REVIEW

From next-generation resequencing reads to a high-quality variant data set

SP Pfeifer[1,2,3]

Sequencing has revolutionized biology by permitting the analysis of genomic variation at an unprecedented resolution. High-throughput sequencing is fast and inexpensive, making it accessible for a wide range of research topics. However, the produced data contain subtle but complex types of errors, biases and uncertainties that impose several statistical and computational challenges to the reliable detection of variants. To tap the full potential of high-throughput sequencing, a thorough understanding of the data produced as well as the available methodologies is required. Here, I review several commonly used methods for generating and processing next-generation resequencing data, discuss the influence of errors and biases together with their resulting implications for downstream analyses and provide general guidelines and recommendations for producing high-quality single-nucleotide polymorphism data sets from raw reads by highlighting several sophisticated reference-based methods representing the current state of the art.
Heredity (2017) **118**, 111–124; doi:10.1038/hdy.2016.102; published online 19 October 2016

INTRODUCTION

Sequencing has entered the scientific zeitgeist and the demand for novel sequences has never been greater, with applications spanning comparative genomics, clinical diagnostics and metagenomics, as well as agricultural, evolutionary, forensic and medical genetic studies. Several hundred species have already been completely sequenced (among them reference genomes for human and numerous major model organisms) (Pagani *et al.*, 2012), shifting the focus of many scientific studies toward resequencing analyses in order to identify and catalog genetic variation in different individuals of a population. These population-scale studies permit insights into natural variation, inference on the demographic and selective history of a population, and variants identified in phenotypically distinct individuals can be used to dissect the relationship between genotype and phenotype.

Until 2005, Sanger sequencing was the dominant technology, but it was prohibitively expensive and time consuming to routinely perform sequencing on a scale required to reach the scientific goals of many modern research projects. For these reasons, several massively parallel, high-throughput 'next-generation' sequencing (NGS) technologies have since been developed, permitting the analyses of genomes and their variation by being hundreds of times faster and over a thousand times cheaper than traditional Sanger sequencing (Metzker, 2010). Whereas the major limiting factor in the Sanger sequencing era was the experimental production of sequence data, data generation using NGS platforms is straightforward, shifting the bottleneck to downstream analyses, with computational costs often surpassing those of data production (Mardis, 2010). In fact, a multitude of bioinformatic algorithms is necessary to efficiently analyze the generated data sets and to answer biologically relevant questions. Thereby, the large amount of data with shorter read lengths, higher per-base error rates and nonuniform coverage, together with platform-specific read error profiles and artifacts (Table 1), imposes several statistical and computational challenges in the reliable detection of variants from NGS data (Harismendy *et al.*, 2009).

The correct identification of variation in genomes from resequencing data strongly relies on both the precise alignment of sequenced reads to a reference genome and reliable, accurate variant calling to avoid errors produced by misaligned reads or sequencing issues. Despite carefully chosen analytical methods, there is often a considerable amount of uncertainty associated with the results (particularly for low coverage sequencing) that necessarily must be accounted for in downstream population genomic analyses (see O'Rawe *et al.*, 2015 for a detailed review of uncertainty in NGS data). Here, I review several commonly used reference-based methods for generating and processing population-scale next-generation resequencing data and provide a general guideline for producing high-quality variant data sets from raw reads, focusing on single-nucleotide polymorphisms (SNPs). It should be noted that the detection of structural variation from NGS data will not be discussed here, but is a topic that has been well reviewed recently (see, for example, Tattini *et al.*, 2015; Guan and Sung, 2016; Ye *et al.*, 2016). In the following sections, I will outline the different stages of a typical workflow in a next-generation resequencing study (Figure 1); namely, sequencing, read processing, alignment and genetic variant detection. I conclude with a discussion on the influence of errors and biases introduced in these steps in downstream data analyses.

PREREQUISITE: DATA GENERATION (SEQUENCING)

NGS protocols commonly start with the preparation of libraries by shearing the DNA (either randomly or systematically (for example,

[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland and [3]School of Life Sciences, Arizona State University, Tempe, AZ, USA
Correspondence: SP Pfeifer, Arizona State University, School of Life Sciences, Tempe, AZ, USA.
E-mail: susanne.pfeifer@asu.edu

Table 1 Characteristics of several commercially available NGS platforms

	Capillary	Next generation					
	Sanger	454		Illumina			Ion Torrent
Platform	3730xl	GS FLX+	GS Jr.[a]	HiSeq X Ten	HiSeq 2500	MiSeq[a]	PGM 318[a]
Template preparation	Plasmid/PCR	emPCR	emPCR	Solid phase	Solid phase	Solid phase	emPCR
Run time	~3 h	~1 Day	~10 h	~3 Days	~6 Days	~65 h	~4–7 h
Output/run	0.08 Mb	700 Mb	35 Mb	1.8 Tb	1 Tb	15 Gb	2 Gb
Read length	1 kb	1 kb	700 b	2×150 b	2×125 b	2×300 b	400 b
No. of reads/run	96 (standard) up to 384 (rare)	1 M	0.1 M	6 B	4 B	25 M	5.5 M
Error rate[b]	0.1–1%	~1%	~1%	~0.1%	~0.1%	~0.1%	~1%
Primary errors	Substitutions	Indels		Substitutions[c]			Indels
Advantages	• Long reads • High quality	• Long reads • Fast run time		• Highest throughput • Low per-base cost			• Unmodified nucleotides • No optical scanning necessary, and thus no photo damage • Fast run time
Limitations	• Low throughput • High costs	• High error rates in homopolymer regions • Low throughput • High costs • Cumbersome emPCR		• Short reads • Random dispersion of clusters can cause poor sequence quality • Underrepresentation of AT-rich and GC-rich regions			• High error rates in homopolymer regions • Cumbersome emPCR

Abbreviations: emPCR, emulsion PCR; NGS, next-generation sequencing.
Information based on company sources.
[a]Inexpensive, compact bench-top sequencers with faster turnover rates but lower sequencing capacity.
[b]Error rates between different platforms might not be exactly comparable.
[c]Biased toward A<->C and G<->T transversions.

restriction site associated DNA sequencing)) into relatively short sequence fragments to which platform-specific adapters, containing primer sites for sequencing, are ligated. Optionally, samples can be indexed by hybridization of additional sequencing primers, allowing multiple samples to be sequenced simultaneously, computationally separable by their barcode. These adapters are used to spatially distribute the fragments by immobilizing them onto a solid surface (usually either a micron-scale bead or a solid planar surface). In contrast to traditional Sanger sequencing that requires a bacterial cloning step, the fragments are amplified *in vitro* by PCR (usually either emulsion PCR (Dressman *et al.*, 2003) used by 454 and Ion Torrent or bridge PCR (Adessi *et al.*, 2000; Fedurco *et al.*, 2006) used by Illumina).

This step results in the presence of clusters (either ordered or randomly dispersed) consisting of multiple identical copies of a DNA sequence of interest, flanked by universal adapter sequences. The presence of clusters ensures that the sequencing reaction produces a signal sufficiently strong to be detected by an optical system. Amplification is followed by a series of repeated steps, alternating between enzymatic manipulation and image-based data acquisition. Specifically, spatially separated clusters are simultaneously decoded by applying a single reagent volume to the array, whereby a cyclic enzymatic process interrogates the identity of one position or type of nucleotide at a time for all clusters in parallel. This process is coupled either to the measurement of H^+ ion release (Ion Torrent), the production of light (454) or the incorporation of a fluorescent group (Illumina), the latter two of which are directly detectable using a charge-coupled device. Thus, after multiple cycles, the continuous sequence from each cluster can be obtained.

Platform-specific software is then used to translate these signals into base calls, represented in FASTQ format (Cock *et al.*, 2010) and associated with ASCII-encoded PHRED-like quality scores (that is, statistical measures of call certainty provided by the logarithm of the expected error probability of the base call: $Q_{PHRED} = -10 \times \log_{10} P(error)$) (Ewing and Green, 1998). These quality scores can be used together with the sequence information for subsequent analyses. Depending on the platform, sequence information can be obtained from one end of the fragment (that is, single-end sequencing) or from both ends of either a linear fragment (that is, paired-end sequencing; reads are usually separated by 300–500 bp) or a previously circularized fragment (that is, mate pair sequencing; reads are usually separated by 1.5–20 kb) (Mardis, 2011).

