Вавиловский журнал генетики и селекции. 2016;20(6):746-755
DOI 10.18699/VJ16.191

Computational problems of analysis of short next generation sequencing reads

R. te Boekhorst[1], F.M. Naumenko[2], N.G. Orlova[2, 3], E.R. Galieva[2, 4], A.M. Spitsina[2], I.V. Chadaeva[2], Y.L. Orlov[2, 4] ✉, I.I. Abnizova[5] ✉

[1] University of Hertfordshire, Hatfield, UK
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Novosibirsk State University of Architecture and Civil Engineering (Sibstrin), Novosibirsk, Russia
[4] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[5] Wellcome Trust Sanger Institute, Cambridge, UK

Short read next generation sequencing (NGS) has significant impacts on modern genomics, genetics, cell biology and medicine, especially on meta-genomics, comparative genomics, polymorphism detection, mutation screening, transcriptome profiling, methylation profiling, chromatin remodelling and many more applications. However, NGS are prone for errors which complicate scientific conclusions. NGS technologies consist of shearing DNA molecules into collection of numerous small fragments, called a 'library', and their further extensive parallel sequencing. These sequenced overlapping fragments are called 'reads', they are assembled into contiguous strings. The contiguous sequences are in turn assembled into genomes for further analysis. Computational sequencing problems are those arising from numerical processing of sequenced samples. The numerical processing involves procedures such as: quality-scoring, mapping/assembling, and surprisingly, error-correction of a data. This paper is reviewing post-processing errors and computational methods to discern them. It also includes sequencing dictionary. We present here quality control of raw data, errors arising at the steps of alignment of sequencing reads to a reference genome and assembly. Finally this work presents identification of mutations ("Variant calling") in sequencing data and its quality control.

Keywords: next generation sequencing (NGS); DNA; sequencing technologies; statistical biases; genome polymorphisms; sequencing errors; review.

Вычислительные проблемы анализа ошибок коротких прочтений ДНК при секвенировании следующего поколения

Р. те Боекхорст[1], Ф.М. Науменко[2], Н.Г. Орлова[2, 3], Э.Р. Галиева[2, 4], А.М. Спицина[2], И.В. Чадаева[2], Ю.Л. Орлов[2, 4] ✉, И.И. Абнизова[5] ✉

[1] Университет Хартфордшира, Хатфилд, Великобритания
[2] Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия
[3] Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Новосибирский государственный архитектурно-строительный университет (Сибстрин)», Новосибирск, Россия
[4] Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
[5] Институт Сэнгера, Велком Траст, Кембридж, Великобритания

Секвенирование следующего поколения (NGS) с помощью коротких прочтений ДНК вносит большой вклад в решение задач современной геномики, генетики, клеточной биологии и медицины, особенно в исследования метагеномики, сравнительной геномики, определение полиморфизмов, скрининг мутаций, транскриптомное профилирование, изучение ремоделирования хроматина и многие другие приложения. Секвенирование неустойчиво к техническим ошибкам, которые могут влиять на научные выводы. NGS технологии состоят из создания коллекции многочисленных коротких фрагментов ДНК, именуемой «библиотекой», получения молекулярных колоний и их дальнейшего массового параллельного секвенирования. Такие секвенированные фрагменты называются «прочтениями», они собираются (ассемблируются) в протяженные строки. Протяженные последовательности, в свою очередь, собираются в геномы для дальнейшего анализа. Вычислительные/процессинговые ошибки и сбои секвенирования – это ошибки, возникающие при последующей цифровой обработке секвенированных образцов. Последующая обработка (процессирование) включает процедуры оценки качества, картирования, ассемблирования и даже корректировки ошибочных данных. Данная статья рассматривает вычислительные ошибки процессирования, компьютерные и статистические подходы для их определения, а также представляет словарь терминологии секвенирования. Рассмотрены задачи идентификации мутаций («Определение вариаций») в данных секвенирования и контроль качества их определения. Определение вариаций включает локальные вариации, такие как одиночные нуклеотидные полиморфизмы, короткие вставки и делеции (инделы), и масштабные вариации (инверсии, транслокации или большие

REVIEW
Received 18.11.2016
Accepted for publication 20.12.2016

✉ e-mail: ia1@sanger.ac.uk; orlov@bionet.nsc.ru

инделы). Обсуждены проблемы контроля качества исходных (сырых) данных, ошибки, возникающие на этапах выравнивания прочтений последовательностей ДНК на референсный геном и последующего выравнивания/ассемблирования.

Ключевые слова: секвенирование следующего поколения; ДНК; технологии секвенирования; статистические неоднородности; геномные полиморфизмы; ошибки секвенирования; обзор.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:
те Боекхорст Р., Науменко Ф.М., Орлова Н.Г., Галиева Э.Р., Спицина А.М., Чадаева И.В., Орлов Ю.Л., Абнизова И.И. Вычислительные проблемы анализа ошибок коротких прочтений ДНК при секвенировании следующего поколения. Вавиловский журнал генетики и селекции. 2016;20(6):746-755. DOI 10.18699/VJ16.191

HOW TO CITE THIS ARTICLE:
te Boekhorst R., Naumenko F.M., Orlova N.G., Galieva E.R., Spitsina A.M., Chadaeva I.V., Orlov Y.L., Abnizova I.I. Computational problems of analysis of short next generation sequencing reads. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016;20(6):746-755. DOI 10.18699/VJ16.191

Here we will concentrate on second generation short read DNA sequencing (Liu et al., 2012), and will drop off the term ‘second generation short read’ while speaking about NGS further. The **terms in bold** are explained in the Glossary.

NGS technologies have as an essential feature of breaking DNA molecules into stack of numerous fragments, called a ‘library’. The end parts of these fragments are sequenced in parallel, and called ‘reads’. They are assembled into contiguous strings. These assembled sub-sequences are in turn assembled into genomes, and are subjects for further analysis.

We refer to (van Dijk et al., 2014; Anders et al., 2015) for detailed characteristics of NGS platforms.

The NGS data processing is arranged in a set of consecutive steps, called a pipeline**.** A common post-sequencing NGS pipeline (Mutarelli et al., 2014; Newell, 2014) consists of:
(1) *Quality Control (QC) of initial data*;
(2) *Mapping to a reference genome and/or assembly*;
(3) *Post-mapping/assembly QC and re-calibration*;
(4) *Variant calling and its QC*;
(5) *Correcting of errors*.

The step 2 may be combined and/or substituted by a de-novo genome assembly (Baker, 2012) in case there is no reference for the sequenced genome.

We refer readers to reviews on NGS computational frameworks (Dolled-Filhart et al., 2013; Guo et al., 2014a; Mutarelli et al., 2014; Pabinger et al., 2014). The on-line sources (Li J. et al., 2012b; Hadfield, 2013; Ignatieva et al., 2015) can help researchers to build up their own pipelines. For ‘meta’ pipelines we refer to (Anders et al., 2015; Wolfinger et al., 2015) which describe tools to build up bespoke pipelines.

