotide variations, other types of variations that include large genomic rearrangements are not infrequent in bacterial genomes (Darling *et al*., 2008; Sun *et al*., 2012). Among the different types of genetic variations found in genomes, structural variants have remained the most difficult to identify and interpret. Structural variations (SVs) encompass a fairly large piece of DNA and can emerge as a result of different cellular mechanisms such as DNA recombination, replication and DNA repair (Hastings *et al*., 2009b). SVs introduce variability in gene copy number, position, orientation and, in several cases, combinations of these events (Freeman *et al*., 2006). The mechanisms of SV formation appears to be similar in prokaryotes and eukaryotes (Hastings *et al*., 2009a, b). But, SVs have not been widely explored and studied in prokaryotic genomes as compared with eukaryotes (Kresse *et al*., 2003; Liang *et al*., 2010; Skovgaard *et al*., 2011).

The comprehensive assessment of SVs has been a challenge largely due to the underlying complex mechanisms that gives rise to them. Accurate and precise identification of SVs would require prediction of three features, namely, copy, content and structure (Alkan *et al*., 2011). Over the years, several approaches have been developed to detect and characterize SVs. Some of the classical methods include ArrayCGH (array comparative genomic hybridization) and single-nucleotide polymorphism (SNP) arrays, which have been extensively reviewed earlier (Alkan *et al*., 2011; Carter, 2007). ArrayCGH is based on hybridizing fluorescently labeled sample with normal DNA immobilized on a glass surface and analyzing hybridization ratios. SNP arrays on the other hand use single sample per array and measure the intensities of the probe signals on the basis of single base difference.

The major understanding on the genomic landscape of SVs was facilitated by the availability of sequencing technologies coupled with computational algorithms to map and identify SVs at a much higher resolution (Skovgaard *et al*., 2011; Sun *et al*., 2012). In sequencing-based approaches, paired-end reads generated with an approximate insert size are mapped onto the reference genome. The pairs mapping at a distance substantially different from expected length or in altered orientation are nominated as structural variants. SVs identified by means of sequencing data offer a strong advantage of detecting 'breakpoints' (i.e. sequence boundaries where an SV begins and ends) (Jiang *et al*., 2012; Sindi *et al*., 2009; Ye *et al*., 2009), which can further aid to unravel the functional impact of a variant and the mechanisms that created it (Bao *et al*., 2014). Though the high-throughput technologies have significantly contributed to the understanding of the repertoire of SVs in prokaryotic genomes, the problem of SV detection has always remained challenging as none of the methods can appropriately address the complexity of repetitive regions found in genomes.

It is widely believed that the SVs arise as a by-product of illegitimate recombination events (homologous recombination) and also by imprecise non-homologous repair mechanism

*To whom correspondence should be addressed.

during aberrant DNA replication to repair broken replication forks (Hastings *et al*., 2009a, b). Apart from being the focus of evolutionary analysis (Lim *et al*., 2012), the phenotypic implications of SVs in bacterial genomes have also been deciphered (Cui *et al*., 2012; Dobinsky *et al*., 2002; Nagarajan *et al*., 2012), which are discussed in detail in the later part of the manuscript.

In this review, we provide a comprehensive overview of the present understanding of SVs in general and in the context of prokaryotic genomes. We briefly describe the various types of SVs, discuss their probable molecular mechanisms of formation, advances in the development of tools and techniques to detect SVs and also their phenotypic consequences in context of prokaryotic genomes. We also discuss currently used methodologies of next-generation sequencing (NGS) and analysis algorithms, which could provide a comprehensive and high-resolution map of SVs and how they could be extensively used for understanding biological phenomena of strain variability and evolution. We also describe their potential applications in the emerging fields of synthetic biology and genome engineering.

2 GENERAL OVERVIEW OF SVS

SVs involve long stretches of DNA that can span from a few kilobases to sometimes up to millions of base pairs in length. Chromosomal rearrangements can result in loss, amplification (Andersson *et al*., 1998), translocation (Block *et al*., 2012) and inversions (Johnson, 1991) of DNA fragments. SVs can contribute to evolution of an organism through disruption of an existing gene (Jasin and Schimmel, 1984), creation of a new gene (Nagarajan *et al*., 2012) or a chimeric gene product through gene fusions (Nogami *et al*., 1985; Roth *et al*., 1996). In bacterial genomes, chromosomal rearrangements can change the distance of a gene from the origin of chromosome replication (oriC) leading to altered gene copy number and thereby affecting its expression (Rebollo *et al*., 1988).

SVs can be broadly classified into five major classes—Deletions, Duplications, Insertions, Inversions and Translocations. Each of these discrete events is caused by a double-strand break involving at least two different locations, followed by a re-ligation of the broken ends to produce a new chromosomal arrangement or context at the ends (Hastings *et al*., 2009a, b; Roth *et al*., 1996). The rearrangements could widely vary in their lengths, ranging from a few thousand nucleotides to a few million nucleotides. In some cases, the rearrangement could encompass genes, even operons or a large number of genes depending on the size of the rearranged fragment (Hastings *et al*., 2009b). The different types of SVs and their genomic contexts are summarized in Supplementary Figure S1. The functional consequences of the SVs could therefore vary widely.