Despite these commonalities of different NGS protocols, platforms vary greatly in their specific characteristics (for example, DNA input requirement, template preparation, throughput and average read length). In fact, each platform is associated with unique biases introduced during library construction, amplification and sequencing, as well as systematic errors, resulting in average per-base error rates (and the underlying reasons for the error) that differ strongly between methods (Table 1). These biases can originate from the experimental sample preparation, where, in addition to unintentional contamination, polymerases frequently introduce errors in fragments because of imperfect *in vitro* amplification (Dohm *et al.*, 2008). Polymerases may also vary in speed and often reach different read lengths because of photodamage, thus increasing background noise. Errors can also be introduced during the sequencing step, where certain DNA sequence characteristics, such as long homopolymer runs or extreme GC-contents, increase error rates in reads (Laehnemann *et al.*, 2016).
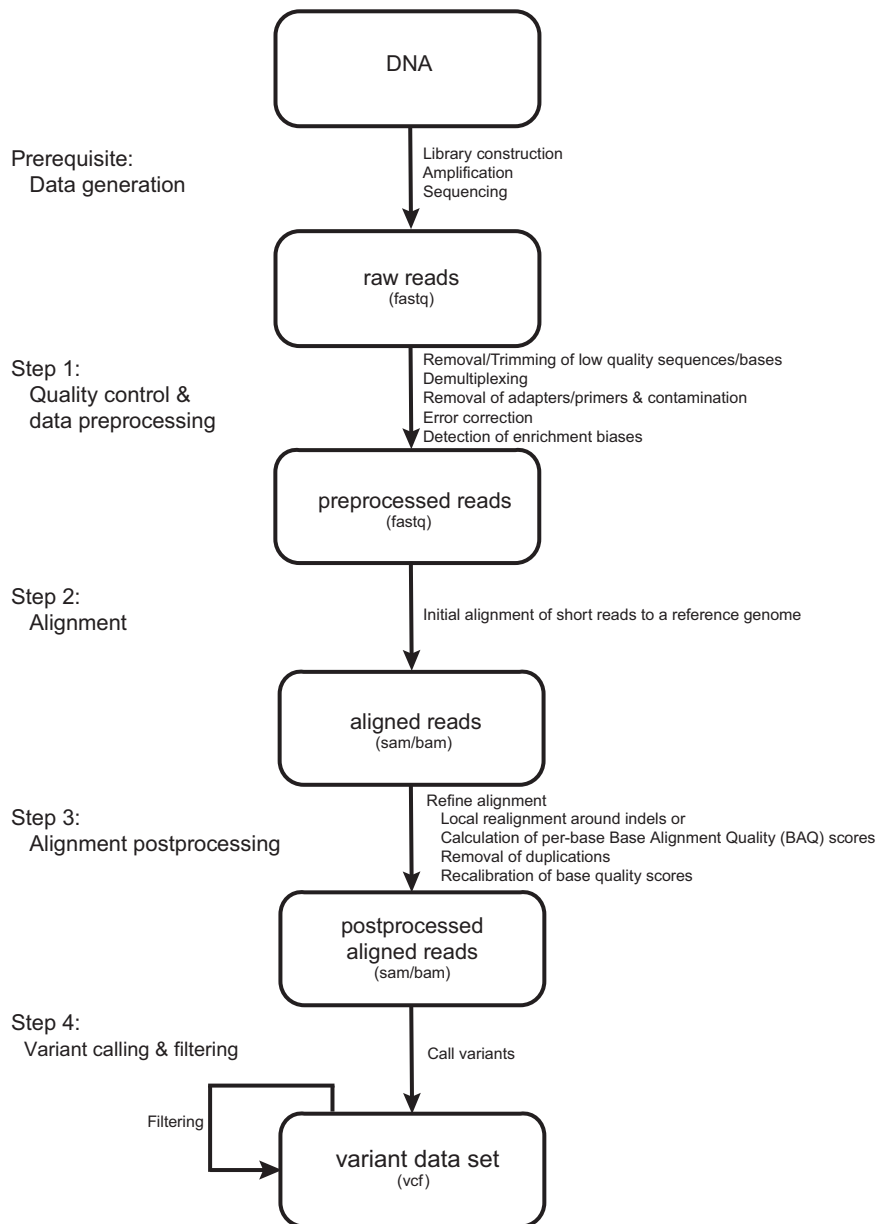
Figure 1 Steps in a typical next-generation resequencing workflow. *De facto* standard file formats are given in parentheses.

For most platforms, errors increase towards the end of the read because of reductions in signal intensity, caused by decreased enzyme activity (Kircher *et al.*, 2009). Incomplete read extension or nonreversible termination desynchronizes clusters of the same template (referred to as dephasing), elevating noise further (Kircher and Kelso, 2010). Dephasing not only results in base calling errors, but it also limits achievable read lengths. Random dispersion of clusters onto a solid surface coupled with limited sensor resolution is an additional source of error, introducing false reads when signals from nearby clusters interfere with the readout (Kircher and Kelso, 2010). Furthermore, chemical crystals, dust and other small particles can be mistaken as clusters in the images, yielding low-quality base calls.

Given these different platform characteristics, a detailed knowledge of the advantages and limitations of each method can help inform decisions about the optimal sequencing technology for a particular project. In general, large amounts of short reads are most appropriate for whole-genome or targeted resequencing studies (as discussed here), chromatin immunoprecipitation with subsequent sequencing and expression analyses. In contrast, longer reads are better suited for an initial characterization of the genome (that is, *de novo* assembly) as well as the study of alternative splicing. An overview of the main technical specifications of several current commercially available NGS platforms is provided in Table 1.

STEP 1: QUALITY CONTROL AND DATA PREPROCESSING
Together with the DNA sequences of interest, raw read data often contain biases (for example, through systematic effects such as Poisson sampling) and complex artifacts arising from the experimental and sequencing steps (Aird *et al.*, 2011; Nakamura *et al.*, 2011; Allhoff *et al.*, 2013). These biases and artifacts strongly interfere with accurate

read alignments that in turn influence variant calling and genotyping. Thus, to increase the reliability of downstream analyses and to simultaneously decrease the required computational resources (that is, RAM, disk space and execution time), raw read data should be preprocessed.

Quality assessment

In order to identify potential problems in the experimental setup and to ensure that the correct samples have been sequenced with minimal contamination and to sufficient coverage, summary statistics assessing the overall quality of the data set—such as nucleotide and quality score distributions, as well as sequence characteristics including GC-content, levels of sequence ambiguity and PCR duplication—should be generated from the raw read data before any analyses. Popular tools for performing this initial quality check include FastQC, htSeqTools (Planet *et al.*, 2012), Kraken (Davis *et al.*, 2013), NGSQC (Dai *et al.*, 2010), PRINSEQ (Schmieder and Edwards, 2011b), qrqc and SAMStat (Lassmann *et al.*, 2011). Subsequently, the results can aid the selection of quality control parameters and thresholds in read processing to circumvent potential problems in the later stages of data analyses.

Potential issue 1: low-quality data. High-quality sequence data are characterized by a majority of reads exhibiting high PHRED-like quality scores along their entire length (Figure 2a). However, examination of overall sequence quality scores frequently indicates some proportion of raw read data that contains sequences with universally low-quality scores. A small number of these low-quality reads might be caused by air bubbles, spot-specific signal noise or problems with the readout during sequencing (for example, imaging of reads on the edge of the flow cell; Kircher *et al.*, 2011) and these reads should be excluded from subsequent analyses. In contrast, a substantial number of low-quality sequences might be indicative of a more systematic problem with the run. Another indicator of a general quality loss is a large proportion of positions without base calls (that is, bases that cannot be accurately called and are indicated as N's in the reads).

For many sequencing platforms, the quality of the read decreases as the run progresses (Figure 2b) because both signal decay and dephasing elevate the background noise (Kircher *et al.*, 2009; Kircher and Kelso, 2010). Two different strategies can be employed to handle these low-quality base calls: error correction or removal of low-quality read regions. First, assuming that errors are both infrequent and random, erroneous base calls in low-quality reads may be corrected by superimposing multiple reads and modifying low-frequency patterns by calling a high-frequency consensus sequence. Sophisticated error-correction methods include:

(a) *k*-spectrum-based approaches that decompose the reads into a set of all *k*-mer fragments (for example, BFC (Li, 2015), BLESS (Heo *et al.*, 2014), CUDA-EC (Shi *et al.*, 2010), DecGPU (Liu *et al.*, 2011), Hammer (Medvedev *et al.*, 2011), Lighter (Song *et al.*, 2014), Musket (Liu *et al.*, 2013b), Quake (Kelley *et al.*, 2010), Reptile (Yang *et al.*, 2010), SOAPec (Li *et al.*, 2010) and Trowel (Lim *et al.*, 2014));

(b) suffix-tree/array-based methods (for example, Fiona (Schulz *et al.*, 2014), HiTEC (Ilie *et al.*, 2011), Hybrid-SHREC (Salmela, 2010), RACER (Ilie and Molnar, 2013) and SHREC (Schröder *et al.*, 2009)); and

(c) multiple sequence alignment methods (for example, Coral (Salmela and Schröder, 2011) and ECHO (Kao *et al.*, 2011)).

See Yang *et al.* (2012) and Laehnemann *et al.* (2016) for reviews of error-correction methods and their limitations, as well as benchmarking data.

However, error-correction approaches require a high coverage (see, for example, Kelley *et al.*, 2010; Gnerre *et al.*, 2011) and are thus not suitable for low to medium coverage studies of most nonmodel organisms. In addition, the correction method necessitates a uniform read distribution, rendering it infeasible for several types of studies (for example, transcriptomics or metagenomics). Therefore, second, low-quality read regions can be removed by estimating error rates and by identifying suitable thresholds, enabling the retention of the longest high-quality sequence possible (referred to as read trimming), inevitably leading to a loss of information. A large variety of techniques have been proposed to trim low-quality regions from NGS data (Table 2), most of which can be classified into (1) window-based methods that locally scan a read using either non-overlapping or sliding windows and (2) running-sum methods that globally scan a read using a cumulative quality score to determine the best position for trimming.

Because of air bubbles passing through the flow cell during sequencing, reads might exhibit low-quality scores at their 5′-end that recover to high-quality calls later in the run (Figure 2c). Under this circumstance, error correction can be applied but trimming is generally not advisable. An alternative to error correction is to mask the low-quality sequence in the read before mapping it to a reference genome. However, in the case of multiple samples having been sequenced simultaneously, which are distinguishable from one another by short, unique barcodes, these reads can cause more severe issues as it becomes necessary to demultiplex the read data in the presence of potential sequencing errors in these barcodes. To circumvent this problem, several methods have been developed for designing barcodes with an error-correction capability that aid correct sample identification in the presence of sequence alterations introduced during synthesis, primer ligation, amplification or sequencing (Buschmann and Bystrykh, 2013). Popular error-correcting techniques include methods based on false discovery rate statistics (see, for example, Buschmann and Bystrykh, 2013) as well as adaptations of both Hamming codes (see, for example, Hamady *et al.*, 2008; Bystrykh, 2012) and Levenshtein codes (for example, implemented in the software Sequence- Levenshtein (Buschmann and Bystrykh, 2013) and TagGD (Costea *et al.*, 2013)). Algorithms based on the latter are capable of correcting not only substitution errors but also insertion and deletion (indel) errors that is particularly important for sequencing technologies where indels are the main source of error (that is, 454 and Ion Torrent).