For each of the steps in a pipeline above we will review (a) what they are and what is their goal; (b) how it is done summarising the methodology, their advantages and pitfalls.

1. QC of initial data

For any platform, an initial unprocessed digital outputs of a sequencing are a **base calls** and their **qualities**.

Compared to a first generation Sanger sequencing (Sanger et al., 1992), NGS technologies are confronted by shorter read length, platform/instrument/sample specific biases (Harismendy et al., 2009), higher error rate, and irregular coverage. These factors lower the accuracy of NGS further analysis (e. g. variant calls and de-novo assembly) by introducing sequencing errors that may direct to mis-interpretation of data.

The main factors utilised in quality control of raw data to characterise sequencer’s performance and library preparation are:

Total read count; *Proportion of high quality data*; *Nucleotide and quality distribution per cycle*; *Duplication rate (can be optical or amplification duplicates)*; ***Adapter’s** counts*; *Proportion of bases per sample for pooled **multiplexed** data*;

◊ *Total read count* shows general library effectiveness. It should be reasonably large to produce results of statistical significance.

◊ Proportion of high quality (Q > 30) bases within *Q value distribution* should be large: at least more than half. A base call is scored with low Q mostly because of sequencer’s preferences and faults (Abnizova et al., 2010; Ledergerber, Dessimoz, 2011). These low quality bases are typically trimmed or corrected (Kelley et al., 2010; Del Fabbro et al., 2013), so low Q and possibly wrong called data will not compromise downstream analysis. However, an error correction should be applied only to high**-**coverage and homogeneous data – an assumption that often fails for NGS data.

◊ *Quality-per-cycle distribution*. A random quality peaks/deeps per cycle point to some problems on machine during sequencing. Quality usually declines gradually with cycle as a result of increasing signal-to-noise ratio.

◊ *Duplicate reads*, appearing due to PCR (polymerase chain reaction) and optical problems, may lead to over-estimating of some variant contribution in the data. Duplicate removing is debated in (Pireddu et al., 2011; Davis et al., 2013). Thus, their proportion should be less than 10 %.

◊ *Proportion if adapters* should less than 10 % as well. Parts of adapter might be erroneously sequenced in the beginning of a read, and thus may bring artificial mutations (Martin, 2011; Li J. et al., 2012a). The popular tools for adapter removing are discussed in (Marroni et al., 2012; Jiang et al., 2014).

◊ *Di-multiplexing,* namely splitting up samples based on their tags, should be even across tags in theory. It is very important that the size of each pool is sufficient and equal (Mir et al., 2013) for pooled multiplexed samples. Fairly even di-multiplexing (Hadfield, 2013) provides less biased data.

◊ There is also a possibility to quality control a library before massive sequencing. The MiSeq QC (Illumina, 2014) enables performing a preliminary run on libraries before deep-sequencing on a bigger machine, such as HiSeq or HiSeqX.

Nevertheless, any individual QC metric should be regarded in context of its project (Guo et al., 2014a).

A lot of sequencers (Cox et al., 2010) generate a QC reports included into their processing pipeline, and these reports investigate mainly a general performance of the corresponding sequencer. They typically do not cover any effects of sample extraction and library preparations.

A special case is fastQC. It is built up to point to problems which developed either in the library preparation or on sequencer. It is a very fast and crude estimation of different metrics formed by stratified samples of the data.

Alternatively, FaQCs (Lo, Chain, 2014) records errors in the whole data. It also takes away low Q-value reads.

The NGS QC Toolkit (Patel, Jain, 2012), except of performing a quality check and generating descriptive statistics, trims low Q ends of reads and removes low Q bases. It also enables a conversion between various file formats of NGS data from Illumina and Roche 454 platforms.

One should be careful: what is removed might be a genuine biological signal. Nevertheless, any fluctuation from expected values for the QC metrics, might be a possible error.

2. Mapping/aligning to a reference genome and/or assembly

The further step is the matching of the reads to positions at the reference genome, so called mapping. This is done by aligning reads to sub-sequences of the reference genome to which they are most close in terms of nucleotide sequence. Computationally, mapping is the most time and memory consuming step (Day-Williams, Zeggini, 2011; Fonseca et al., 2012). It is also critical: any mistake in alignment will be subject to further processing and hence spread errors to the further stages of sequencing and analysis.

For the short reads of NGS, it is too inefficient in time and memory to use the well-known BLAST (Altschul et al., 1990) algorithm to map reads to genome. Therefore a particular memory and time optimised mapping algorithms are developed.

NGS mappers/aligners can be classified based on their methods: hash table indexing (Shang et al., 2014) or Burrows-Wheeler Transform (BWT) (Li, Durbin, 2010). They also differ by computer resource usage and sensitivity. Thus, they may lead to a different mapping results. Here we define sensitivity as a proportion of genome which is covered by at least one read after mapping. Mapping algorithms also vary in their ability to deal with particular sequencing platforms, quality of base, protocols and in the dealing with structural features of the DNA subject to sequencing, such as repetitive motives, gaps, deletions and insertions.

Both types of aligners typically pre-process and index both reference and reads before a search of matching read positions (in the reference genome) itself. A hash table is a kind of a look up table, only supplied with advanced structure of indexing. BWT usually compresses data in a particular way (modification of a suffix array) before matching. BWT aligners are less sensitive than hash table methods, but are faster and use less memory (Newell, 2014).

A sequence assembly refers to aligning and integrating short fragments from a sequenced DNA in order to recreate the original sequence. If the genome of an organism has not been sequenced before, the assembly results in the first form of its reference genome. This procedure is called "de-novo assembly". Sometimes a de-novo assembly is used together with alignment to reconstruct previously insufficiently covered and untrustworthy sequenced genome loci.

Present-day assembling algorithms for NGS comprise two main groups (Li Z. et al., 2012): (i) Overlap-layout-consensus (OLC) methods; and (ii) Eulerian/de Bruijn Graph (DBG) methods. Both groups apply a graph theory to deal with NGS data, but in OLC notation reads are nodes, while in DBG notation a *k*-mer is a node. A read's overlapping sequences stand for graph edges in both groups of assemblers.

We refer to (Li, Homer, 2010; Nagarajan, Pop, 2013; Otto et al., 2014; Pightling et al., 2014) or comparisons and benchmark aligners tests; for exhaustive literature on assemblers read (Nagarajan, Pop, 2013; Chin et al., 2014; Shang et al., 2014). An online list of aligners is regularly updated at (http://www.ebi.ac.uk).

What can go wrong?