The unbalanced SVs having a net gain or loss of genetic material include deletions, duplications and insertions (Bentley and Parkhill, 2004). Deletions entail loss of a genomic segment, and could be intragenic, wherein they result in inactivation of a gene or the loss of one or more functional domains or an altered gene function. In case of intergenic deletions, they could potentially affect the regulatory regions, thereby affecting the expression of neighboring genes (Angov and Brusilow, 1994). Duplications are marked by the presence of two or more copies of a genomic region or a genomic segment (Anderson and Roth, 1977). The duplicated regions may either lie adjacent to each other, referred to as tandem duplication (Wang *et al*., 1982) or could occur at a different genomic location termed as insertional duplication. Duplication generally results in gain of a copy of the DNA segment carrying information (Roth *et al*., 1996). The functional consequence of the duplication could vary depending on the information content of the duplicated genomic segment and also on the context in which it is inserted (Reams and Neidle, 2004). The third class of unbalanced SVs is the insertions. Insertion involves gain of a genomic segment through a double-stranded break (DSB). The well-studied example of insertions is the Horizontal Gene Transfer (HGT) (Nelson *et al*., 1999; Ochman *et al*., 2000). HGT results in the gain of a new genomic segment in a new genomic context. In addition to novel sequence insertion, mobile-element insertion can also lead to SVs (Xing *et al*., 2009). Mobile elements jump from one position to another within a genome often resulting in duplication. The functional consequences of insertions are governed by the information content of the inserted fragment and the context of the genomic segment of insertion (Dobinsky *et al*., 2002).

The balanced SVs comprise inversions. Inversions are variations that involve a rearrangement of the orientation of a genomic segment (Johnson, 1991). These are copy-invariant SVs because there is no net gain or loss of genomic information. Typically, inversions involve two breakpoints and realignment of the flipped ends. The functional consequences of inversions are potentially guided by their new genomic contexts (Johnson, 1991).

SVs could also involve exchange of a genomic segment from one context to another within the same chromosome or between chromosomes and are classified as translocations. Translocations can be either balanced with the retention of full genetic functionality or unbalanced with loss or gain of functional elements (Block *et al*., 2012). This type of SV is particularly more evident and common in multi-chromosomal bacteria, where the smaller secondary chromosomes evolve more rapidly (Morrow and Cooper, 2012).

3 MOLECULAR PREDISPOSITION AND MECHANISM OF STRUCTURAL VARIABILITY IN PROKARYOTIC GENOMES

Apart from the above well-defined classes of SVs, complex SVs that include combinations of two or more of these broad classes are not uncommon to observe in real-life situations (Hastings *et al*., 2009b). A number of studies have highlighted the molecular predispositions that enable SVs to occur. This includes a wide variety of chromosomal contexts such as sequence and structural motifs, repeat elements, insertion sequence (IS) elements and transposon elements (TE) (Mahillon and Chandler, 1998; Treangen *et al*., 2009). In organisms with repetitive DNA, homologous repetitive segments within one chromosome or on different chromosomes can serve as sites for illegitimate crossing-over. Bacterial DNA consists of an extensive array of repetitive sequences, which significantly underlie genomic instability and contain recombination hotspots (Aras *et al*., 2003; Treangen *et al*., 2009). DNA repeats increases the chances of

rearrangement through recombination, amplification and deletion of genetic material, thereby leading to genome plasticity (Aras *et al*., 2003; Bao *et al*., 2014). Generic repeats may arise by HGT whereby the incoming DNA fragment contains the information already present in the host genome and integrates seamlessly into the host genome using site-specific recombination (Treangen *et al*., 2009).

There have been some important functional consequences of repetitive elements in chromosome rearrangements. Naas *et al*. (1995) suggested that in resting *Escherichia coli* strain, the spontaneous mutagenesis is IS-specific and the observed genetic polymorphism is a consequence of DNA rearrangements. The effects of deletions are highly irreversible and can be explained by the loss of functions. The effects of large inversions on the other hand result from selection for chromosome organization (Rocha, 2008). The *Neisseria* species contains an extensive array of repetitive sequences such as tandem repeats and IS elements spread throughout its genome. Comparative genome analysis of *Neisseria meningitidis* revealed that repeats are involved in three major inversion events (Bentley *et al*., 2007). Additionally, the bacterial species *Neisseria gonorrhoeae*, which contains fewer repeat elements than *N.meningitidis*, also showed that rearrangements are associated with IS elements leading to inversion (Spencer-Smith *et al*., 2012). In addition to the sequence elements, other structural elements including Z-DNA (Freund *et al*., 1989) and G-quadruplex motifs have been implicated in predisposing specific genomic loci to recombination (Katapadi *et al*., 2012).

4 PHENOTYPIC IMPACT OF SV AND CHROMOSOMAL REARRANGEMENTS

Genome rearrangements in prokaryotes have also been studied in relation to their phenotypic outcomes. Recent studies suggest that the genomic rearrangements and SVs have a profound impact on the phenotypic outcomes in a number of organisms (Cui *et al*., 2012; Gaudriault *et al*., 2008; Okinaka *et al*., 2011). Both balanced and unbalanced forms of variation have remained difficult to interpret with respect to their functional consequence. Though many variant calling technologies have enabled the identification and characterization of SVs (Skovgaard *et al*., 2011; Sun *et al*., 2012), the large size of the SVs and the complexities of their rearrangements have posed challenges in deciphering their functional consequences. In addition, the molecular, cellular and mechanistic insights into their formation and resultant phenotype remain largely obscure (Weischenfeldt *et al*., 2013). Some important studies emphasizing the functional impact of SVs in prokaryotic genomes have been established (Darling *et al*., 2008). Large-scale rearrangements in closely related strains of a species, for example, in the case of *Yersinia pestis*, have shown to significantly contribute to the evolution, divergence and pathogenicity of the organism (Liang *et al*., 2010).

Deletions and duplications can potentially lead to altered doses of otherwise functionally intact elements. Phenotypic effects of deletions depend on the size and the location of deleted chromosomal segments on the genome. Larger deletions are likely to involve many genes, thereby resulting in more drastically altered phenotypes (Srivatsan *et al*., 2008). Deletions encompassing loss of essential genes or gene components may significantly hamper cell viability (Jasin and Schimmel, 1984). On the other hand, gene duplication can have four possible outcomes (Treangen *et al*., 2009): (i) non-functionalization or loss of the duplicated gene through deletion or degeneration; (ii) subfunctionalization, resulting in adoption of complementary roles; (iii) neo-functionalization, resulting in new functions (Blount *et al*., 2012); and (iv) differential regulation of duplicated genes allowing spatiotemporal expression. The last possibility has been explored by Nagarajan *et al*. where they show that both the operons of duplicated hik31 are temporally and differentially regulated and share an integrated regulatory relationship (Nagarajan *et al*., 2012).