Potential issue 2: presence of adapter sequences or contaminants. Reads longer than the targeted DNA fragments result in the sequencing of partial or complete adapter or primer sequences, present at either the 3′- or 5′-end depending on the protocol used to build the library. These adapter/primer sequences will lead to mismatches between the read and the reference sequence, either entirely inhibiting the alignment or resulting in false positive variant calls. This is particularly problematic when the adapter occurs at the 5′-end, as many commonly used aligners require a high similarity within this region —an issue that becomes more severe the shorter the fragment (for example, for ancient DNA or forensic samples). As a result, regions of nongenomic origin should be removed before mapping the reads to a reference genome. However, the identification of known library sequences is often a nontrivial task, complicated by errors occurring during the PCR and sequencing steps that can alternate

known adapter and primer sequences. Furthermore, genomic contamination caused by nontarget DNA (both within species and between species, as well as control DNA utilized in the experiment, such as PhiX used in Illumina sequencing) introduced during the experimental preparation should be eliminated before downstream analyses.
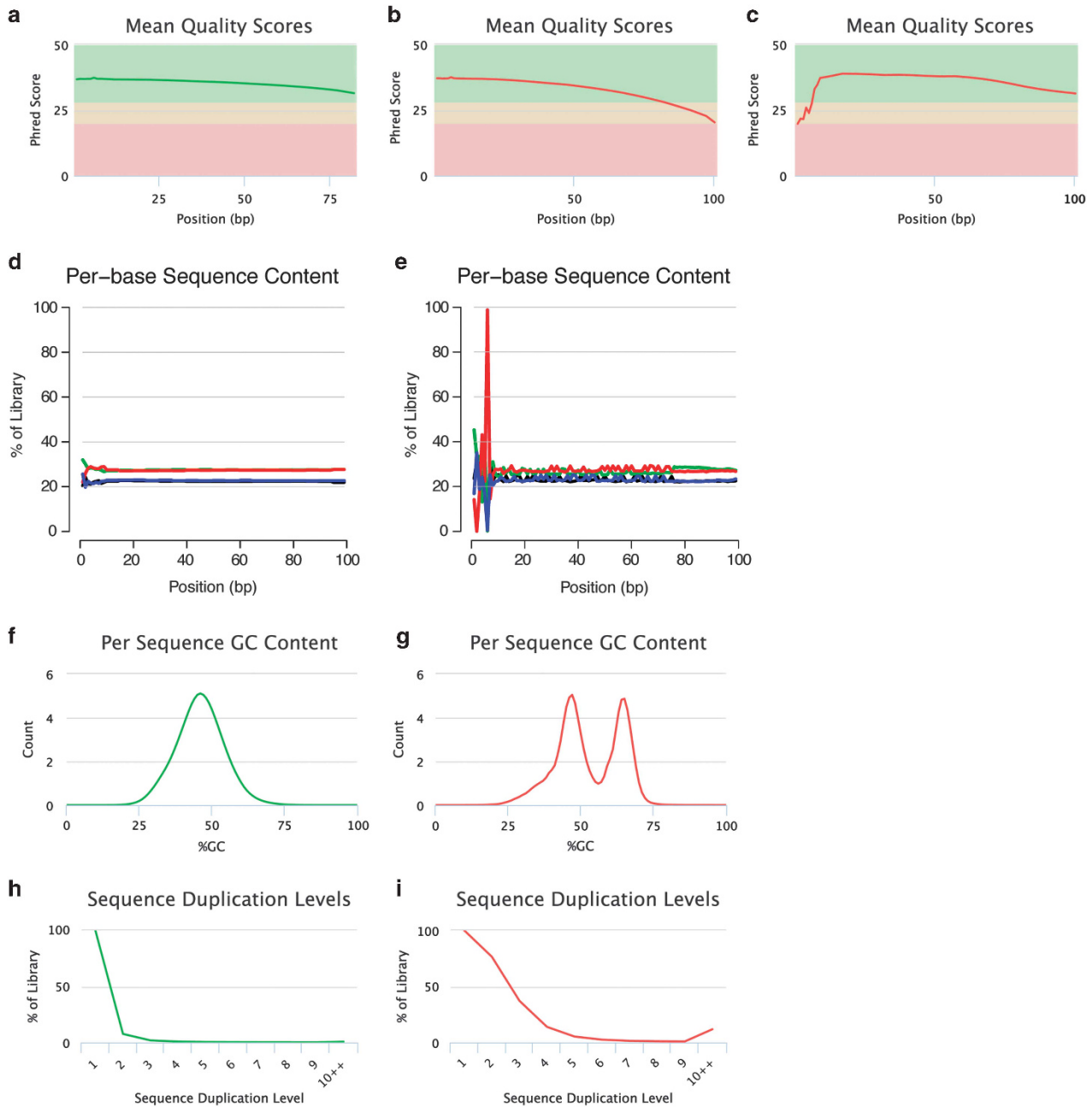

Figure 2 Read quality assessment. *Per-base sequence quality* plots indicate the mean quality scores for each nucleotide position in all reads. Background colors highlight the quality of the call (that is, green: high quality; yellow: reasonable quality; red: low quality). Examples: (**a**) Reads exhibit high base quality scores at each position. (**b**) Sequenced nucleotides initially exhibit high-quality scores but the per-base read quality decreases with increasing read length, reaching lower quality values toward the read end, necessitating read trimming. (**c**) Initial low per-base quality that recovers to a high quality later in the run. Under this circumstance, error correction can be applied but trimming is generally not advisable. *Per-base sequence content* plots indicate the proportion of each nucleotide for each read position. Examples: (**d**) A random library with little difference of base composition (colors indicate different nucleotides: green: A; blue: C; black: G; red: T) between single read positions. (**e**) Imbalance of different bases, potentially caused by overrepresented sequences (for example, adapters). *Per sequence GC-content* plots indicate the observed GC-content of all reads. Examples: (**f**) The GC-content of the reads is normally distributed with a peak that corresponds to the overall genomic GC-content of the studied species. (**g**) The bimodal shape of the distribution of the reads' GC-content suggests that the sequenced library may have been contaminated or that adapter sequences may still be present. *Sequence duplication* plots indicate the level of duplication among all reads in the library (reads with more than 10 duplicates are binned). Examples: (**h**) The low level of sequence duplication suggests that a diverse library has been sequenced with a high coverage of target sequence. (**i**) A high level of sequence duplications often suggests either a technical artifact (for example, because of PCR overamplification) or biological duplications. All examples were generated using FastQC and plotted using MultiQC (Ewels *et al.*, 2016).

Table 2 Characteristics of several popular tools for preprocessing NGS reads

Software	Ability of the software to:			Reference
	Handle multiple adapter sequences	Trim low-quality bases	Demultiplex barcodes	
AdapterRemoval	−	+	−	Lindgreen (2012)
AlienTrimmer	+	+	−	Criscuolo and Brisse (2013)
Btrim	−	+	+	Kong (2011)
CONDETRI	−	+	−	Smeds and Künstner (2011)
Cutadapt	+	+	−	NA
EA-Utils	−	+	+	NA
ERNE-FILTER	−	+	−	Del Fabbro et al. (2013)
FASTX-Toolkit	−	−	+	NA
Flexbar	+	+	+	Dodt et al. (2012)
Kraken	+	+	+	Davis et al. (2013)
NGSQC	+	+	−	Dai et al. (2010)
ngsShoRT	+	+	−	Chen et al. (2014)
PEAT	+	−	−	Li et al. (2015)
PRINSEQ	−	+	−	Schmieder and Edwards (2011b)
QC-Chain	+	+	−	Zhou et al. (2013)
QcReads	+	+	−	Ma et al. (2013)
Reaper	+	+	+	NA
SeqTrim	+	+	−	Falgueras et al. (2010)
Sickle	−	+	−	NA
Skewer	+	+	+	Jiang et al. (2014)
TrimGalore!	−	+	−	NA
Trimmomatic	+	+	−	Bolger et al. (2014)

Abbreviations: NA, no publication available; NGS, next-generation sequencing.
Popularity was assessed by the number of citations of the software.

There are several ways to visually detect the presence of adapter sequences or potential contaminations. First, the per-base sequence content can be examined. As the relative proportion of each nucleotide at each read position should reflect the overall sequence composition of the genome, there should not be any significant difference in base composition along the read in a random library (Figure 2d). Imbalances of different bases could be caused by overrepresented sequences such as adapters (Figure 2e). Second, the per-sequence GC-content should be normally distributed with a peak corresponding to the overall genomic GC-content of the studied species (Figure 2f). Any deviation from the normal distribution suggests that the sequenced library might either have been contaminated or adapter sequences might still be present (Figure 2g). Unusually shaped distributions with sharp peaks often suggest the presence of adapter sequences, whereas broader or multiple peaks indicate a contamination with a different species. In contrast, a shift in the distribution can suggest a systematic bias. In addition, adapter sequences can be automatically detected and removed from genomic data sets (often simultaneously with a quality trimming of the data) using one of the software packages described in Table 2. Different mitigation techniques also exist for the removal of genomic contaminants, and some of the most popular ones include ContEst (Cibulskis et al., 2011), DeconSeq (Schmieder and Edwards, 2011a), QC-Chain (Zhou et al., 2013) as well as the set of methods developed by Jun et al., 2012 used in the 1000 Genomes Project.

Potential issue 3: enrichment bias. In a diverse library, low levels of sequence duplication suggest that the library has been sequenced with a high coverage of target sequence (Figure 2h). In contrast, high levels of sequence duplications often arise from either a technical artifact (for example, because of PCR overamplification) or biological duplications (Figure 2i). Technical duplicates should be removed during alignment postprocessing as they manifest themselves as high read coverage support, potentially leading to erroneous variant calls.