- *Reference mistakes.* One should understand that an alignment step is apparently dependent on a reference's accuracy. In the case of incorrect reference many reference errors could be mistreated as high quality genetic variants.
- A bias *shared by* most technologies is that their accuracy decreases with the number of sequencing cycles, thus an error of mapping the end of a read grows (Balint, 2016; Sameith et al., 2016).
- Of the more specific defects, we refer to: *platform-dependent issues*; *the type of protocol used*; *complications due to the functional and structural complexity of the sample DNA.*
- *Read Length and Error Rates*

Read lengths span from 70–1500 bp (Newell, 2014) depending on the sequencing platform. If reads are short it is harder to match them precisely to a unique genomic location.

Some sequencing platforms allow for longer read lengths than others (for example 200 bp by Ion Torrent and 700 bp by Roche's 454, while Illumina's reads of 100–250 bp) which makes mapping more precise. However, this advantage is defeated by their higher mismatch-error rate; aligners throw away reads with too many mismatches on the basis of a preset mismatch error rate.

- *Platform-dependent issues*

The technology on which a platform is founded may be prone to a certain sequencing mistakes, resulting in platform-specific error characteristic.

A 'light-based' sequencing platforms, Illumina, SOLiD, and Complete Genomics, employ fluorescent dye's labelling to measure a signal strength for a successive sequencing cycle. The light-based platforms are known to be impaired by GC-bias, i. e. a low coverage of either **GC-rich** or **GC-poor** DNA regions (Chen et al., 2013; Rieber et al., 2013). Its origin is likely to be a fragmentation or/and cloning procedures during library preparation (Benjamini, Speed, 2012; Ross et al., 2013).

The light-based platforms typically are disadvantaged by single nucleotide miss-identifications. The SOLiD platform is known to have difficulties with sequencing palindromic sequences (Huang et al., 2012).

Ion Torrent's Personal Genome Machine (PGM) (Niu et al., 2010) utilises semiconductor sequencing technology that operates on acidity (pH) rather than on light. Roche's 454 (Niu et al., 2010) employs a pyro-sequencing technology. An accuracy of both technologies depends on the length of sub-sequences of identical nucleotides ("homo-polymers") because of similar computational approaches to evaluate a homo-polymer length. Defective flow-calls result in insertion/deletion (indel) errors: they are largely homo-polymer-associated errors in case when short homo-polymers are frequent while long are rare (Bragg et al., 2013; Li et al., 2013).

Recognizing indels from NGS is known to be very daring (Li et al., 2013), because 'indel by itself obstructs with precise mapping'. To map indels precisely, pair-end (PE) information is employed (Albers et al., 2011). It is valid for indels half a size of reads. Longer deletions are detected by a split-read method.

To distinguish long insertions a de-novo assembly of weakly covered regions is required (Li et al., 2013).

- *Sequence-specific errors*: For pyrosequencing platforms, a 'homopolymer-associated errors' result into throwing away repetitive DNA after mapping. Indel errors are known to be context-dependent. Moreover, for Ion Torrent, GC-poor organisms have higher error rate and poorer coverage than GC-balanced. The nucleotide context of Illumina errors is reported in (Minoche et al., 2011).
- *DNA complexity*: *DNA functionality causes aligning biases*

The study of NGS artefacts in (Schwartz et al., 2011) showed that less linguistically complex sequences of introns are less covered with reads than more complex sequences of exons. The authors discovered that peaks of mapped reads were associated with biological features, such as intron-exon junction, expression level, splice sites and transcription length.

Similarly, the authors of (Auerbach et al., 2009) found that regions proximal to promoters are prone for sonication breakage, and hence are the subjects of regional bias. These regions are the primary cause of an uneven read coverage, retaining a large peaks of aligned reads.

- *DNA complexity*: *Repetitive DNA causes assembly problem*

A particular troublesome feature of the sequential structure of many genomes is the occurrence of long chunks of repetitive DNA (so-called "repeats"): repetitive DNA is frequently overlooked, miss-mapped and miss-assembled by all platforms (McCoy et al., 2014).

Around half of human genome is comprised of repetitive DNA (de Koning et al., 2011), the fraction of repeats is even larger for some plant genomes (Feschotte et al., 2002). Even though repetitive DNA is functionally important, NGS sequencing often fails to sequence it flawlessly (Alkan et al., 2011b; Ye et al., 2011). Most current technologies are error-prone while handling repeats.

But granting a repetitive DNA stretch is sequenced correctly, it might be compromised by similar DNA in other genome location, and lead to mis-alignment. And finally, repetitive DNA is often a hot-spot of real biological mutations and structural variations (Orlov et al., 2006; Medvedev et al., 2009; Safronova et al., 2015, 2016).

In addition to various repetitive DNA, a short indels and segmental duplications are also difficult to align and assemble (McCoy et al., 2014) because of ambiguity at which location to map an identical DNA subsequence.

The main assumption of assembly (similar reads belong to the same location) is breached by various repeats and polymorphisms. An assembly is computationally not tractable for genomes where the ratio of repeat length to read length is large (Nagarajan, Pop, 2013).

When whole long repetitive stretch were sequenced together with their flanking regions, it would be easier to detect it within genome. Therefore, longer reads could solve this problem (Huddleston et al., 2014).

- *Diversity of protocols*: *PE and MP methods*

The types of sequencing protocols depend on a researcher's question: e. g. reads sequenced in pairs (pair end, PE) (Medvedev et al., 2009) or singles (SE). PE reads are designed to detect direction and distance between reads, therefore reads containing complex DNA can be mapped uniquely (Miller et al., 2010; Alkan et al., 2011a).

A sub-type of PE reads, the long inserts reads (up to 5 KB), frequently named as **mate-pair libraries (MP)** (Park, 2013) are valuable to connect long repeats (including **repetitive transposable elements**) and structural variations.

Longer reads can solve the assembly and mapping problems. With longer reads it is easier to establish a correct genomic location for a sequenced DNA. Therefore, a new synthetic long reads (McCoy et al., 2014) from the Illumina TruSeq are developed. They are as long as 3d generation PacBio (Sharon et al., 2013), but much more accurate, having as low error rate as 0.03 % per base.

These synthetic long reads are assembled from Illumina short reads, by combination of laboratory and computational efforts (Voskoboynik et al., 2013). Nonetheless, there are still some imperfections left: gaps in assembly and a low coverage for repetitive AT-rich regions.

Regrettably, when some problems are reduced, a new ones arise. The essential problems of MP (Park, 2013) are: (i) extremely elaborated construction of their libraries, and (ii) common mistakes of mapping: 'inward facing' reads as a substitute of 'outward facing'. This mistake results into **chimeric reads** (Illumina). Another problems are: unexpectedly small **insert sizes** (Nextera), underrepresentation of the AT-rich sequences (SOLiD) and unplanned spontaneous secondary fragmentation (Roche).

- *Sequencing errors* (Abnizova et al., 2012; Ross et al., 2013) is another threat for aligners. Clearly, if a read encloses more mismatches than allowed by aligner settings, than it will be discarded, even if it accommodates biological signal.