Gene deletions could also arise from recombination events involving repeats (Gaudriault *et al*., 2008). It has been proposed that these deletions arise because these genes are particularly rich in closely spaced repeats (Achaz *et al*., 2002), for example, the deletion of 54 nt between two 8 nt length direct repeats in the mismatch repair gene, mutS of *Pseudomonas aeruginosa* show close association of the deletion with the repeats at the edges (Oliver *et al*., 2002). In another study, experimental deletion of the mutS gene of *E.coli* showed that the truncated protein introduces variability in DNA binding, dimerization and its interactions with mutL (Wu and Marinus, 1999).

Large chromosomal inversions were initially considered to be rare in bacteria (Roth *et al*., 1996), but recent studies show these to be among the major forces responsible for genome evolution (Hughes, 2000). The two closely related species *Salmonella typhimurium* and *E.coli* displayed contrasting behavior in terms of permissibility of inversions. Though in *E.coli*, it was earlier shown that inversions could lead to viable cells with reduced cell growth (Hill and Harnish, 1981) but later it was also demonstrated that the inversions introduced in the non-permissive regions of the chromosome were refractory in nature meaning that they are mechanistically feasible but lethal (Guijo *et al*., 2001; Rebollo *et al*., 1988). However, in *S.typhimurium* no such region was identified and the failure to introduce inversions was suggested to be a mechanistic problem (Miesel *et al*., 1994; Segall *et al*., 1988). In yet another supporting evidence, Campo *et al*. (2004) characterized the two constrained chromosomal regions in gram-positive bacteria (the oriC and ter domain) where they showed that introduction of inversions is mechanistically possible but leads to reduced cell fitness and lethal cell phenotype.

Acquisition of mobile genetic elements through HGT in *Staphylococcus aureus* contributes to its genotypic and phenotypic diversity (Deurenberg *et al*., 2007; Diep *et al*., 2006). Clones of *S.aureus* are thought to evolve by point mutations instead of genetic recombination, which are considered as rare in this species (Feil *et al*., 2003). Introduction of mobile genetic elements by site-specific recombinases can bestow epidemiological advantage to the pathogen with traits such as survival under low pH conditions, and stressed environments, or drug-resistant strains (Deurenberg *et al*., 2007; Diep *et al*., 2006). Genomic rearrangements can confer drug resistance and aid in pathogen evolution such as the evolution of pandemic strains in *Y.pestis* as a result of accumulation of rearrangements (Liang *et al*., 2010). Alteration in the gene pool of a genome is central to adaptive evolution. Reconstructing the genome synteny evolution can contribute to understanding of the dynamics of

genome evolution. Genome rearrangements can lead to gain or loss of genes and can help to understand the long- and short-term genome evolution (Furuta *et al*., 2011). Although evolutionary implications of SVs have been shown at large, few studies have highlighted their functional and biological importance. For instance, the reversible switching on–off of colony variation in *S.aureus* suggest a survival strategy for coping with uncertain and variable environments (Cui *et al*., 2012). Some key examples highlighting phenotypic changes associated with SVs in bacterial genomes are presented in Supplementary Table S1.

Apart from creating SVs in the genome, TEs and IS elements can also influence the expression of the genes, depending on the context of the gene in relation to the element. TEs can affect the expression of closely placed and nearby genes by either interrupting them or enhancing their expression (Brown and Evans, 1991). SVs involving IS elements have been shown to activate the expression of neighboring genes (Hubner and Hendrickson, 1997; Mahillon and Chandler, 1998). A conceptual overview of the potential functional effects of SVs is summarized in Figure 1.

5 SYMMETRY OF GENOME REARRANGEMENTS

Large rearrangements have been shown to be highly deleterious in prokaryotic genomes (Rocha, 2008). The symmetrical organization of bacterial chromosome along the replichores (Eisen *et al*., 2000) leads to biased symmetrical genome rearrangements. Three selection forces have been hypothesized by Mackiewicz *et al*. (2001) for the biased symmetrical genome rearrangements. Firstly, the distance of a gene from oriC has been shown to be a major selection force. This arises from the fact that the copy number of transcripts is significantly variable and dependent on the distance of the gene from the oriC, and a strong positional bias for genes with specific functional attributes have been observed. The selection pressure leads to optimally placed genes with respect to oriC for genes with certain functional attributes. Second, difference in replication associated mutational pressure in leading and lagging strands (Mackiewicz *et al*., 2001). Symmetrical inversion of genes encompassing the oriC does not change gene location with respect to the leading and lagging DNA strands (Eisen *et al*., 2000). Thirdly, the constraint of keeping both replichores of same size leads to symmetrical inversions at oriC and ter. Large inversions, which disturb the symmetry around the constrained chromosomal regions, i.e. the oriC and ter domains, have been suggested to result in reduced cell fitness (Campo *et al*., 2004). Symmetrical inter-replichore inversions are more frequently observed than any other rearrangement. Chromosome breakpoint analysis in naturally evolving population of *Yersinia* showed that the rate of inversion is three times higher near oriC, thereby, displaying recombination bias and selective forces reducing the inversion rate near ter (Darling *et al*., 2008).