STEP 2: ALIGNMENT

Fundamentally, the most important step, upon which any subsequent next-generation resequencing data analysis is based, is the accurate alignment of the generated reads to a reference genome. NGS technologies have the capacity to yield billions of reads per experiment—an amount of sequencing data orders of magnitude greater than those produced by capillary-based techniques, rendering alignment tools developed in the Sanger sequencing era insufficient to analyze the generated data. Importantly however, NGS reads are also much shorter and less accurate than the reads obtained by traditional Sanger sequencing and, as a consequence, experimentally induced artifacts, sequencing errors as well as true polymorphisms have a larger influence on the alignment.

Over the past decade, more than a hundred sophisticated alignment algorithms have been specifically designed to handle the computational challenges imposed by modern NGS platforms (see Table 3 for a summary of the characteristics of several popular NGS aligners; a technical overview is provided by Reinert et al., 2015). These algorithms are optimized for their efficiency (that is, speed), scalability (that is, storage space) and accuracy (for example, by taking specific technological biases of the different platforms and protocols into account). At the time NGS platforms first entered the market, reads were substantially shorter (~25 bp) and, thus, many of the earlier proposed NGS mappers used ungapped alignments to avoid the computational costs associated with allowing gaps (Li and Homer, 2010). However, even when mapping reads to the correct position in

Table 3 Characteristics of several popular open-source/binary NGS aligners

Software	Sequencing platform	Ability to perform gapped alignment	Quality awareness	Ability to align PE reads	Reference
BFAST	I,4	+	−	+	Homer et al. (2009)
Bowtie	I,4,Sa	−	+	+	Langmead et al. (2009)
Bowtie 2	I,4,Ion	+	+	+	Langmead and Salzberg (2012)
BWA	I,4,Sa	+	+	+	Li and Durbin (2009)
CloudBurst	non-specific	+	−	−	Schatz (2009)
GSNAP	I,4,Sa,Ion	+	−	+	Wu and Nacu (2010)
MAQ	I	−	+	+	Li et al. (2008)
MOSAIK	I,4,Sa,Ion	+	+	+	NA
mrFAST	I	−	+	+	Alkan et al. (2009)
mrsFAST	I	−	+	+	Hach et al. (2010)
NextGenMap	I,4,Ion	+	−	+	Sedlazeck et al. (2013)
PASS	I,4	+	+	+	Campagna et al. (2009)
RazerS	I,4	+	−	+	Weese et al. (2009)
segemehl	I,4,Sa,Ion	+	−	+	Hoffmann et al. (2009)
SHRiMP	I,4	+	−	+	Rumble et al. (2009)
SHRiMP 2	I,4	−	+	+	David et al. (2011)
SOAP2	I	+	−	+	Li et al. (2009b)
Stampy	I	+	+	+	Lunter and Goodson (2011)

Abbreviations: I, Illumina; Ion, Ion Torrent; NA, no publication available; NGS, next-generation sequencing; PE, paired end; Sa, ABI Sanger; 4, Roche 454.
Information obtained from http://www.ebi.ac.uk/~nf/hts_mappers/ (last accessed August 2016). Popularity was assessed by the number of citations of the software.

the reference, ungapped alignments can produce false positive SNP calls when multiple reads support consecutive mismatches at a locus of a true indel polymorphism (Li and Homer, 2010). These false positive SNPs are difficult to detect, even with sophisticated filtering strategies (Li and Homer, 2010). To circumvent this issue, modern NGS mappers implement gapped alignments. In addition, many algorithms enhance alignments in error-prone, low-quality regions by integrating base quality scores emitted by the sequencing platform into their algorithms (Smith et al., 2008). Resulting alignments are usually stored in the sequence alignment/map (SAM) format, or its binary, compressed version (BAM format) (Li et al., 2009b), containing information about the location, orientation and quality of each read alignment. Tools have been specifically designed to manipulate SAM/BAM files (for example, to quickly sort, merge or retrieve alignments), including the widely used SAMtools (Li et al., 2009b) and Picard.

Several benchmarking studies have empirically compared short read alignment methods with respect to various metrics (that is, runtime, sensitivity and accuracy) using both simulated and real data sets from different organisms (Holtgrewe et al., 2011; Ruffalo et al., 2011; Fonseca et al., 2012; Lindner and Friedel, 2012; Schbath et al., 2012; Hatem et al., 2013; Caboche et al., 2014; Shang et al., 2014; Highnam et al., 2015). These analyses demonstrated that results depend strongly on the properties of the input data, and thus there is no single method best suited for all scenarios. In fact, most tools are highly configurable, making them flexible to accommodate different research applications, but parameter choice has been shown to have a large influence on mapper performance (Lindner and Friedel, 2012). Currently, there are no gold-standard test data sets available for different sequencing technologies or applications, and determining the ideal parameter settings is often nontrivial, requiring an in-depth understanding of both the data and the alignment algorithm. Therefore, researchers working with NGS data face the challenge of choosing a method that best suits their specific requirements and research goals. Recently developed dedicated software (for example, Teaser; Smolka et al., 2015) can lend guidance when choosing an aligner or the parameters suitable for a particular data set and application. In addition, there are certain best practices that should be encouraged to prevent problems in downstream analyses.

Potential issue 1: alignment in low-complexity or repetitive regions
A particularly challenging task is the alignment of a short read originating from a repetitive or low-complexity genomic region that is longer than the read itself. In this case, the read often maps equally well to multiple locations in the genome, resulting in an ambiguous alignment that potentially leads to biases and errors in the variant calling. One way to aid read alignment in these regions is the utilization of paired-end or mate pair reads. They provide information about the relative position and orientation of a pair of reads in the genome, allowing the approximately known physical chromosomal distance between the read pair to be used to increase both the sensitivity and specificity of an alignment (that is, a repetitive read that cannot be confidently mapped on its own can often be placed using the information provided by its partner originating from a nonrepetitive region). Although the incorporation of mate pair or paired-end information aids the alignment, mapping to highly repetitive genomes continues to present a serious challenge (not at last because repetitive regions are often poorly resolved in the reference assembly). Focusing on uniquely mapped reads addresses issues in the SNP discovery, but biologically important variants might be missed and correctly resolving copy number variations remains difficult. Longer reads with varying insert sizes, sequenced using multiple technologies, can help to overcome some of the problems posed by repeats. However, the need for more sophisticated methods in both the de novo assembly and alignment of repetitive reads is ongoing (see Treangen and Salzberg, 2012 for an in-depth discussion on the computational challenges and potential solutions of handling repetitive DNA in NGS data analyses).

Potential issue 2: alignment in the presence of contamination or missing information
The goal of any alignment algorithm is to map individual reads to the position in the reference genome from which they most likely

originated. Unfortunately, even the highest quality reference assemblies have gaps and regions of uncertainty—missing sequence information that will inevitably result in off-target alignments. In addition, DNA sequences of interest are often contaminated (for example, with human herpesvirus 4 type 1, also known as Epstein–Barr virus—a DNA virus frequently used to immortalize cell lines). As a consequence, including a decoy genome, which enables the absorption of reads that do not originate from the reference, often improves alignment accuracy. The usage of a decoy genome will not only reduce false positive variant calls, but will also speed up the alignment by eliminating long computation cycles where the algorithm tries to identify an ideal position for a read in a reference genome from which it did not originate.

Potential issue 3: alignment in species with high mutation rates

Mapping reads in species with high mutation rates or reads originated from nonmodel organisms, for which only low-quality draft reference assemblies are available, pose challenges. For species with a large genetic diversity among individuals, the genome of a sequenced individual might be considerably different from the available reference assembly. As a consequence, reads containing multiple alternative alleles might not be mapped correctly, causing an underrepresentation of particular haplotypes. To tackle the problem of biasing the results to a specific, arbitrarily chosen genome, previously cataloged information about polymorphisms (for example, variant information obtained from a previous study or freely available variant data from public databases such as dbSNP; Sherry et al., 2001) can be integrated to enable a simultaneous alignment against multiple genomes (referred to as SNP-aware or SNP-tolerant alignment; see Schneeberger et al., 2009; Wu and Nacu, 2010; Hach et al., 2014), allowing minor alleles to be considered as matches rather than mismatches during mapping.

Quality assessment

Alignment quality can be assessed visually (in a small target region using an alignment viewer such as BamView (Carver et al., 2010; Carver et al., 2013), Gap5 (Bonfield and Whitwham, 2010), the Broad Institute's Integrative Genomics Viewer (IGV) (Robinson et al., 2011), LookSeq (Manske and Kwiatkowski, 2009a), MapView (Boa et al., 2009), SAMtools' Text Alignment Viewer (tview) (Li et al., 2009a) or Tablet (Milne et al., 2010)) as well as by using the information provided in the SAM/BAM file (see specifications in Li et al., 2009a for a format description). Thereby, mapping statistics will provide a first overview of the fraction of reads that were successfully mapped to the reference genome, and PHRED-scaled mapping quality scores indicate whether or not the mapping is likely to be correct (Figure 3a). In addition, the FLAG field in the SAM/BAM file (that is, a bitwise encoded set of information describing the alignment) relays important information such as whether the read is aligned properly (for example, in paired-end mapping, the second read in the pair should map on the same chromosome in the opposite strand direction), whether it passed all quality control checks or whether the read is likely either a PCR or optical duplicate (that is, the same fragment has been read twice). CIGAR strings, indicating which bases align with the reference (either match or mismatch), or are inserted/deleted compared with the reference sequence, as well as the edit distance to the reference can highlight regions with an unusual behavior (for example, a region where short reads map with many small insertions and deletions (Figure 3b) frequently hints toward a spurious alignment).