Another objection significant discordance of assemblers (Magoc et al., 2013): different assemblers yield very unequal amount of assembled reads for the same data sets, specifically for homologous genome regions.

3. Post-mapping/assembly QC and re-calibration

Mapping is known (Li H. et al., 2009) to be the a primary cause of sequencing biases. Therefore it is recommended that one reviews the quality of mapped reads before in-depth scientific analysis.

3.1. Mapping metrics

To safeguard an adequate aligners' performance, there are several QC metrics:

◊ Number and proportion of high quality reads mapped;
◊ Coverage uniformity and average/median depth of co-verage;
◊ Quality-associated metrics: mapping quality, base quality score distribution, etc.;
◊ Insert size distribution; most frequent *k*-mers;
◊ Protocol-linked metrics: fraction of discordant pairs; capture efficiency;
◊ Forward and reverse read strand symmetry;
◊ Error rate per cycle;
◊ Possible contaminations metrics, see Chapter 4; GC-bias metrics;
◊ For a confident variant call two metrics are of great importance: sequencing coverage depth and uniformity.

Ideally, an even read coverage is expected along genome, to escape local biases. On the other hand, coverage is known (Minoche et al., 2011) to be non-uniform along genome, depending on the regional function, composition (Rieber et al., 2013) and many other features.

◊ The Q-value/score is a commonly used measure of base call quality (Bonfield, Staden, 1995; Ewing et al., 1998). The quality Q-scores compress different types of information about the quality of base calls into a confidence (of error) value. Quality score is commonly accepted input for majority of analysis tools, assemblers and aligners in order to produce accurate results.

However, in a raw fastq/bam files these Qs are inferred or predicted. The predictions are based on a set of measurements of a base call, and on previous observations of the values of the measurements. The inferred Q-values are assigned by the means of pre-computed look up table, so called calibration table (Brockman et al., 2008; Abnizova et al., 2010).

A sequencer's errors are typically of low Q, and come from technological and hardware shortcomings.

The infamous sources of errors for Illumina sequencers are: phasing and pre-phasing, dye label X-talk, molecule degradation with time and G-quenching (IDT, 2011). The phase inaccuracy results from base-incorporation errors on a sequencer machine. A G-quenching is an effect of previous nucleotide G; a base quality is typically low for this G-preceded base call (Abnizova et al., 2010, 2012). It was strongly pronounced for the v3 version HiSeq, and dramatically reduced for HiSeqX10 and X5.

◊ Contaminated sequences (due to different reasons) may bring up artefacts during variant calling (Schmieder, Edwards, 2011).
◊ A capture efficiency for exome sequencing is a proportion of useful data (Garcia-Garcia et al., 2016). It is normally 40–75 % (Guo et al., 2014a), and should not be too small for statistically sound results.

And likewise to the section 1, any inconsistency with expected values for a sample investigated should be cautioning.

3.2. Assembly metrics

In the non-existence of reference genome, the assembly metrics are:

◊ Total number of contigs or scaffolds: the less the better;
◊ Contig or scaffolds sizes: max, mean and N50. N50 is defined as the length of the scaffold/contig, which overlaps the midpoint of length-ordered concatenation of scaffolds/contigs;
◊ Total size of scaffolds. It should be close to an expected size of a genome sequenced;
◊ Number of Ns should be limited. (The created gaps in assembly are filled with the uninformative base-pair character '**N**'.)

An assembly accuracy and several normalised metrics are possible to assess in case when a reference genome exits. Note that normalization accounts only on those parts of assembly, which can be mapped to a reference genome by standard local alignment tools.

◊ Sensitivity of assembly is defines as a percent of genome assembled.
◊ Normalised N50 for contigs and for scaffolds is more complicated than for contigs because of N-filler of gaps (Makinen et al., 2012).

3.3. Q re-calibration

A predicted Qs often do not correspond to an actual Qs for a certain run/lane/library. In this incident (and in case heterogeneous data are combined) it is suggested to re-calibrate the data (Ewing et al., 1998; Massingham, Goldman, 2012). In the WTSI we implemented the in-house recalibration and error analysis tools (Abnizova et al., 2010). Instead of trimming an ambiguous base calls, we warn (low Q) about possible sequencing errors. Trustworthy Q-value is known to increase SNP call accuracy (Li, Stoneking, 2012) more than hard filtering.

4. Variant calling and its QC

Variant calling from NGS data is defined as a computational methods for establishing an event of genetic variant resulting from NGS experiments (Lawrence, 2014; Zhang et al., 2015).

Variant calling involves small-range variants (Kojima et al., 2013), such as single nucleotide polymorphisms (SNPs), short insertions and deletions (indels), and large-range structural variants, copy number variants (CNV) and structural variants (SV). A SVs are inversions, translocations, or large indels. All types of variants are identified by comparison to a reference genome.

Fraction of variation in genomes is significant: e. g. for human genome, SNPs comprise around 0.1 %, although SV's contribution is estimated as 1.2 % (Tattini et al., 2015) and CNV's contribution is as large as 15 % (Wong et al., 2010).

A variant calling is crucial for comparative genomics and genetics of human diseases. A valuable variant calling application is clinical testing: identifying disease-associated mutations (Chin et al., 2013).

Variant calls are implemented in two ways: (i) after aligning reads, or (ii) after assembling reads. Sometimes these steps are combined. SNPs and small indels can be identified by alignment of short sequencing reads to a reference genome. However, larger structural variants and repetitive regions in the genome are harder to find.

Structural variation can disturb genes or regulatory elements, therefore whole-genome sequencing is not complete without assembly and detection of structural variation (Li H. et al., 2009). In the (i) case, a position of each read relative to the reference genome if identified first. After reads are aligned, a set of QC steps, involving recalibration, duplicate removing, and indel-realignment, are done before variant calling.

In the (ii) case, an assembly of un-processed reads is performed first, and only after this the assembly is set against a reference genome (if the later exists). Variant detection after assembling is beneficial to individual genes (Olson et al., 2015), but it loses power when applied to a whole genome: in the absence of a reference genome it is not possible to identify other genome's contaminations; spurious variants can not be verified by raw reads after assembling.

Somatic versus germline mutation

Variant calling from NGS is well utilised in genetics of human diseases. There are three typical ways how NGS data is applied in the area: (a) detection of causal germline mutations in Mendelian disorders (Lettice et al., 2008; Stitziel et al., 2011); (b) detection of putative genes for complex diseases with GWAS (Day-Williams, Zeggini, 2011; Lander, 2011; Marian, 2012); (c) detection of somatic and constitutional mutations in cancer (Walther et al., 2009).

It is more complicated to identify a somatic mutation than a germline mutation (Pabinger et al., 2014).

To identify somatic mutations in cancer, they typically compare tumor vs/and normal samples for the same individual (Vissers et al., 2011; Yan et al., 2011).