6 METHODS FOR DETECTING SVS

6.1 Classical approaches

Technological advancements for detecting genomic SVs have facilitated our ability to investigate details of genome structure, patterns and extent of genome organization in various organisms (Quail *et al*., 2012; Sun *et al*., 2012). The classical approaches used to detect SVs include conventional cytogenetic methods of 'chromosome banding', which is staining of condensed chromosomes for precise identification of individual chromosomes or parts of chromosomes, and 'karyotyping', which is the process of pairing and ordering all the chromosomes of an organism. These methods were limited by their low throughput and low resolution (Alkan *et al*., 2011). High-density genotyping microarrays involved hybridization of nucleic acid sample to a large set of probes to detect variations in genes. Although they were high on throughput, they were limited by detecting only small copy-number variants (Iafrate *et al*., 2004). The resolution and scalability of SV detection was improved with use of an 'optical mapping' technique, which was based on restriction mapping and allowed identification of fine-scale structural analysis of genomes but on other hand was limited by its dependency on a reference genome (Teague *et al*., 2010). Other techniques include 'DNA barcoding', which uses a short genetic marker in an

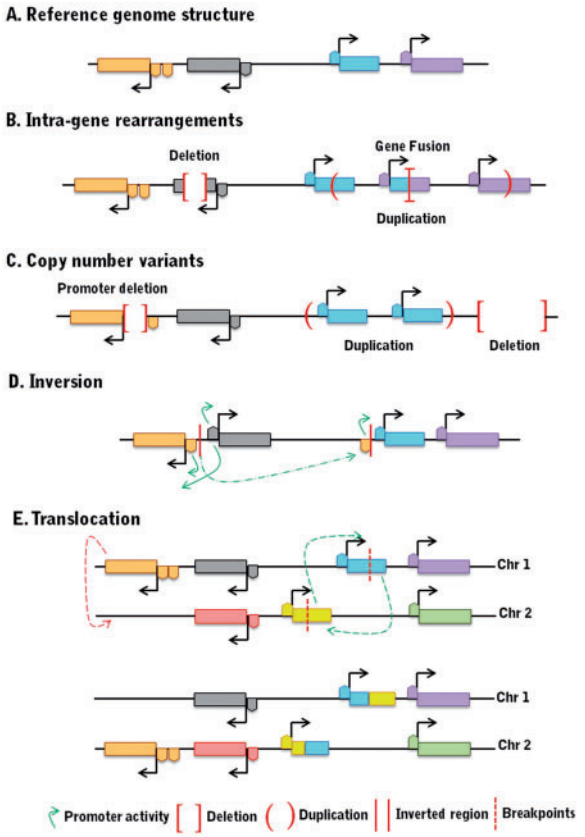

Fig. 1. Conceptual overview of functional consequences of SVs. (**A**) Genomic region without any SV. The boxes represent genes, and the small connectors beside them represent the promoter of respective gene. (**B**) Intragene rearrangement depicting deletion and gene fusion as a result of partial gene duplication. Altered coding regions result in aberrant transcripts. Deletion or duplication can lead to altered gene dosage of otherwise functionally intact regions (**C**) Change in copy number as a result of deletion (altered regulation) and duplication. (**D**) Inversions affecting gene structure, the gene gets inverted and flipped and rearranges, thereby pushing one of the promoters of first gene (orange) away from it. (**E**) Translocations affecting genic context

organism's DNA to identify species. It allows detection of balanced rearrangements by making use of fluorescently labeled nano-channel flow cells (Das *et al*., 2010). The 'emulsion droplet PCR' method generates water-in-oil emulsion droplets containing all polymerase chain reaction components and has been used for estimating absolute copy number (Beer *et al*., 2007). The major impetus toward understanding genomic SVs sprang from the availability of whole-genome sequences of multiple prokaryotic species and, in many cases, multiple closely related strains of a given species (Tatusova *et al*., 2014).

6.2 Detecting rearrangements by whole genome comparisons

A large number of tools and methods offer information on rearrangements by pairwise comparative analysis of sequences. 'GeneOrder3.0' developed by Celamkoti *et al*. (2004) allows the rapid identification and visualization of gene order and synteny between two bacterial genomes under comparison without any a priori knowledge or information of their phylogenetic relatedness. It has been tested on the smallest bacterial genomes: *Mycoplasma*, *Haemophilus* and few other bacteria of size <2 MB. It is an enhanced version of previous algorithms GeneOrder (Mazumder *et al*., 2001).

Structural changes in genome are common occurrences during evolution of prokaryotic species. The earlier methodologies for annotating SVs in prokaryotic genomes have been majorly focused on pairwise comparison of genomes. In 2004, Darling *et al* introduced a multiple genome comparison and visualization method called Mauve, which identifies rearrangements and inversions in conserved regions, small in-dels and the exact sequence breakpoints. When used to align nine enterobacterial genomes it was able to resurrect most of the known inversions (Darling *et al*., 2004). Comparing rearrangements in multi-chromosome genomes proved to be a more daunting task. In 2006, Lu *et al* proposed a novel algorithm called 'FFBI' (Fusion, fission and block interchanges), to compare circular multi-chromosome genomes such as the *Vibrio* and *Burkholderia* pathogens (Lu *et al*., 2006). The FFBI algorithm was designed to analyze genome rearrangements arising as a result of chromosome fusion, fission and blocks interchanges. Genome rearrangements detected in the three *Vibrio* genomes were coherent with earlier studies.

In 2011, a highly computationally efficient whole genome alignment tool called Mugsy was introduced, which detected duplications, rearrangements and large gain/loss in genomes (Angiuoli and Salzberg, 2011). Mugsy's performance was evaluated on 57 *E.coli* and 31 *Streptococcus pneumoniae* genomes. It works without reference genomes.

Recently, Marttinen *et al*. (2012) developed a statistical algorithm called 'BratNextGen' that allows large-scale comparison of recombination events in hundreds of complete bacterial genomes. They compared 241 whole genomes of *S.pneumoniae* to detect ~39k polymorphic sites over the 2 MB genome aligned. BratNextGen functions by creating a Bayesian clustering model, to detect recombination in taxa along with resampling.

Large chromosomal rearrangements studies done using traditional genetic mapping or whole genome comparisons have been challenging in terms of accuracy and throughput. The availability of a gamut of new technologies for high-throughput nucleotide sequencing have opened up new opportunities toward understanding genome structure and their variations.