Another useful measurement of alignment quality is the number of perfect hits of a read in the reference sequence. In case of an ambiguous mapping where reads map equally well in multiple

locations in the genome, these reads might either be excluded or one alignment chosen at random for further analyses. However, both scenarios are associated with potential problems: in the first case, only uniquely mapped reads will be included in the data set, thus potentially missing biologically important variants, whereas systematic misalignments in the latter case may lead to the erroneous inference of polymorphisms.

STEP 3: ALIGNMENT POSTPROCESSING

Before variant calling, read alignments should be preprocessed to detect and correct spurious alignments in order to minimize artifacts in the downstream analyses.

Potential issue 1: high ambiguity in the local alignment

Because of the fact that alignment algorithms map reads individually to the reference genome, reads spanning insertions or deletions are often misaligned as most aligners have the tendency to introduce SNPs rather than structural variants in the alignments, as they occur more frequently in the genomes of most species. Thus, at positions of real but unidentified indels, alignment artifacts result in spurious variant calls (Homer and Nelson, 2010). One possibility to improve variant calling is multiple sequence alignment that locally realigns reads such that the number of mismatching bases is minimized across all reads. These realignment methods (popular tools include the Genome Analysis ToolKit (GATK) IndelRealigner (McKenna et al., 2010; DePristo et al., 2011; Van der Auwera et al., 2013), and the Short Read Micro re-Aligner (SRMA) (Homer and Nelson, 2010)) first identify suspicious-looking intervals that might be in need of realignment (for example, a site at which a cluster of mismatching nucleotides exists and at least one of the aligned reads contains an indel, or a site of a previously known indel; Figure 3c), before locally realigning these reads in order to obtain a more concise consensus alignment (Figure 3d). Unfortunately, current implementations are computationally intense, rendering them infeasible for high coverage sequencing (Ni and Stoneking, 2016). In addition, filtering SNPs around predicted indels is hindered by the challenge of correctly identifying indels initially. Another approach to avoid these artifacts is calculating per-base Base Alignment Quality scores (Li, 2011b) that decreases false positive calls in low complexity regions of the genome by down weighting base qualities with high ambiguity in the local alignment.

Potential issue 2: artificial high coverage read support

Overrepresentations of certain sequences, such as sequence duplications introduced during the PCR amplification step, substantially influence variant discovery by skewing coverage distributions. Duplications manifest as high coverage read support (Figure 3e), thereby often giving rise to false positive variant calls, as errors that occurred during the sample and library preparation have been propagated to all PCR duplicates. Because of the fact that PCR duplicates arise from the same DNA fragment, their positions in the reference (and, depending on the software, their sequence identity) can be used in shotgun sequencing to readily identify and either mark or entirely remove these duplicates (Figure 3f), retaining only the highest quality read (commonly used software for this task include SAMtools (Li et al., 2009b) and Picard). However, it is worth keeping in mind that this duplicate removal strategy is not perfect: it neither accounts for sequencing errors nor for biological duplications and nor for PCR duplicates that align to different positions in the genome. Furthermore, it cannot be applied to restriction site associated DNA sequencing or amplicon sequencing data, where reads begin and end
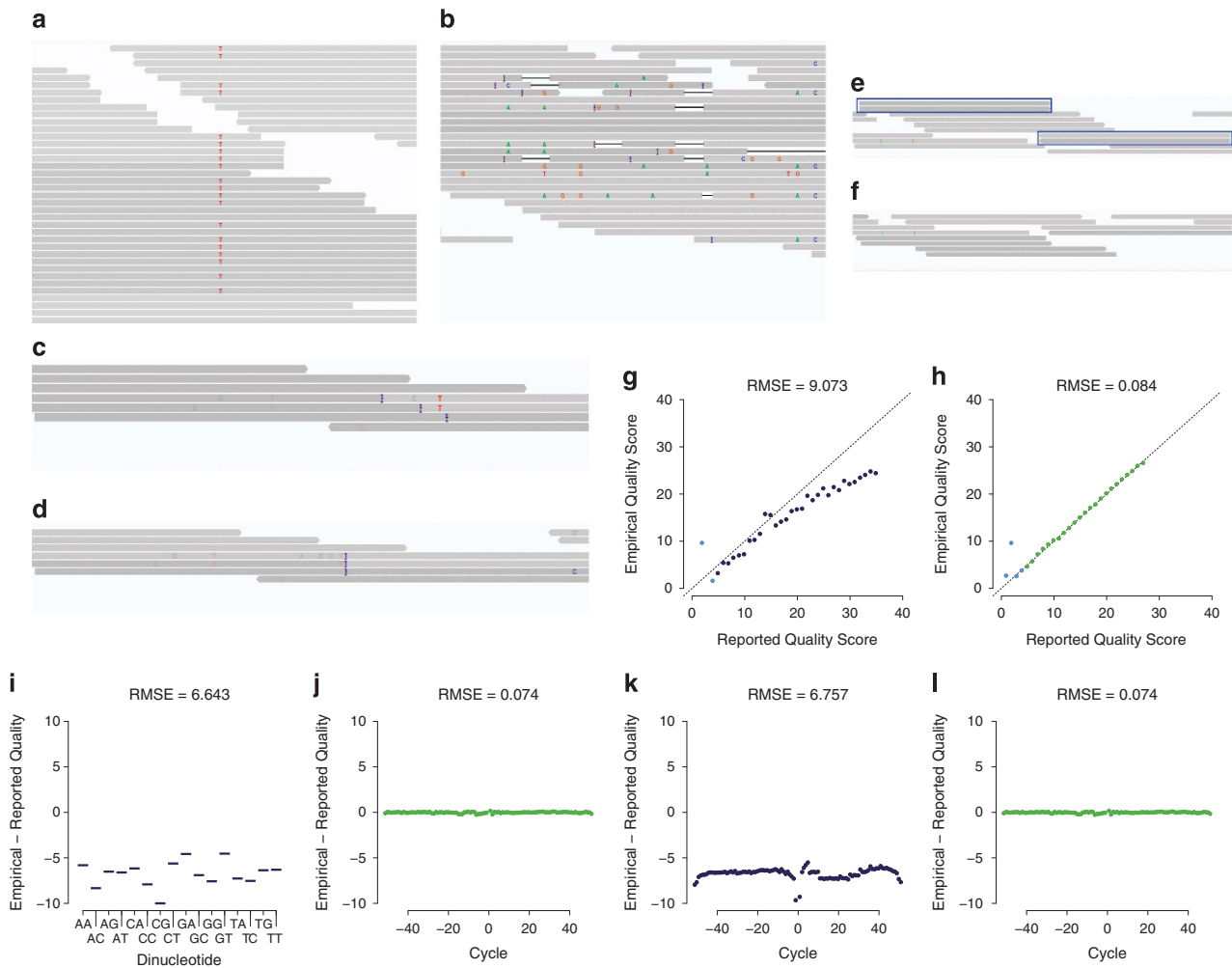
Figure 3 Alignment: quality assessment and postprocessing. *Quality assessment*: Examples: (**a**) A correctly aligned region (reads are shown as gray vertical bars with SNPs indicated as colored letters). (**b**) A spurious alignment where reads exhibit many small insertions (indicated as purple *I*s), deletions (shown as black horizontal lines) and SNPs. *Local realignment*: Examples: (**c**) Pre-realignment: suspicious-looking interval, covered by reads exhibiting both mismatching nucleotides and small indels at different positions that would benefit from a local realignment. (**d**) Post-realignment: reads were locally realigned such that the number of mismatching bases is minimized across all reads. *Duplicate removal*: Examples: (**e**) Pre-removal of duplications: duplicates (blue boxes) manifest themselves as high coverage read support. (**f**) Post-removal of duplications: no excess coverage because of identical duplicates. *Base quality score recalibration*: Examples: Reported versus empirical quality scores (**g**) before and (**h**) after recalibration. Empirical quality scores were calculated by PHRED-scaling the observed rate of mismatches with the reference genome. Bases with a quality value of < 5 (indicated in light blue) were ignored during the recalibration. Residual error for each of the 16 genomic dinucleotide contexts (for example, the AC contexts refers to a site in a read where the current nucleotide, a cytosine (C), is preceded by an adenine (A)) (**i**) before and (**j**) after recalibration. Residual error by machine cycle (with positive and negative cycle numbers given for the first and second read in a pair) (**k**) before and (**l**) after recalibration. Examples (**a**–**f**) were plotted using IGV (Robinson *et al.*, 2011). Examples (**g**–**l**) were plotted using GATK (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013).

at the same positions by design. In these cases, an alternative approach has recently been proposed by Tin *et al.* (2015) that makes use of degenerated bases in the adapter sequence to identify PCR duplicates.

Potential issue 3: poorly calibrated base quality scores
Most variant and genotype calling algorithms incorporate PHRED-scaled base quality scores into their probabilistic framework to enable an improved calling in low coverage regions and to decrease errors. Unfortunately, because of quality variations (caused by, for example, machine cycle and sequence context), raw base quality scores are often systematically biased and inaccurately convey the actual probability of mismatching the reference assembly (Li, 2011b; Liu *et al.*, 2013a). Therefore, in order to effectively utilize quality scores in the variant

calling step, they should be recalibrated to more accurately reflect true error rates.

One of the most widely applied base recalibration techniques has been implemented in the software GATK (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013), with others including SOAPsnp (Li *et al.*, 2009b) and ReQON (Cabanski *et al.*, 2012). The machine learning algorithm of GATK initially groups all loci that are not known to vary within a population into different categories with respect to several features (including their quality score, their position within the read (that is, machine cycle) and their dinucleotide sequence context). Next, an empirical mismatch rate is calculated for each category that is subsequently used to recalibrate base quality scores by adding the difference between the empirical quality scores and the mismatch rate to the raw quality scores

(Figures 3g–l). Thereby, the algorithm requires a set of known variants as a control, limiting the ability to detect *de novo* minor alleles (Ni and Stoneking, 2016). As a result, quality score recalibration strongly depends on the quality of previously available polymorphism data, restricting its usage to organisms with a public variant database. An alternative for species without a comprehensive SNP database is the utilization of highly confident candidate polymorphic sites from an initial SNP call for recalibration, followed by a second round of variant calling (Nielsen *et al.*, 2011).