An annotation step is commonly performed after a variant call (Li R. et al., 2009; Wang et al., 2010; Yang, Wang, 2015). Annotation is done by utilising a public databases, e. g. dbSNP, dbVar (Lin et al., 2015). The next step is visualization (Nielsen et al., 2010).

A set of metrics to assess a quality of variant call is listed below (Guo et al., 2014b; Jun et al., 2015):

◊ **Ti/Tv ratio**, individually for whole genome sequencing (WGS) and whole exome sequencing (WES) (should be 2 and 3);
◊ **Heterozygocity ratio**;
◊ Number of known and of new SNPs per person: should be not more than 200;
◊ Cross species and within species contamination; **genotype** consistency;
◊ SNP spatial density; QC per SNP;
◊ Strand, cycle, allele balance, reference allele biases; haplotype scores;
◊ Performance metrics: Sensitivity and specificity of single nucleotide variant call.

One can combine these metrics by a machine learning methods (DePristo et al., 2011; Jun et al., 2015). In order to minimise false positive (FP), some variant callers do a lot of filtering and trimming using metrics above: by applying a minimum depth of coverage threshold, by masking of homo-polymers and repeats, by trimming poor quality bases from a read etc. Unfortunately, while reducing FP, one can increase false negative (FN) by applying these filters (Olson et al., 2015).

◊ To assess a goodness of a variant caller, one should use a performance metrics: accuracy, sensitivity and specificity, (Olson et al., 2015) given a reliable benchmarking test sets and reference.

A comprehensive review of post-map QC is performed in (Wyllie, 2013; Guo et al., 2014b). The GATK (DePristo et al., 2011) utilises variant QC metrics for their variant calls, applying genotyping and known SNP information for a variant QC and annotation. However, there seems to be no a standard evaluation of a variant caller (Olson et al., 2015) so far.

5. Correction of errors

A definite amount of errors is the result of sequencing and post-processing imperfections. One way to tackle them is to Q-score possible known artefacts low, so they would be not used by further analysis. Another way is to correct errors using a knowledge about error sources for various platforms' errors (Edgar, Flyvbjerg, 2015; Olson et al., 2015) and computational biases.

An error correction *after mapping* is correction of a mismatch between sequenced read and a reference. *After/during assembling* error correction is a general agreement of base calls across all reads belonging to the same assembled location.

There are multiple attempts to correct sequencing errors. However, an error correction might introduce new type of errors: mis-correction errors (Yang et al., 2013; Fujimoto et al., 2014). And these errors are more difficult to correct back than technological errors.

A sound comparison of NGS platforms is done in (Yang et al., 2013) together with very good explanation of modern error-correction methods. Surprisingly, the paper is very convincing that one should NOT introduce new mis-correcting errors. Additionally, it also does not look promising to correct reads without understanding causes of sequencing/library errors. The work (Fujimoto et al., 2014) confirms that error correction methods can not handle heterozygosity, and they introduce new mis-correction errors.

There are approaches to correct for known context biases, such as GGGGT error patterns for Illumina (Minoche et al., 2011; Nakamura et al., 2011). However, new Illumina releases (e. g. HiSeqX10) are almost free from old type motif-dependency, and new artefacts (such as larger context dependence on a next base) appear.

Error models are used in (Janin et al., 2014) to realistically simulate individual sequencing runs and/or technologies. These models are mostly empirically derived and context-based. A comparison of genomes without assembling them is introduced by (Patro, Kingsford, 2015).

It might be beneficial to do so for de novo sequenced genomes. However, possible PCR biases in coverage are not included in the model. Some studies, such as (Orton et al., 2015) developed a computational error model of Illumina's sample processing, which involves experimental steps. This model infers possible genomic genome locations of PCR errors.

As a conclusion, one should be informed of possible biases, and make decisions depending on their study's aim. Overall conclusion is in necessity to use short sequencing reads error correction for the mapping and processing NGS data, depending on sequencing platforms. Details of error corrections publications will be presented in next paper.

Acknowledgement

Authors are grateful to M.P. Ponomarenko, G.V. Orlova, G.V. Vasiliev and M. Chen for science discussion. Development and testing of genomic computer pipelines was supported by Russian Science Foundation (project 14-24-00123). Application of software for plant transcriptomics was supported by RFBR (16-54-53064).

Glossary

The well-known **Sanger sequencing method** (Sanger et al., 1992) is called a first-generation DNA sequencing technology. The next generation sequencing technologies (Liu et al., 2012) include: (i) 2nd generation sequencing, the massive parallel sequencing of relatively short DNA fragments (Dolled-Filhart et al., 2013); and (ii) 3d generation sequencing, in which single DNA molecules hence much longer fragments (Schadt et al., 2010) are sequenced. In this paper we will focus on 2nd generation DNA sequencing, and will omit the term '2nd generation' while mentioning NGS further.

With NGS technologies, bases are inferred from light/chemistry intensity signals, a process commonly referred to as **base-calling.** The sequenced bases are assigned A, C, G or T letters depending on the intensity.

The **Q-value/score** is the most well accepted measure of base call quality (Bonfield, Staden, 1995).

The quality Q-scores compress a variety of types of information about the quality of base calls into a probability-of-error value.

Mapping or **aligning** is the matching of the reads to locations at the reference genome. This is done by aligning reads to stretches of the reference genome to which they are most similar in terms of nucleotide sequence.

A **sequence assembly** refers to aligning and merging short fragments from a DNA sequence in order to reconstruct the original sequence.

If the genome of a species has not been sequenced before, the assembly of the reads results in the first version of its reference genome. This is called "**de-novo assembly**".

Multiplex is a library containing various samples labelled with bar codes.

Sample **multiplexing** is a useful technique when targeting specific genomic regions or working with smaller genomes. To accomplish this, individual **"barcode"** sequences are added to each sample so they can be distinguished and sorted during data analysis. **Pooling** samples exponentially increases the number of samples analyzed in a single run, without drastically increasing cost or time.

Di-multiplexing is separating samples based on their tags, ideally should be even across tags.

Adapter. The vast majority of next-generation sequencing experiments will attach adapter sequence to the sequencing construct. In many cases these are standard sequences that can be obtained from the vendor and/or sequencing centre. Unfortunately sometimes adapter information is not properly tracked and attached as metadata to the raw sequencing data and may not be known for a given sample.

PF (**purity-filtered**) data: PF-filtering is known as throwing away data with low maximum intensity signal (purity, Illumina terminology).

GC-content is a measure of the relative frequency of the cytosine (C) and guanine (G) bases, in comparison with the adenine (A) and thymine (T) bases. A genome is called **GC-rich** if significantly more than 50 % of its bases are G or C.