6.3 Detecting SVs using NGS

Whole genome sequencing (WGS) by means of NGS platforms has been used to detect mutations (Srivatsan *et al*., 2008). Though tools for mapping of paired-end information for SV calling are plenty, however, each may have their own limitations, advantages and overheads of usage. Additionally, most of them have been tested on eukaryotic (specifically humans) genomes and require additional confirmation so as to confidently rely on their output (Chen *et al*., 2009; Hormozdiari *et al*., 2009; Korbel *et al*., 2009; Zeitouni *et al*., 2010). Nevertheless, as the mechanism of SV formation in both eukaryotes and prokaryotes appears similar (Hastings *et al*., 2009a, b), the tools and algorithms could essentially be applied to study SVs in prokaryotes as well. Data from short-read sequencing generally lead to incomplete genome assemblies because of the intractable complexities such as long repetitive regions found in genomes (Huddleston *et al*., 2014). However, with the availability of long reads sequencing technologies with read lengths sometimes extending to tens of kilobases from single molecule sequencing approaches (Quail *et al*., 2012) to generate finished microbial genomes (Chin *et al*., 2013), provides a new opportunity toward identifying SVs and genome dynamics at a higher resolution.

A plethora of analytical algorithms and techniques have been developed over the years to precisely detect SV boundaries (Fig. 2) with increased resolution ranging from few mega base pairs to kilo base pairs and recently to even at the single nucleotide resolution (Chen *et al*., 2009; Medvedev *et al*., 2009; Zeitouni *et al*., 2010). A comprehensive list of sequence signatures that can be used to efficiently call SVs is detailed by Alkan *et al*. (2011). In the following section, we review some of these algorithms and techniques with their potential applications and limitations.

6.3.1 Tools and algorithms to detect SVs using NGS data A selected list of available software and resources for detecting and analyzing gene rearrangement and SVs is presented in

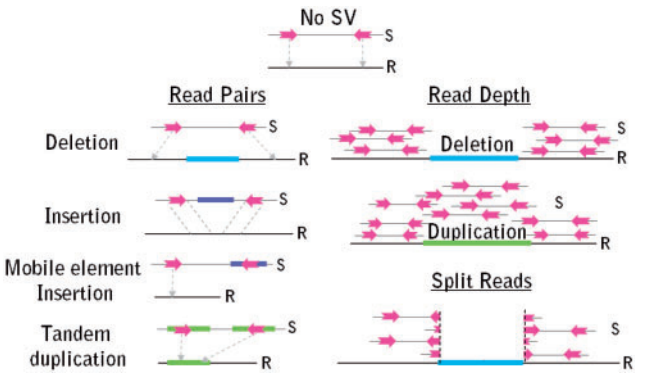

Fig. 2. Some of the commonly used analytical ways of detecting SVs. Paired-end reads, Read depth and Split reads could be used to call SVs. Read depth refers to the number of reads mapping onto a particular part of the genome. In Split reads, a single read maps to two different parts of the genome that lie far away from each other

Table 1 and a comprehensive list of same is provided in Table S2. In this section, we would discuss some of the algorithms developed in the past 5 years. The algorithm *PEMer* (*Paired-End Mapper*) could process data from several NGS platforms and mapped SVs at a high resolution (Korbel *et al*., 2009). *PEMer* maps SVs at a higher resolution with a confidence measure and allows storage, display and manipulation of SV data. *PEMer* can detect insertion, deletion and inversions. Despite a few limitations of *PEMer* such as its inability to detect breakpoints in repetitive regions, SNP-based misalignment errors and missing out large insertions, it appears to be a useful method for calling variants.

Next-generation VariationHunter introduced in 2010 (Hormozdiari *et al*., 2010) was an improved version of previously published combinatorial algorithm VariationHunter (Hormozdiari *et al*., 2009). It makes use of maximum parsimony algorithm to map paired-end reads obtained from NGS. Compared with its earlier release, the new algorithm could resolve incompatible SV calls and requires no post-processing of results. Besides insertions, inversions and deletion events, it also has the capability to detect mobile element insertions. It has better accuracy than MoDIL, which can detect 20–50 bp indels (Lee *et al*., 2009) and BreakDancer (Chen *et al*., 2009).

Skovgaard introduced a novel way of using this technology by combining it with copy-number analysis of template DNA in fast-growing bacterial cultures of *E.coli* (Skovgaard *et al*., 2011). Instead of mapping reads obtained from stationary culture where copy number is constant, they obtained clear contrasting behavior of reads from DNA of exponentially growing cultures confirming presence of a large inversion.

In 2012, Sun *et al* (2012) used 454 pyro-sequencing combined with a 'split mapping' computational method to detect spontaneously occurring genome rearrangements (SGRs) from fast-growing culture of *Salmonella sp*.. Breakpoints were determined to base pair resolution, and experimental verification of breakpoints of SGRs was carried out by padlock probe hybridization (Sun *et al*., 2012).

Detecting SVs accurately from real data can be challenged by limitations in experiment designs, few validated SVs and low resolution of breakpoints. In contrast to this, there is equal probability of imperfections present in the algorithms design rather than the laboratory issues such as reporting of false positives or missing out true positives. In light of this, in 2013, Bartenhagen devised a tool RSVSim, which is a fully simulated approach for detecting SVs (Bartenhagen and Dugas, 2013). It can simulate five types of most common SVs for practically any type of genome. Artificially rearranged genomes from RSVSim can serve as performance evaluators for NGS-based algorithms.

Apart from the above discussed tools and algorithms for SV detection, there are a number of other tools (Supplementary Table S2) as well that can be explored in context of prokaryotic genomes and are worth mentioning such as Pindel (Ye *et al*., 2009), SegSeq (Chiang *et al*., 2009), BreakDancer (Chen *et al*., 2009), SVDetect (Zeitouni *et al*., 2010), DELLY (Rausch *et al*., 2012), SVM^2 (Chiara *et al*., 2012) and PRISM (Jiang *et al*., 2012). More details about these tools, what they detect and their advantages and limitations are provided in Supplementary Table S2.