STEP 4: VARIANT CALLING AND FILTERING

Generating a high-quality variant call set is challenging because of many subtle but complex types of errors, biases and uncertainties in the data. The analysis workflow involves a multistep procedure, utilizing advanced statistical models to capture features of library-, machine- and run-dependent error structures. First, positions or regions where at least one of the samples differs from the reference sequence need to be identified (variant calling) and individual alleles at all variant sites estimated (genotyping). Next, false positives should be removed from the initial variant data set to improve specificity (filtering). Here, I will focus on providing general guidelines for variant calling, genotyping and filtering (the probabilistic background for these methods is covered in detail in Nielsen *et al.*, 2011).

Discovery and genotyping

A multitude of bioinformatic tools have been developed to facilitate variant discovery from NGS read data (Table 4). Earlier proposed methods call variants by counting high-quality alleles at each site and applying simple cutoff rules—a technique that works well for high coverage ($>20\times$) sequencing data. However, this typically leads to problems, such as an undercalling of heterozygous genotypes in studies with a large number of individuals sequenced to low or medium coverage (Nielsen *et al.*, 2011)—a design that is frequently used to detect rare variants (Le and Durbin, 2011). Better suited for low to medium coverage data are more sophisticated approaches based on Bayesian, likelihood or machine learning statistical methods. They

Table 4 Characteristics of several popular SNP callers for reference-mapped data

Software	Method	Sample	Reference
Atlas-SNP2	Bayesian	Single	Challis *et al.* (2012)
CRISP	Testing	Pooled	Bansal (2010)
Dindel	Hidden Markov model	Pooled	Albers *et al.* (2011)
FreeBayes	Bayesian	Multiple	NA
GATK	Bayesian	Multiple	McKenna *et al*, (2010) DePristo *et al.* (2011) Van der Auwera *et al.* (2013)
QCALL	Bayesian	Multiple	Le and Durbin (2011)
SAMtools	Bayesian	Multiple	Li *et al.* (2009a)
SeqEM	Bayesian	Multiple	Martin *et al.* (2010)
SLIDERII	Counting	Single	Malhis and Jones (2010)
SNP-o-matic	Counting	Single	Manske and Kwiatkowski (2009b)
SNVer	Testing	Single and pooled	Wei *et al.* (2011)
SOAPsnp	Bayesian	Single	Li *et al.* (2009b)
SYZYGY	Bayesian	Pooled	NA

Abbreviations: NA, no publication available; SNP, single-nucleotide polymorphism.
Popularity was assessed by the number of citations of the software.

calculate the likelihood that a given locus is homozygous or heterozygous for a variant allele, taking additional information (for example, base and alignment quality scores, error profiles of NGS platforms and read coverage) into account, leading to calls with associated uncertainty information (Nielsen *et al.*, 2011). In addition, several recent techniques have been proposed that combine information from an initial read alignment with local *de novo* assembly in small regions of the genome (for example, GATK's HaplotypeCaller (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013) or Platypus (Rimmer *et al.*, 2014)), resulting in both high sensitivity and high specificity. Variant data are generally reported in the variant calling format (Danecek *et al.*, 2011).

Benchmarking studies indicate that a tool's performance strongly depends on the underlying study design, and thus no technique outcompetes all others under all circumstances (see, for example, O'Rawe *et al.*, 2013; Park *et al.*, 2014). In fact, no single method captures all existing variation and many tools provide contrasting and complementary information (Liu *et al.*, 2013a; Neuman *et al.*, 2013; O'Rawe *et al.*, 2013; Yu and Sun, 2013; Cheng *et al.*, 2014; Pirooznia *et al.*, 2014). As a consequence, it is recommended to combine independent data sets from multiple methods to achieve better specificity or sensitivity. Several tools have recently been designed to automatically merge variant call sets (see, for example, Cantarel *et al.*, 2014; Gao *et al.*, 2015; Gézsi *et al.*, 2015). However, these methods are computationally intensive, currently limiting their usage to small genomic regions.

In general, study design plays an important role in variant calling and genotyping: more specific variant calls and more accurate genotype estimations can be obtained by using high coverage sequence data (Fumagalli, 2013). However, given a finite research budget, sequencing samples to a high coverage often coincides with fewer samples being sequenced. Because of the limited sample size, these data sets are often a poor representation of a population's true genetic variation as many heterozygous individuals will be missed and many rare mutations will not have been sequenced. In contrast, although low coverage sequencing of a large number of individuals commonly provides a better picture of the variation in an entire population, it frequently results in a nonnegligible amount of genotype uncertainty that is exacerbated by sequencing and alignment errors (Crawford and Lazzaro, 2012).

Prior probabilities on the genotype can be improved and genotype uncertainties decreased by simultaneously calling multiple individuals of a population (based on allele frequency information and tests of Hardy–Weinberg equilibrium; Nielsen *et al.*, 2011). This joint calling does not only allow the filtering of systematic errors, but also increases the power to call poorly supported variants, provided sufficient support in other samples is available (Liu *et al.*, 2013a; Rimmer *et al.*, 2014). As a result, joint calling decreases the sampling bias and leads to substantial increases in sensitivity and accuracy compared with calls obtained from each individual independently. Another approach to more reliably infer genotypes in low coverage samples is the incorporation of information on haplotype structure (Nielsen *et al.*, 2011). However, it should be noted that this strategy may lead to biases in population genetic inference, and no additional information can be gained regarding the frequency of singletons (Li, 2011a).

The estimation of population genetic parameters strongly depends on the study design, whereby the highest accuracy is achieved by employing a large sample size, even if individuals are sequenced at a low coverage (Fumagalli, 2013). In recent years, several methods have been proposed that incorporate statistical uncertainty into their models by utilizing genotype likelihoods to directly estimate

population genetic parameters, such as allele frequencies at a single loci (Lynch, 2009; Keightley and Halligan, 2011; Kim *et al.*, 2011) or jointly across many loci (Keightley and Halligan, 2011; Li, 2011a; Nielsen *et al.*, 2012), mutation rates (Kang and Marjoram, 2011) as well as other summary statistics (Gompert and Buerkle, 2011; Li, 2011a; Nielsen *et al.*, 2012; Fumagalli *et al.*, 2013).

Filtering

Initial calls often contain many false positive variants caused by sequencing errors and incorrect alignments. As a result, it is important to apply filter criteria to achieve specificity in the data set. Thereby, filter strategies can be classified into two main techniques: hard filtering (for example, as implemented in GATK VariantFiltration (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Van der Auwera *et al.*, 2013) or VCFtools (Danecek *et al.*, 2011)) and soft filtering (for example, GATK VQRS).

Hard filtering is based on the assumption that false positives frequently exhibit unusual variant properties. For example, variants called in regions with exceptionally low or high coverage might result from an incomplete reference sequence, unresolved collapsed copy number variations or repeats—in general, regions where reads might be poorly aligned to the reference assembly (Figure 4a). Additional indicators of false positives include low-quality scores (for example, in duplicated regions, leading to ambiguous alignments; Figure 4b),

imbalanced strand specificity (that is, true variants should have a roughly equal coverage on both forward and reverse strands; Figure 4c), skewed allelic imbalance and a high clustering with other variants (Figure 4d). Hard filtering attempts to remove these false positives by identifying variants with characteristics outside their normal distributions. Specifically, it requires the setting of specific thresholds on these different variant characteristics, excluding variants that do not fulfill these criteria. Thereby, sex chromosomes will need to be considered separately, requiring different thresholds in order to deal with the different coverage patterns and rate of homozygous calls. However, as many characteristics are interdependent, simultaneous optimization of well-suited thresholds for sensitivity and specificity using hard filters is nontrivial. Furthermore, hard filtering is often complicated by uneven coverage, particularly in highly divergent (for example, within the *HLA* or *MHC* in humans) and repetitive regions (for example, centromeres and telomeres), where alignment is difficult (Lunter *et al.*, 2008).