Mate-pair libraries. *Mate-pair* is different from "*paired-end*" in the sense of how the sequence library is made. In "Mate-pair" sequencing, **2–5 kb** fragments are selected and sequenced from both end, thus giving information how nucleotides far apart are linked together. Mate-pairs are more ideal for studying genomic structural rearrangement and help de novo genome assembly. They also facilitate sensitive structural variant (SV) detection across a widened SV size-spectrum and in repetitive areas of the genome.

Insert size = DNA fragment size.

Ti/Tv (sometimes called Ts/Tv): the ratio of transitions vs. transversions in SNPs. Transitions are mutations within the same type of nucleotide: pyrimidine-pyrimidine mutations (C <–> T) and purine-purine mutations (A <–> G). Transversions are mutations from a pyrimidine to a purine or vice versa.

The heterozygosity ratio is the number of heterozygous sites in an individual divided by the number of non-reference homozygous sites.

Error-correction is an attempt to correct a mismatch between sequenced reads and/or reference (if it is available).

Genomic variant or mutation is a permanent alteration of the nucleotide sequence of the genome of an organism.

A **single nucleotide polymorphism** or simple nucleotide polymorphism, (SNP), is a variation in a single nucleotide which may occur at some specific position in the genome, where each variation is present to some appreciable degree within a population (e. g. >1 %).

Structural variation (also genomic structural variation) is the variation in structure of an organism's chromosome.

Structural variation consists of many kinds of variation in the genome of one species, and usually includes microscopic and submicroscopic types, such as deletions, duplications, copy-number variants, insertions, inversions and translocations.

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule.

WGS – whole genome sequencing.

WES – whole exome sequencing.

In Illumina, **PCR and size selection** steps have been implicated in GC-bias. PCR is known to preferentially amplify GC-moderate sequences, while size selection involves DNA heating which leads to a GC-poor fragment's underrepresentation. Avoiding these steps helps to limit the GC-bias.

BAM File – binary version of SAM file, a typical output of the secondary phase of data analysis.

Coverage – this value indicates the coverage of an analysed sequence with respect to its length, usually expressed as a percentage; sometimes the term is also used for the depth of reading.

Long-Reads – strategy for sequencing samples prepared by Mate-Pair-End method.

Mate Pair-End-Read – strategy for sample preparation where the longer fragment (thousands of bases) is circularized using labelled adapters, the molecule is subsequently fragmented, but only the fragments containing the labelled adapters are sequenced.

Paired-End-Read – a method of reading a fragment where the fragment is first read from one end and then from the other.

Read Depth – DNA = number of times a nucleotide is read; RNA = total number of reads per sample.

Read Length – the number of read bases per fragment, respectively the maximum length of the fragment, which can be sequenced at a time (indicated in bases).

Single-Read – a method of reading a fragment where the fragment is read from one end only during sequencing.

SNP – Single-Nucleotide Polymorphism = sequence divergence in the range of a single base.

SNP Calling – process of detecting SNPs in the sequences obtained.

Variant Calling is a process of detection of sequence variants in the sequences obtained.

Heterozygosity occurs when an individual has two different alleles of a gene/loci.

Chimeric reads are reads with DNA sequences originating from two different samples.

Conflict of interest

The authors declare no conflict of interest.

References

Abnizova I., Leonard S., Skelly T., Brown A., Jackson D., Gourtovaia M., Qi G., Te Boekhorst R., Faruque N., Lewis K., Cox T. Analysis of context-dependent errors for illumina sequencing. J. Bioinform. Comput. Biol. 2012;10(2):1241005.

Abnizova I., Skelly T., Naumenko F., Whiteford N., Brown C., Cox T. Statistical comparison of methods to estimate the error probability in short-read Illumina sequencing. J. Bioinform. Comput. Biol. 2010;8(3):579-591.

Albers C.A., Lunter G., MacArthur D.G., McVean G., Ouwehand W.H., Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011;21(6):961-973.

Alkan C., Cardone M.F., Catacchio C.R., Antonacci F., O'Brien S.J., Ryder O.A., Purgato S., Zoli M., Della Valle G., Eichler E.E., Ventura M. Genome-wide characterization of centromeric satellites from multiple mammalian genomes. Genome Res. 2011a;21(1):137-145.

Alkan C., Sajjadian S., Eichler E.E. Limitations of next-generation genome sequence assembly. Nat. Methods. 2011b;8(1):61-65.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. J. Mol. Biol. 1990;215(3):403-410.

Anders S., Pyl P.T., Huber W. HTSeq – a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-169.

Auerbach R.K., Euskirchen G., Rozowsky J., Lamarre-Vincent N., Moqtaderi Z., Lefrancois P., Struhl K., Gerstein M., Snyder M. Mapping accessible chromatin regions using Sono-Seq. Proc. Natl. Acad. Sci. USA. 2009;106(35):14926-14931.

Baker M. De novo genome assembly: what every biologist should know. Nature Methods. 2012;9.

Balint B. Decreased sequencing accuracy at the 3′ end of SBS Illumina Reads. 2016.

Benjamini Y., Speed T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012; 40(10):e72.

Bonfield J.K., Staden R. The application of numerical estimates of base calling accuracy to DNA sequencing projects. Nucleic Acids Res. 1995;23(8):1406-1410.

Bragg L.M., Stone G., Butler M.K., Hugenholtz P., Tyson G.W. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput. Biol. 2013;9(4):e1003031.

Brockman W., Alvarez P., Young S., Garber M., Giannoukos G., Lee W.L., Russ C., Lander E.S., Nusbaum C., Jaffe D.B. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res. 2008;18(5):763-770.

Chen Y.C., Liu T., Yu C.H., Chiang T.Y., Hwang C.C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLoS One. 2013; 8(4):e62856.

Chin E.L.H., da Silva C., Hegde M. Assessment of clinical analytical sensitivity and specificity of next-generation sequencing for detection of simple and complex mutations. BMC Genet. 2013;14:6.

Chin F.Y.L., Leung H.C.M., Yiu S.M. Sequence assembly using next generation sequencing data – challenges and solutions. Sci. China Life Sciences. 2014;57(11):1140-1148.

Cox M.P., Peterson D.A., Biggs P.J. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics. 2010;11:485.

Davis M.P., van Dongen S., Abreu-Goodger C., Bartonicek N., Enright A.J. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods. 2013;63(1):41-49.

Day-Williams A.G., Zeggini E. The effect of next-generation sequencing technology on complex trait research. Eur. J. Clin. Invest. 2011; 41(5):561-567.

de Koning A.P., Gu W., Castoe T.A., Batzer M.A., Pollock D.D. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011;7(12):e1002384.

Del Fabbro C., Scalabrin S., Morgante M., Giorgi F.M. An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. 2013;8(12):e85024.

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennel T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011;43(5):491-498.

Dolled-Filhart M.P., Lee M., Jr., Ou-Yang C.W., Haraksingh R.R., Lin J.C. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. Sci. World J. 2013:730210.