7 CONCLUSIONS AND FUTURE DIRECTIONS

The present knowledge and repertoire of SVs in prokaryotic genomes is limited to a handful of examples. Although the phenotypic consequences of many of the SVs are not well understood, the known repertoire of phenotypic consequences suggests their role in a wide spectrum of physiological and phenotypic outcomes (Deurenberg *et al*., 2007). The lack in the understanding

Table 1. Major computational approaches/tools for detecting and analyzing SVs (also see Supplementary Table S2 for a comprehensive list)

Approach/tool	Description	Detection	Advantages	Limitations
PEMer (Korbel *et al*., 2009)	Modularized framework for detecting SVs. Includes read mapping, filtering low-quality reads, signature detection and clustering	Insertions, deletions, inversions and more complex events	High resolution and also high sensitivity	Can not detect segmental duplications particularly in repetitive regions; can include SNP-based misalignment errors; large insertions (>3 kb) could be missed
Next-generation VariationHunter (Hormozdiari *et al*., 2010)	Identifies the most parsimonious mapping of paired end reads. Calculates probability of each SV.	Insertion, inversion, deletion and mobile element insertion	Identifies transposition events and removes ambiguity in variation discovery	—
NGS and copy-number analysis (Skovgaard *et al*., 2011)	Combines WGS and copy-number analysis for detecting rearrangements	Inversion, duplication	Detects point mutations, single and dinucleotide indels and major genomic rearrangements	Efficient for cultivable and fast-growing prokaryotes.
454 pyrosequencing and 'split mapping' (Sun *et al*., 2012)	Identifies unique junction sequences formed by spontaneous genome rearrangements	Deletion, duplication and inversion	Identifies junction sequences at base pair resolution	Efficient for cultivable and fast-growing prokaryotes
RSVSim (Bartenhagen, 2013)	SVs are simulated randomly, based on user-supplied genomic coordinates or associated to various kinds of repeats	Deletions, insertions, inversions, tandem duplications and translocations	Covers a wide range of SVs and considers size of SVs and their mechanism of formation	—

of genomic landscape of structural variability and its phenotypic consequences in prokaryotes was primarily due to the paucity of large-scale genome data for closely related organisms, and recent evidence suggests that this is rapidly changing (Skovgaard *et al*., 2011; Srivatsan *et al*., 2008). The availability of high-throughput sequencing technologies (NGS) offers the throughput, scale and cost-effectiveness required to do genome-wide associations for specific traits or phenotypes in prokaryotes. This would require two major gaps in the area to be addressed. The primary one being methodologies to phenotypically screen large populations of prokaryotes and secondly the computational algorithms, which can decipher, map and report genomic variations at large scale for these screens in short time.

Also the availability of longer read lengths that could encompass repeat regions could also provide immense insights into SVs in prokaryotic genomes. The deciphering of the genomic variability in *E.coli* strain from the German outbreak is worth mentioning (Rasko *et al*., 2011) where longer read lengths and computational algorithms were extensively used to decipher additional SVs, which were not previously annotated, and could provide new insights into the evolution and pathogenicity of the strain. The availability of datasets in the public domain for a number of prokaryotic species (Tatusova *et al*., 2014) provides a unique opportunity to understand the structural variability in prokaryotes. Although the phenotypic correlates for many of these strains would not be available, the extent of information could, however, provide enormous insights toward deriving a baseline map of polymorphic SVs in these genomes, a much needed resource toward inferring phenotypic correlates.

The in-depth understanding of SVs and its phenotypic consequences would see its widespread applications in a number of areas. The major area that could benefit from this knowledge is Synthetic Biology (Marguet *et al*., 2007). The understanding of gene arrangement, organization and positioning could extensively be used to engineer new genes and pathways and contribute to a set of rules for designing and engineering genomes. As interest in engineering bacterial genomes increases, the need for developing efficient tools for their successful manipulation also increases. Site-specific recombination system such as the Cre-lox system has been used to create deletions in *E.coli* (Fukiya *et al*., 2004) and large inversions in *Lactococcus lactis* (Campo *et al*., 2004). Recently, a new technology called GETR (Genome Editing via Targetrons and Recombinases) has been introduced for genome engineering of practically any bacteria (Enyeart *et al*., 2013). The technique has been efficiently used to introduce insertions, deletions, inversions and translocations in *E.coli*, *S.aureus* and *Bacillus subtilis*. For efficient genome engineering, it is required to induce DSBs in the DNA to initiate the recombination process. For a long time, it was not possible to induce DSBs owing to the lack of means to target DSBs to specific sites, however, with the introduction of technologies to synthesize and assemble large fragments of DNA (Ellis *et al*., 2011), and tools capable of accurately editing genomic regions have made it possible. The availability of genome editing tools like ZF-TFs (zinc finger transcription factors; Gommans *et al*., 2005), CRISPR/CAS [clustered regularly interspaced short palindromic repeats/CRISPR-associated (Cas) systems; Cong *et al*., 2013] and TALENS (Transcription activator-like effector nucleases; Miller *et al*., 2011) could provide the much necessary technological prowess to be able to accurately engineer genomes for strain improvements toward specific applications. It may not be excessively optimistic to believe that the rules of genome organization would find extensive application in genome engineering and genome design.

ACKNOWLEDGEMENTS

The authors acknowledge critical comments from Dr Srinivasan Ramachandran and Dr Sridhar Sivasubbu from CSIR-IGIB.

Funding: Authors acknowledge funding from Council for Scientific and Industrial Research (CSIR), India (OLP1105).

Conflict of interest: none declared.

REFERENCES

Achaz,G. *et al.* (2002) Origin and fate of repeats in bacteria. *Nucleic Acids Res.*, **30**, 2987–2994.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Anderson,R.P. and Roth,J.R. (1977) Tandem genetic duplications in phage and bacteria. *Annu. Rev. Microbiol.*, **31**, 473–505.