Whereas simple hard filters have been shown to provide reliable results for high coverage data sets, soft filtering using machine learning methods has better specificity at low coverage (Cheng *et al.*, 2014). Soft filtering techniques initially build a statistical model based on a set of known high-quality variant calls (along with variant error annotations) that is subsequently used to evaluate the probability of each new variant being real (DePristo *et al.*, 2011). Unfortunately, because of the
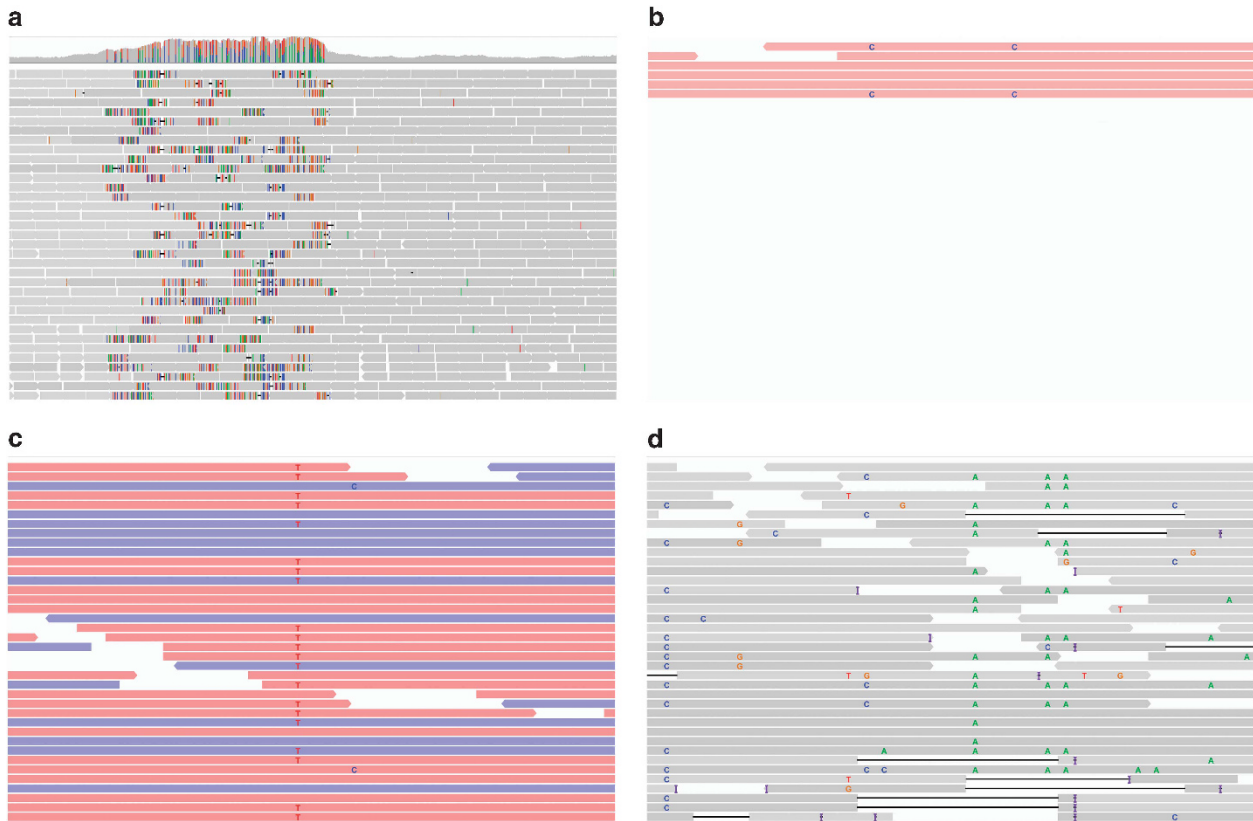

Figure 4 Indicators of false positive variant calls. Examples: (**a**) Variants in a region with exceptionally high coverage that might result from an incomplete reference sequence, unresolved collapsed copy number variations or repeats. Coverage is shown as vertical bars in the top panel. Alignments are provided in the lower panel with reads indicated as gray horizontal bars, variants as colored ticks, insertions as purple *I*s and deletions as black horizontal lines. (**b**) Variants only supported by low-quality reads (reads with a mapping quality of <5 are shown in red). (**c**) Variant supported by reads that show a strand bias (that is, the variant is stronger supported by the forward strand (red) than the reverse strand (blue)). (**d**) Region with a high clustering of variants. All examples were plotted using IGV (Robinson *et al.*, 2011).

requirement of a large training data set of known high-quality variants in the underlying model, these techniques are currently limited to model organisms for which extensive variation databases are available.

Quality assessment

Individual variants can be checked either by visual inspection (using dedicated software packages such as IGV (Robinson *et al.*, 2011), GenomeView (Abeel *et al.*, 2012) or SAVANT (Fiume *et al.*, 2010)), or by conducting a follow-up validation experiment using an independent technology, whereas the overall quality of the data set can be assessed using several simple metrics. First, the number of observed SNPs should be in line with the species' expected diversity. Second, the number of SNPs per chromosome should roughly correspond to its length. Third, every diploid individual carrying a SNP should (based on the mapped reads) exhibit an allele frequency of either 0.5 (that is, a heterozygous individual) or 1.0 (that is, an individual that is homozygous for the alternative allele) at the locus of the SNP.

SUMMARY

Over the past years, advances in NGS technologies together with a considerable decrease in costs has permitted high-throughput sequencing to become accessible for a wide range of research topics, including population-scale studies of genetic diversity in a multitude of organisms. Generating high-quality variant and genotype call sets from the obtained sequences is a nontrivial task, complicated by many subtle but complex types of errors, biases and uncertainties in the data. However, many statistical methods and computational tools have been proposed to tackle these challenges. Here, I outlined the main analytic steps of a SNP calling workflow required to optimize the accuracy of variant and genotype calling from raw resequencing data. Because of the fact that available software rapidly evolves in response to the changing sequencing technologies and protocols, I pointed out important factors to consider when choosing tools suitable for a particular study design as a guideline for researchers interested in generating their own variant data sets, rather than recommending a single established variant calling pipeline.

Nevertheless, there are still challenges. In particular, the correct identification of variants in highly diverse species or genomic regions, or in repetitive low-complexity regions, remains difficult. Under these circumstances, a valuable alternative is reference-free, assembly-based variant calling (for example, Cortex (Iqbal *et al.*, 2012)) that is agnostic to both variant type and divergence from the reference sequence, at the cost of a generally lower sensitivity and higher computational requirements. In addition, single-molecule, real-time, third-generation sequencing and mapping technologies can aid the detection of variation in these cases. Although not yet as well established, they have already been successfully applied to study genetic diversity (see, for example; Chaisson *et al.*, 2015; Gordon *et al.*, 2016) and indeed show great promise for future genomic research.

CONFLICT OF INTEREST

The author declares no conflict of interest.

Abeel T, Van Parys T, Saeys Y, Galagan J, Van de Peer Y (2012). GenomeView: a next-generation genome browser. *Nucleic Acids Res* **40**: e12.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P *et al.* (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res* **28**: E87.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C *et al.* (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R (2011). Dindel: accurate indel calls from short-read data. *Genome Res* **21**: 961–973.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F *et al.* (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**: 1061–1067.

Allhoff M, Schönhuth A, Martin M, Costa IG, Rahmann S, Marschall T (2013). Discovering motifs that induce sequencing errors. *BMC Bioinformatics* **14** (Suppl 5): S1.

Bansal V (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* **26**: i318–i324.

Bao H, Guo H, Wang J, Zhou R, Lu X, Shi S (2009). MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* **25**: 1554–1555.

Bolger AM, Lohse M, Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bonfield JK, Whitwham A (2010). Gap5 - editing the billion fragment sequence assembly. *Bioinformatics* **26**: 1699–1703.

Buschmann T, Bystrykh LV (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics* **14**: 272.

Bystrykh LV (2012). Generalized DNA barcode design based on Hamming codes. *PLoS One* **7**: e36852.

Cabanski CR, Cavin K, Bizon C, Wilkerson MD, Parker JS, Wilhelmsen KC *et al.* (2012). ReQON: a Bioconductor package for recalibrating quality scores from next-generation sequencing data. *BMC Bioinformatics* **13**: 221.

Caboche S, Audebert C, Lemoine Y, Hot D (2014). Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics* **15**: 264.

Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S *et al.* (2009). PASS: a program to align short sequences. *Bioinformatics* **25**: 967–968.

Cantarel BL, Weaver D, McNeill N, Zhang J, Mackey AJ, Reese J (2014). BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics* **15**: 104.

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

Carver T, Harris SR, Otto TD, Berriman M, Parkhill J, McQuillan JA (2013). BamView: visualizing and interpretation of next-generation sequencing read alignments. *Brief Bioinformatics* **14**: 203–212.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F *et al.* (2015). Resolving the complexity of the human genome using single molecule sequencing. *Nature* **517**: 608–611.

Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C *et al.* (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* **13**: 8.

Chen C, Khaleel SS, Huang H, Wu CH (2014). Software for pre-processing Illumina next-generation sequencing short read sequences. *Source Code Biol Med* **9**: 8.

Cheng AY, Teo YY, Ong RT (2014). Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics* **30**: 1707–1713.

Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**: 2601–2602.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**: 1767–1771.

Costea PI, Lundeberg J, Akan P (2013). TagGD: fast and accurate software for DNA Tag generation and demultiplexing. *PLoS One* **8**: e57521.

Crawford JE, Lazzaro BP (2012). Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet* **3**: 66.

Criscuolo A, Brisse S (2013). AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102**: 500–506.

Dai M, Thompson RC, Maher C, Contreras-Galindo R, Kaplan MH, Markovitz DM *et al.* (2010). NGSQC: cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics* **11** (Suppl 4): S7.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

David M, Dzamba M, Lister D, Ilie L, Brudno M (2011). SHRiMP2: sensitive yet practical SHort Read Mapping. *Bioinformatics* **27**: 1011–1012.

Davis MP, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**: 41–49.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* **8**: e85024.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C *et al.* (2011). A framework for variation discovery and genotyping using next generation DNA sequencing data. *Nat Genet* **43**: 491–498.

Dodt M, Roehr JT, Ahmed R, Dieterich C (2012). FLEXBAR - flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel)* **1**: 895–905.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36**: E105.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* **100**: 8817–8822.

Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Ewing B, Green P (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Res* **8**: 186–194.

Falgueras J, Lara AJ, Fernández-Pozo N, Cantón FR, Pérez-Trabado G, Claros MG (2010). SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics* **11**: 38.

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* **34**: e22.

Fiume M, Williams V, Brook A, Brudno M (2010). Savant: genome browser for high-throughput sequencing data. *Bioinformatics* **26**: 1938–1944.

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

Fumagalli M (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One* **8**: e79667.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A et al. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics* **195**: 979–992.

Gao X, Xu J, Starmer J (2015). Fastq2vcf: a concise and transparent pipeline for whole-exome sequencing data analyses. *BMC Res Notes* **8**: 72.

Gézsi A, Bolgár B, Marx P, Sarkozy P, Szalai C, Antal P (2015). VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. *BMC Genomics* **16**: 875.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* **108**: 1513–1518.