Edgar R.C., Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics. 2015; 31(21):3476-3482.

Ewing B., Hillier L., Wendl M.C., Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8(3):175-185.

Feschotte C., Jiang N., Wessler S.R. Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. 2002;3(5):329-341.

Fonseca N.A., Rung J., Brazma A., Marioni J.C. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012;28(24):3169-3177.

Fujimoto M., Bodily P.M., Okuda N., Clement M.J., Snell Q. Effects of error-correction of heterozygous next-generation sequencing data. BMC Bioinformatics. 2014;15(Suppl.7):S3.

Garcia-Garcia G., Baux D., Faugere V., Moclyn M., Koenig M., Claustres M., Roux A.F. Assessment of the latest NGS enrichment capture methods in clinical context. Sci. Rep. 2016;6:20948.

Guo Y., Ye F., Sheng Q., Clark T., Samuels D.C. Three-stage quality control strategies for DNA re-sequencing data. Brief Bioinform. 2014a;15(6):879-889.

Guo Y., Zhao S., Sheng Q., Ye F., Li J., Lehmann B., Pietenpol J., Samuels D.C., Shyr Y. Multi-perspective quality control of Illumina exome sequencing data using QC3. Genomics. 2014b;103(5-6):323-328.

Hadfield J. Quality control for your NGS data. 2013.

Harismendy O., Ng P.C., Strausberg R.L., Wang X., Stockwell T.B., Beeson K.Y., Schork N.J., Murray S.S., Topol E.J., Levy S., Frazer K.A. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10(3):R32.

Huang Y.F., Chen S.C., Chiang Y.S., Chen T.H., Chiu K.P. Palindromic sequence impedes sequencing-by-ligation mechanism. BMC Syst. Biol. 2012;6(Suppl.2):S10.

Huddleston J., Ranade S., Malig M., Antonacci F., Chaisson M., Hon L., Sudmant P.H., Graves T.A., Alkan C., Dennis M.Y., Wilson R.K., Turner S.W., Korlach J., Eichler E.E. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;24(4):688-696.

IDT. 2011. G-quenching.

Ignatieva E.V., Podkolodnaya O.A., Orlov Y.L., Vasiliev G.V., Kolchanov N.A. Regulatory genomics: Integrated experimental and computer approaches. Genetika. 2015;51(4):409-429.

Illumina. 2014. Sequencing Library QC on the MiSeq® System.

Janin L., Schulz-Trieglaff O., Cox A.J. BEETL-fastq: a searchable compressed archive for DNA reads. Bioinformatics. 2014;30(19):2796-2801.

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

Jun G., Wing M.K., Abecasis G.R., Kang H.M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015;25(6):918-925.

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

Kojima K., Nariai N., Mimori T., Takahashi M., Yamaguchi-Kabata Y., Sato Y., Nagasaki M. A statistical variant calling approach from pedigree information and local haplotyping with phase informative reads. Bioinformatics. 2013;29(22):2835-2843.

Lander E.S. Initial impact of the sequencing of the human genome. Nature. 2011;470(7333):187-197.

Lawrence M. Introduction to variant calling. Bioconductor. 2014.

Ledergerber C., Dessimoz C. Base-calling for next-generation sequencing platforms. Brief Bioinform. 2011;12(5):489-497.

Lettice L.A., Hill A.E., Devenney P.S., Hill R.E. Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Hum. Mol. Genet. 2008;17(7):978-985.

Li H., Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-595.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079.

Li H., Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473-483.

Li J.W., Robison K., Martin M., Sjodin A., Usadel B., Young M., Olivares E.C., Bolser D.M. The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. Nucleic Acids Res. 2012a;40(Database iss.):D1313-1317.

Li J.W., Schmieder R., Ward R.M., Delenick J., Olivares E.C., Mittelman D. SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics. 2012b;28(9):1272-1273.

Li M.K., Stoneking M. A new approach for detecting low-level mutations in next-generation sequence data. Genome Biology. 2012;13(5).

Li R., Yu C., Li Y., Lam T.W., Yiu S.M., Kristiansen K., Wang J. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-1967.

Li S., Li R., Li H., Lu J., Li Y., Bolund L., Schierup M.H., Wang J. SOAPindel: efficient identification of indels from short paired reads. Genome Res. 2013;23(1):195-200.

Li Z., Chen Y., Mu D., Yuan J., Shi Y., Zhang H., Gan J., Li N., Hu X., Liu B., Yang B., Fan W. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct. Genomics. 2012;11(1):25-37.

Lin K., Smit S., Bonnema G., Sanchez-Perez G., de Ridder D. Making the difference: integrating structural variation detection tools. Brief Bioinform. 2015;16(5):852-864.

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., Law M. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012;251364.

Lo C.C., Chain P.S. Rapid evaluation and quality control of next generation sequencing data with FaQCs. BMC Bioinformatics. 2014;15:366.

Magoc T., Pabinger S., Canzar S., Liu X., Su Q., Puiu D., Tallon L.J., Salzberg S.L. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics. 2013;29(14):1718-1725.

Makinen V., Salmela L., Ylinen J. Normalized N50 assembly metric using gap-restricted co-linear chaining. BMC Bioinformatics. 2012; 13:255.

Marian A.J. Molecular genetic studies of complex phenotypes. Transl. Res. 2012;159(2):64-79.

Marroni F., Pinosio S., Morgante M. The quest for rare variants: pooled multiplexed next generation sequencing in plants. Front Plant. Sci. 2012;3:33.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal. 2011;17(1).

Massingham T., Goldman N. All your Base: a fast and accurate probabilistic approach to base calling. Genome Biology. 2012;13(2).

McCoy R.C., Taylor R.W., Blauwkamp T.A., Kelley J.L., Kertesz M., Pushkarev D., Petrov D.A., Fiston-Lavier A.S. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS One. 2014;9(9):e106689.

Medvedev P., Stanciu M., Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods. 2009;6(Suppl.11):S13-20.

Miller J.R., Koren S., Sutton G. Assembly algorithms for next-generation sequencing data. Genomics. 2010;95(6):315-327.

Minoche A.E., Dohm J.C., Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 2011;12(11):R112.

Mir K., Neuhaus K., Bossert M., Schober S. Short barcodes for next generation sequencing. PLoS One. 2013;8(12):e82933.

Mutarelli M., Marwah V., Rispoli R., Carrella D., Dharmalingam G., Oliva G., di Bernardo D. A community-based resource for automatic exome variant-calling and annotation in Mendelian disorders. BMC Genomics. 2014;15(Suppl.3):S5.

Nagarajan N., Pop M. Sequence assembly demystified. Nat. Rev. Genet. 2013;14(3):157-167.

Nakamura K., Oshima T., Morimoto T., Ikeda S., Yoshikawa H., Shiwa Y., Ishikawa S., Linak M.C., Hirai A., Takahashi H., Altaf-Ul-Amin Md., Ogasawara N., Kanaya S. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39(13):e90.