Andersson,D.I. *et al.* (1998) Evidence that gene amplification underlies adaptive mutability of the bacterial lac operon. *Science*, **282**, 1133–1135.

Angiuoli,S.V. and Salzberg,S.L. (2011) Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, **27**, 334–342.

Angov,E. and Brusilow,W.S. (1994) Effects of deletions in the uncA-uncG intergenic regions on expression of uncG, the gene for the gamma subunit of the *Escherichia coli* F1Fo-ATPase. *Biochim. Biophys. Acta*, **1183**, 499–503.

Aras,R.A. *et al.* (2003) Extensive repetitive DNA facilitates prokaryotic genome plasticity. *Proc. Natl Acad. Sci. USA*, **100**, 13579–13584.

Bao,Z. *et al.* (2014) Genomic plasticity enables phenotypic variation of *Pseudomonas syringae* pv. *tomato DC3000*. *PLoS One*, **9**, e86628.

Bartenhagen,C. and Dugas,M. (2013) RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics*, **29**, 1679–1681.

Beer,N.R. *et al.* (2007) On-chip, real-time, single-copy polymerase chain reaction in picoliter droplets. *Anal. Chem.*, **79**, 8471–8475.

Bentley,S.D. and Parkhill,J. (2004) Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.*, **38**, 771–792.

Bentley,S.D. *et al.* (2007) Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet.*, **3**, e23.

Block,D.H. *et al.* (2012) Regulatory consequences of gene translocation in bacteria. *Nucleic Acids Res.*, **40**, 8979–8992.

Blount,Z.D. *et al.* (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, **489**, 513–518.

Brown,N.L. and Evans,L.R. (1991) Transposition in prokaryotes: transposon Tn501. *Res. Microbiol.*, **142**, 689–700.

Campo,N. *et al.* (2004) Chromosomal constraints in Gram-positive bacteria revealed by artificial inversions. *Mol. Microbiol.*, **51**, 511–522.

Carter,N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.*, **39**, S16–S21.

Celamkoti,S. *et al.* (2004) GeneOrder3.0: software for comparing the order of genes in pairs of small bacterial genomes. *BMC Bioinformatics*, **5**, 52.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Chiang,D.Y. *et al.* (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Chiara,M. *et al.* (2012) SVM(2): an improved paired-end-based tool for the detection of small genomic structural variations using high-throughput single-genome resequencing data. *Nucleic Acids Res.*, **40**, e145.

Chin,C.S. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.

Cong,L. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.