Gompert Z, Buerkle CA (2011). A hierarchical Bayesian model for next-generation population genomics. *Genetics* **187**: 903–917.

Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM et al. (2016). Long-read sequence assembly of the gorilla genome. *Science* **352**: aae0344.

Guan P, Sung WK (2016). Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* **102**: 36–49.

Hach F, Hormozdiari F, Alkan C, Hormozdiari F, Birol I, Eichler EE et al. (2010). mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**: 576–577.

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2014). mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* **42** ((Web Server issue)): W494–W500.

Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235–237.

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* **10**: R32.

Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14**: 184.

Heo Y, Wu X-L, Chen D, Ma J, Hwu W-M (2014). BLESS: bloom-filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics* **30**: 1354–1362.

Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N et al. (2015). An analytical framework for optimizing variant discovery from personal genomes. *Nat Commun* **6**: 6275.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J et al. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* **5**: e1000502.

Holtgrewe M, Emde AK, Weese D, Reinert K (2011). A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics* **12**: 210.

Homer N, Merriman B, Nelson SF (2009). BFAST: an alignment tool for large scale genome resequencing. *PLoS One* **4**: e7767.

Homer N, Nelson SF (2010). Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol* **11**: R99.

Ilie L, Fazayeli F, Ilie S (2011). HiTEC: accurate error correction in high-throughput sequencing data. *Bioinformatics* **27**: 295–302.

Ilie L, Molnar M (2013). RACER: rapid and accurate correction of errors in reads. *Bioinformatics* **29**: 2490–2493.

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

Jiang H, Lei R, Ding SW, Zhu S (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**: 182.

Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR et al. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**: 839–848.

Kang CJ, Marjoram P (2011). Inference of population mutation rate and detection of segregating sites from next-generation sequence data. *Genetics* **189**: 595–605.

Kao W, Chan AH, Song YS (2011). ECHO: a reference-free short-read error correction algorithm. *Genome Res* **21**: 1181–1192.

Keightley PD, Halligan DL (2011). Inference of site frequency spectra from high-throughput sequence data: quantification of selection on nonsynonymous and synonymous sites in humans. *Genetics* **188**: 931–940.

Kelley DR, Schatz MC, Salzberg SL (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**: R116.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G et al. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**: 231.

Kircher M, Heyn P, Kelso J (2011). Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* **12**: 382.

Kircher M, Kelso J (2010). High-throughput DNA sequencing - concepts and limitations. *Bioessays* **32**: 524–526.

Kircher M, Stenzel U, Kelso J (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol* **10**: R83.

Kong Y (2011). Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* **98**: 152–153.

Laehnemann D, Borkhardt A, McHardy AC (2016). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinformatics* **17**: 154–179.

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

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

Lassmann T, Hayashizaki Y, Daub CO (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* **27**: 130–131.

Le SQ, Durbin R (2011). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* **21**: 952–960.

Li H (2011a). A framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

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

Li H (2015). BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**: 2885–2887.

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

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

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

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

Li YL, Weng JC, Hsiao CC, Chou MT, Tseng CW, Hung JH (2015). PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. *BMC Bioinformatics* **16** (Suppl 1): S2.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K et al. (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966–1967.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**: 265–272.

Lim EC, Müller J, Hagmann J, Henz SR, Kim ST, Weigel D (2014). Trowel: a fast and accurate error correction module for Illumina sequencing reads. *Bioinformatics* **30**: 3264–3265.

Lindgreen S (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res Notes* **5**: 337.

Lindner R, Friedel CC (2012). A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS One* **7**: e52403.

Liu X, Han S, Wang Z, Gelernter J, Yang BZ (2013a). Variant callers for next-generation sequencing data: a comparison study. *PLoS One* **8**: e75619.

Liu Y, Schmidt B, Maskell DL (2011). DecGPU: distributed error correction on massively parallel graphics processing units using CUDA and MPI. *BMC Bioinformatics* **12**: 85.

Liu Y, Schröder J, Schmidt B (2013b). Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**: 308–315.

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

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J (2008). Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res* **18**: 298–309.

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

Ma Y, Xie H, Han X, Irwin DM, Zhang YP (2013). QcReads: an adapter and quality trimming tool for next-generation sequencing data. *J Genet Genomics* **40**: 639–642.

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

Manske HM, Kwiatkowski DP (2009a). LookSeq: a browser-based viewer for deep sequencing data. *Genome Res* **19**: 2125–2132.

Manske HM, Kwiatkowski DP (2009b). SNP-o-matic. *Bioinformatics* **25**: 2434–2435.

Mardis ER (2010). The $1,000 genome, the $100,000 analysis? *Genome Med* **2**: 84.

Mardis ER (2011). A decade's perspective on DNA sequencing technology. *Nature* **470**: 198–203.

Martin ER, Kinnamon DD, Schmidt MA, Powell EH, Zuchner S, Morris RW (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* **26**: 2803–2810.

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

Medvedev P, Scott E, Kakaradov B, Pevzner P (2011). Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics* **27**: i137–i141.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet* **11**: 31–46.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F *et al.* (2010). Tablet - next generation sequence assembly visualization. *Bioinformatics* **26**: 401–402.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y *et al.* (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**: e90.

Neuman JA, Isakov O, Shomron N (2013). Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinformatics* **14**: 46–55.

Ni S, Stoneking M (2016). Improvement in detection of minor alleles in next generation sequencing by base quality recalibration. *BMC Genomics* **17**: 139.

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

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

O'Rawe JA, Ferson S, Lyon GJ (2015). Accounting for uncertainty in DNA sequencing data. *Trends Genet* **31**: 61–66.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J *et al.* (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5**: 28.

Pagani I, Liolios K, Jansson J, Chen IA, Smirnova T, Nosrat B *et al.* (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* **40**: 571–579.

Park MH, Rhee H, Park JH, Woo HM, Choi BO, Kim BY *et al.* (2014). Comprehensive analysis to improve the validation rate for single nucleotide variants detected by next-generation sequencing. *PLoS One* **9**: e86664.

Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR *et al.* (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Hum Genomics* **8**: 14.

Planet E, Attolini CS, Reina O, Flores O, Rossell D (2012). htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28**: 589–590.

Reinert K, Langmead B, Weese D, Evers DJ (2015). Alignment of next-generation sequencing reads. *Annu Rev Genomics Hum Genet* **16**: 133–151.

Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SR, Wilkie AO *et al.* (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet* **46**: 912–918.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G *et al.* (2011). Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Ruffalo M, LaFramboise T, Koyutürk M (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**: 2790–2796.

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M (2009). SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**: e1000386.

Salmela L (2010). Correction of sequencing errors in a mixed set of reads. *Bioinformatics* **26**: 1284–1290.

Salmela L, Schröder J (2011). Correcting errors in short reads by multiple alignments. *Bioinformatics* **27**: 1455–1461.

Schatz MC (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* **25**: 1363–1369.

Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, Gibrat JF (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J Comput Biol* **19**: 796–813.

Schmieder R, Edwards R (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* **6**: e17288.

Schmieder R, Edwards R (2011b). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863–864.

Schneeberger K, Hagmann J, Ossowski S, Warthmann N, Gesing S, Kohlbacher O *et al.* (2009). Simultaneous alignment of short reads against multiple genomes. *Genome Biol* **10**: R98.

Schröder J, Schröder H, Puglisi SJ, Sinha R, Schmidt B (2009). SHREC: a short-read error correction method. *Bioinformatics* **25**: 2157–2163.

Schulz MH, Weese D, Holtgrewe M, Dimitrova V, Niu S, Reinert K *et al.* (2014). Fiona: a parallel and automatic strategy for read error correction. *Bioinformatics* **30**: i356–i363.

Sedlazeck FJ, Rescheneder P, von Haeseler A (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**: 2790–2791.

Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B (2014). Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res Int*; e-pub ahead of print 23 March 2014; doi:10.1155/2014/309650.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM *et al.* (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308–311.

Shi H, Schmidt B, Liu W, Müller-Wittig W (2010). A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware. *J Comput Biol* **17**: 603–615.

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

Smith AD, Xuan Z, Zhang MQ (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9**: 128.

Smolka M, Rescheneder P, Schatz MC, von Haeseler A, Sedlazeck FJ (2015). Teaser: individualized benchmarking and optimization of read mapping results for NGS data. *Genome Biol* **16**: 235.

Song L, Florea L, Langmead B (2014). Lighter: fast and memory-efficient sequencing error correction without counting. *Genome Biol* **15**: 509.

Tattini L, D'Aurizio R, Magi A (2015). Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng. Biotechnol* **3**: 92.

Tin MM, Rheindt FE, Cros E, Mikheyev AS (2015). Degenerate adaptor sequences for detecting PCR duplicates in reduced representation sequencing data improve genotype calling accuracy. *Mol Ecol Resour* **15**: 329–336.

Treangen TJ, Salzberg SL (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A *et al.* (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11**: 11.10.1–11.10.33.

Weese D, Emde AK, Rausch T, Döring A, Reinert K (2009). RazerS - fast read mapping with sensitivity control. *Genome Res* **19**: 1646–1654.

Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* **39**: e132.

Wu TD, Nacu S (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881.

Yang X, Chockalingam SP, Aluru S (2012). A survey of error-correction methods for next-generation sequencing. *Brief Bioinformatics* **14**: 56–66.

Yang X, Dorman KS, Aluru S (2010). Reptile: representative tiling for short read error correction. *Bioinformatics* **26**: 2526–2533.

Ye K, Hall G, Ning Z (2016). Structural variation detection from next generation sequencing. *Next Generat Sequenc Applic* **S1**: 007.

Yu X, Sun S (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* **14**: 274.

Zhou Q, Su X, Wang A, Xu J, Ning K (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One* **8**: e60234.