Newell F. NGS mapping, errors and quality control. Australia: Univ. of Queensland, 2014.

Nielsen C.B., Cantor M., Dubchak I., Gordon D., Wang T. Visualizing genomes: techniques and challenges. Nat. Methods. 2010; 7(Suppl.3):S5-S15.

Niu B., Fu L., Sun S., Li W. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics. 2010; 11:187.

Olson N.D., Lund S.P., Colman R.E., Foster J.T., Sahl J.W., Schupp J.M., Keim P., Morrow J.B., Salit M.L., Zook J.M. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. Front. Genet. 2015;6:235.

Orlov Y.L., Te Boekhorst R., Abnizova I.I. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J. Bioinform. Comput. Biol. 2006;4:523-36.

Orton R.J., Wright C.F., Morelli M.J., King D.J., Paton D.J., King D.P., Haydon D.T. Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. BMC Genomics. 2015;16:229.

Otto C., Stadler P.F., Hoffmann S. Lacking alignments? The next-generation sequencing mapper segemehl revisited. Bioinformatics. 2014;30(13):1837-1843.

Pabinger S., Dander A., Fischer M., Snajder R., Sperk M., Efremova M., Krabichler B., Speicher M.R., Zschocke J., Trajanoski Z. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014;15(2):256-278.

Park N., Shirley L., Gu Y., Keane T.M., Swerdlow H., Quail M.A. An improved approach to mate-paired library preparation for Illumina sequencing. Methods Next-Generation Sequencing. 2013;1(1): 10-20.

Patel R.K., Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7(2):e30619.

Patro R., Kingsford C. Data-dependent bucketing improves reference-free compression of sequencing reads. Bioinformatics. 2015;31(17): 2770-2777.

Pightling A.W., Petronella N., Pagotto F. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. PLoS One. 2014;9(8):e104579.

Pireddu L., Leo S., Zanetti G. SEAL: a distributed short read mapping and duplicate removal tool. Bioinformatics. 2011;27(15):2159-2160.

Rieber N., Zapatka M., Lasitschka B., Jones D., Northcott P., Hutter B., Jäger N., Kool M., Taylor M., Lichter P., Pfister S., Wolf S., Brors B., Eils R. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. PLoS One. 2013; 8(6):e66621.

Ross M.G., Russ C., Costello M., Hollinger A., Lennon N.J., Hegarty R., Nusbaum C., Jaffe D.B. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.

Safronova N.S., Babenko V.N., Orlov Y.L. 117 analysis of SNP containing sites in human genome using text complexity estimates. J. Biomol. Structure Dynamics. 2015;33(Supp1.):73-74. DOI 10.1080/07391102.2015.1032750.

Safronova N.S., Ponomarenko M.P., Abnizova I.I., Orlova G.V., Chadaeva I.V., Orlov Y.L. Flanking monomer repeats determine decreased context complexity of single nucleotide polymorphism sites in the human genome. Rus. J. Genet. Appl. Res. 2016;6(8):809-815. DOI 10.1134/S2079059716070121.

Sameith K., Roscito J.G., Hiller M. Iterative error correction of long sequencing reads maximizes accuracy and improves contig assembly. Brief Bioinform. 2016:1-8.

Sanger F., Nicklen S., Coulson A.R. DNA sequencing with chain-terminating inhibitors. (1977). Biotechnology. 1992;24:104-108.

Schadt E.E., Turner S., Kasarskis A. A window into third-generation sequencing. Hum. Mol. Genet. 2010;19(R2):R227-240.

Schmieder R., Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6(3):e17288.

Schwartz S., Oren R., Ast G. Detection and removal of biases in the analysis of next-generation sequencing reads. PLoS One. 2011;6(1): e16685.

Shang J., Zhu F., Vongsangnak W., Tang Y., Zhang W., Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. Biomed. Res. Int. 2014:309650.

Sharon D., Tilgner H., Grubert F., Snyder M. A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. 2013; 31(11):1009-1014.

Stitziel N.O., Kiezun A., Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol. 2011;12(9):227.

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

van Dijk E.L., Auger H., Jaszczyszyn Y., Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418-426.

Vissers L.E., Fano V., Martinelli D., Campos-Xavier B., Barbuti D., Cho T.J., Dursun A., Kim O.H., Lee S.H., Timpani G., Nishimura G., Unger S., Sass J.O., Veltman J.A., Brunner H.G., Bonafe L., Dionisi-Vici S., Superti-Furga A. Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). Am. J. Med. Genet. A. 2011;155A(11):2609-2616.

Voskoboynik A., Neff N.F., Sahoo D., Newman A.M., Pushkarev D., Koh W., Passarelli B., Fan H.C., Mantalas G.L., Palmeri K.J., Ishizuka K.J., Gissi C., Griggio F., Ben-Shlomo R., Corey D.M., Penland L., White R.A.III, Weissman I.L., Quake S.R. The genome sequence of the colonial chordate, *Botryllus schlosseri*. ELife. 2013;2: e00569.

Walther A., Johnstone E., Swanton C., Midgley R., Tomlinson I., Kerr D. Genetic prognostic and predictive markers in colorectal cancer. Nat. Rev. Cancer. 2009;9(7):489-499.

Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

Wolfinger M.T., Fallmann J., Eggenhofer F., Amman F. ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines. F1000Res. 2015;4:50.

Wong K., Keane T.M., Stalker J., Adams D.J. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. Genome Biology. 2010;11(12).

Wyllie M. Comprehensive analysis of clinical trials data shows unequivocally that Phosphodiesterase Inhibitors (PDEi) improve orgasm. The power of meta-analysis? BJU Int. 2013;111(2): 190-191.

Yan X.J., Xu J., Gu Z.H., Pan C.M., Lu G., Shen Y., Shi J.Y., Zhu Y.M., Tang L., Zhang X.W., Liang W.-X., Mi J.-Q., Song H.-D., Li K.-Q., Chen Z., Chen S.-J. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat. Genet. 2011;43(4):309-315.

Yang H., Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc. 2015;10(10):1556-1566.

Yang X., Chockalingam S.P., Aluru S. A survey of error-correction methods for next-generation sequencing. Brief Bioinform. 2013; 14(1):56-66.

Ye L., Hillier L.W., Minx P., Thane N., Locke D.P., Martin J.C., Chen L., Mitreva M., Miller J.R., Haub K.V., Dooling D.J., Mardis E.R., Wilson R.K., Weinstock G.M., Warren W.C. A vertebrate case study of the quality of assemblies derived from next-generation sequences. Genome Biol. 2011;12(3):R31.

Zhang W., Ng H.W., Shu M., Luo H., Su Z., Ge W., Perkins R., Tong W., Hong H. Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium. J. Genetics. 2015;94(4):731-740.