Cui,L. *et al.* (2012) Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria. *Proc. Natl Acad. Sci. USA*, **109**, E1647–E1656.
Darling,A.C. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.
Darling,A.E. *et al.* (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.*, **4**, e1000128.
Das,S.K. *et al.* (2010) Single molecule linear analysis of DNA in nano-channel labeled with sequence specific fluorescent probes. *Nucleic Acids Res.*, **38**, e177.
Deurenberg,R.H. *et al.* (2007) The molecular evolution of methicillin-resistant *Staphylococcus aureus*. *Clin. Microbiol. Infect.*, **13**, 222–235.
Diep,B.A. *et al.* (2006) Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet*, **367**, 731–739.
Dobinsky,S. *et al.* (2002) Influence of Tn917 insertion on transcription of the icaADBC operon in six biofilm-negative transposon mutants of *Staphylococcus epidermidis*. *Plasmid*, **47**, 10–17.
Eisen,J.A. *et al.* (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biol.*, **1**, RESEARCH0011.
Ellis,T. *et al.* (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol. (Camb)*, **3**, 109–118.
Enyeart,P.J. *et al.* (2013) Generalized bacterial genome editing using mobile group II introns and Cre-lox. *Mol. Syst. Biol.*, **9**, 685.
Feil,E.J. *et al.* (2003) How clonal is *Staphylococcus aureus*? *J. Bacteriol.*, **185**, 3307–3316.
Freeman,J.L. *et al.* (2006) Copy number variation: new insights in genome diversity. *Genome Res.*, **16**, 949–961.
Freund,A.M. *et al.* (1989) Z-DNA-forming sequences are spontaneous deletion hot spots. *Proc. Natl Acad. Sci. USA*, **86**, 7465–7469.
Fukiya,S. *et al.* (2004) An improved method for deleting large regions of *Escherichia coli* K-12 chromosome using a combination of Cre/loxP and lambda Red. *FEMS Microbiol.Lett.*, **234**, 325–331.
Furuta,Y. *et al.* (2011) Birth and death of genes linked to chromosomal inversion. *Proc. Natl Acad. Sci. USA*, **108**, 1501–1506.
Gaudriault,S. *et al.* (2008) Plastic architecture of bacterial genome revealed by comparative genomics of Photorhabdus variants. *Genome Biol.*, **9**, R117.
Gommans,W.M. *et al.* (2005) Engineering zinc finger protein transcription factors: the therapeutic relevance of switching endogenous gene expression on or off at command. *J. Mol. Biol.*, **354**, 507–519.
Guijo,M.I. *et al.* (2001) Localized remodeling of the *Escherichia coli* chromosome: the patchwork of segments refractory and tolerant to inversion near the replication terminus. *Genetics*, **157**, 1413–1423.
Hastings,P.J. *et al.* (2009a) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.*, **5**, e1000327.
Hastings,P.J. *et al.* (2009b) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
Hill,C.W. and Harnish,B.W. (1981) Inversions between ribosomal RNA genes of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **78**, 7069–7072.
Hormozdiari,F. *et al.* (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.*, **19**, 1270–1278.
Hormozdiari,F. *et al.* (2010) Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics*, **26**, i350–i357.
Hubner,A. and Hendrickson,W. (1997) A fusion promoter created by a new insertion sequence, IS1490, activates transcription of 2,4,5-trichlorophenoxyacetic acid catabolic genes in *Burkholderia cepacia* AC1100. *J. Bacteriol.*, **179**, 2717–2723.
Huddleston,J. *et al.* (2014) Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.*, **24**, 688–696.
Hughes,D. (2000) Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biol.*, **1**, REVIEWS0006.
Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
Jasin,M. and Schimmel,P. (1984) Deletion of an essential gene in *Escherichia coli* by site-specific recombination with linear DNA fragments. *J. Bacteriol.*, **159**, 783–786.
Jiang,Y. *et al.* (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, **28**, 2576–2583.
Johnson,R.C. (1991) Mechanism of site-specific DNA inversion in bacteria. *Curr. Opin. Genet. Dev.*, **1**, 404–411.
Katapadi,V.K. *et al.* (2012) Potential G-quadruplex formation at breakpoint regions of chromosomal translocations in cancer may explain their fragility. *Genomics*, **100**, 72–80.
Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.
Kresse,A.U. *et al.* (2003) Impact of large chromosomal inversions on the adaptation and evolution of *Pseudomonas aeruginosa* chronically colonizing cystic fibrosis lungs. *Mol. Microbiol.*, **47**, 145–158.
Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.
Liang,Y. *et al.* (2010) Genome rearrangements of completely sequenced strains of *Yersinia pestis*. *J. Clin. Microbiol.*, **48**, 1619–1623.
Lim,K. *et al.* (2012) Large variations in bacterial ribosomal RNA genes. *Mol. Biol. Evol.*, **29**, 2937–2948.
Lu,C.L. *et al.* (2006) Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. *BMC Bioinformatics*, **7**, 295.
Mackiewicz,P. *et al.* (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes. *Genome Biol.*, **2**, INTERACTIONS1004.
Mahillon,J. and Chandler,M. (1998) Insertion sequences. *Microbiol. Mol. Biol. Rev.*, **62**, 725–774.
Marguet,P. *et al.* (2007) Biology by design: reduction and synthesis of cellular components and behaviour. *J. R. Soc. Interface*, **4**, 607–623.
Marttinen,P. *et al.* (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.*, **40**, e6.
Mazumder,R. *et al.* (2001) GeneOrder: comparing the order of genes in small genomes. *Bioinformatics*, **17**, 162–166.
Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**, S13–S20.
Miesel,L. *et al.* (1994) Construction of chromosomal rearrangements in *Salmonella* by transduction: inversions of non-permissive segments are not lethal. *Genetics*, **137**, 919–932.
Miller,J.C. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
Morrow,J.D. and Cooper,V.S. (2012) Evolutionary effects of translocations in bacterial genomes. *Genome Biol. Evol.*, **4**, 1256–1262.
Naas,T. *et al.* (1995) Dynamics of IS-related genetic rearrangements in resting *Escherichia coli* K-12. *Mol. Biol. Evol.*, **12**, 198–207.
Nagarajan,S. *et al.* (2012) Functions of the duplicated hik31 operons in central metabolism and responses to light, dark, and carbon sources in *Synechocystis sp.* strain PCC 6803. *J. Bacteriol.*, **194**, 448–459.
Nelson,K.E. *et al.* (1999) Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of Thermotoga maritime. *Nature*, **399**, 323–329.
Nogami,T. *et al.* (1985) Construction of a series of ompF-ompC chimeric genes by *in vivo* homologous recombination in *Escherichia coli* and characterization of the translational products. *J. Bacteriol.*, **164**, 797–801.
Ochman,H. *et al.* (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature*, **405**, 299–304.
Okinaka,R.T. *et al.* (2011) An attenuated strain of *Bacillus anthracis* (CDC 684) has a large chromosomal inversion and altered growth kinetics. *BMC Genomics*, **12**, 477.
Oliver,A. *et al.* (2002) The mismatch repair system (mutS, mutL and uvrD genes) in *Pseudomonas aeruginosa*: molecular characterization of naturally occurring mutants. *Mol. Microbiol.*, **43**, 1641–1650.
Quail,M.A. *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
Rasko,D.A. *et al.* (2011) Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.*, **365**, 709–717.
Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
Reams,A.B. and Neidle,E.L. (2004) Selection for gene clustering by tandem duplication. *Annu. Rev. Microbiol.*, **58**, 119–142.
Rebollo,J.E. *et al.* (1988) Detection and possible role of two large nondivisible zones on the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA*, **85**, 9391–9395.
Rocha,E.P. (2008) The organization of the bacterial genome. *Annu. Rev. Genet.*, **42**, 211–233.
Roth,J. *et al.* (1996) Rearrangements of the bacterial chromosome: formation and applications. In: Neidhardt,F.C. (ed.) *Escherichia coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, DC, pp. 2256–2276.
Segall,A. *et al.* (1988) Rearrangement of the bacterial chromosome: forbidden inversions. *Science*, **241**, 1314–1318.

Sindi,S. *et al.* (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics*, **25**, i222–i230.

Skovgaard,O. *et al.* (2011) Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.*, **21**, 1388–1393.

Spencer-Smith,R. *et al.* (2012) Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*. *PLoS One*, **7**, e46023.

Srivatsan,A. *et al.* (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.*, **4**, e1000139.

Sun,S. *et al.* (2012) Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS One*, **7**, e42639.

Tatusova,T. *et al.* (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.

Teague,B. *et al.* (2010) High-resolution human genome structure by single-molecule analysis. *Proc. Natl Acad. Sci. USA*, **107**, 10848–10853.

Treangen,T.J. *et al.* (2009) Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol. Rev.*, **33**, 539–571.

Wang,E.A. *et al.* (1982) Tandem duplication and multiple functions of a receptor gene in bacterial chemotaxis. *J. Biol. Chem.*, **257**, 4673–4676.

Weischenfeldt,J. *et al.* (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, **14**, 125–138.

Wu,T.H. and Marinus,M.G. (1999) Deletion mutation analysis of the mutS gene in *Escherichia coli*. *J. Biol. Chem.*, **274**, 5948–5952.

Xing,J. *et al.* (2009) Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.*, **19**, 1516–1526.

Ye,K. *et al.* (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.

Zeitouni,B. *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.