Demystification of RNAseq Quality Control

Dragana Dudić[1], Bojana Banović Đeri[2], Vesna Pajić[3] and Gordana Pavlović-Lažetić[4]

[1]Faculty of Informatics and Computer Science, University Union Nikola Tesla, Belgrade, Serbia, ddudic@unionnikolatesla.edu.rs

[2]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia, bojanabanovic@imgge.bg.ac.rs

[3]Seven Bridges Genomics, Belgrade, Serbia, vesna.pajic@gmail.com

[4]Faculty of Mathematics, University of Belgrade, Serbia, gordana@matf.bg.ac.rs

Contribution to the State of the Art

https://doi.org/10.7251/JIT2102073D UDC: 004.62.032.26:577.216.3

Abstract: Next Generation Sequencing (NGS) analysis has become a widely used method for studying the structure of DNA and RNA, but complexity of the procedure leads to obtaining error-prone datasets which need to be cleansed in order to avoid misinterpretation of data. We address the usage and proper interpretations of characteristic metrics for RNA sequencing (RNAseq) quality control, implemented in and reported by FastQC, and provide a comprehensive guidance for their assessment in the context of total RNAseq quality control of Illumina raw reads. Additionally, we give recommendations how to adequately perform the quality control preprocessing step of raw total RNAseq Illumina reads according to the obtained results of the quality control evaluation step; the aim is to provide the best dataset to downstream analysis, rather than to get better FastQC results. We also tested effects of different preprocessing approaches to the downstream analysis and recommended the most suitable approach.

Keywords: data preprocessing, Illumina sequencing, NGS analysis, quality control, sequence analysis, total RNAseq.

Introduction

High throughput sequencing technologies provide a way of studying the structure of genetic material, both DNA and RNA. Complexity of the sequencing procedures results in error-prone data sets, which need to be properly treated in order to obtain relevant results from downstream analyses. This process is more complex in the case of RNA sequencing (RNAseq), as it includes additional step of reverse transcription of RNA molecules to complementary DNA (cDNA), after which the common steps (as for DNA sequencing) are performed: amplification, fragmentation, purification, adaptor ligation, and sequencing. On the other hand, RNAseq today is intensively used for a number of analyses, such as characterization of transcriptional activity, quantification of gene expression, differential gene expression, analysis of alternative splicing, functional analysis, gene fusion detection, etc. Because of demands for the greatest possible reliability of data that will be used in such analysis, it is of the highest importance to estimate the quality of obtained reads and how the quality of reads will or could affect final results of the analysis [1]. In many cases, it is better to omit low quality reads from further processing than to cause the misinterpretation of data.

Quality control consists of two steps: evaluation and preprocessing. The evaluation step consists of a number of metrics which indicate quality of assessed raw reads, while the preprocessing step includes sequence filtering according to results of the evaluation step. Since there is a plethora of available tools, it is challenging for new users to appropriately choose between these tools and to adapt to new ones. Some of them are general tools, meant for both DNA and RNAseq raw reads, including both steps of

quality control like NGS QC Toolkit, FASTX-Toolkit, PRINSEQ, QC-Chain, FaQCs, HTQC and others are designed just for one of the tasks. While there are plenty of tools for preprocessing available, i.e. Trimmomatic [2], SolexaQA, Trim Galore!, Cutadapt, DeconSeq, ConDeTri, Sicle, Scythe, Seqtk, SortMeRNA, BBDuk and BBSplit, VecScreen, Kraken, none of them performs all the preprocessing tasks. On the other hand, there are just several tools designed for the purpose of evaluation of raw reads (FastQC [3], seqTools, fastqp) that provide standard metrics for estimation of raw data quality. Discovery of contaminant sequences (sequences that originated from other organisms than the one that was sequenced) demands specific evaluation tools and they include tools like FastQ Screen and VecScreen. Because it is necessary for quality preprocessing and further steps of downstream RNAseq analysis to estimate error probability in raw reads, we focus on raw reads quality evaluation. The vast majority of researchers today use FastQC [3] in combination with some of the tools specially designed for filtering NGS data. However, FastQC operates by creating flags on datasets based on several metrics and their expected values in the case of DNA experiments which is often misleading for RNAseq experiments. Additionally, special attention should be paid in the preprocessing step in order to enable the assessment of reliable results for further analysis of RNAseq data because the stringent approach, which is widely used, might not be the best choice for transcriptome data.

In this paper, we address the usage and proper interpretation of metrics for RNAseq quality control, implemented in and reported by FastQC, such as Per base sequence content, Per sequence GC content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, and Kmer Content and provide a comprehensive guidance for their assessment in the context of total RNAseq quality control of Illumina raw reads. Also, we give recommendations on how to adequately perform the preprocessing step of raw total RNAseq Illumina reads according to the obtained results of the evaluation step, with an aim not to get better FastQC results, but to provide the best dataset to downstream analysis.

QC evaluation

Evaluation of raw reads is the first step in sequencing analysis and serves to determine the validity of sequenced data. One of the mostly used tools for quality control is FastQC, developed by the Babraham Institute in Cambridge. It is included in many bioinformatics software (Galaxy, Illumina BaseSpace, GenePattern, Chipster, Yabi, Taverna, KNIME, Tavaxy, BioDT), and in that way it became a sort of a standard tool for evaluation of NGS data.

This tool takes fastq, SAM or BAM file as input, quickly performs quality analysis of provided data and outputs results in html format. Quality analysis consists of 12 metrics (Basic Statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, Sequence Length Distribution, Sequence Duplication Levels, Overrepresented sequences, Adapter Content, Kmer Content). Most of the metrics are represented graphically, only Basic Statistics and Overrepresented sequences are presented in tabular format. The exception is the Kmer content module, which is shown in both ways (graphically and in tabular format). For each metric a section in the resulting report is made, flagged as 'PASSED', 'WARNING' or 'FAILED'. The flags are given based on the expected values for sequencing DNA data. At FastQC download page examples of good and bad data for different sequencing platforms are available. Illumina platform is widely used for conducting RNAseq experiments, and for Illumina data the FastQC author suggests that a good report should be mostly flagged as PASSED - only one checkpoint (Kmer content) is flagged as WARNING. However, since the FastQC tool expects diversity and randomness in data even a slight deviation will issue a warning, while severe one will result in failure. Because the correct interpretation of the FastQC report is crucial for other steps of the analysis, flags provided by the program itself cannot be taken for granted. Although the checkpoints like Basic Statistics, Per base sequence quality, Per tile sequence quality, Per sequence quality scores, Per base N content and Adapter Content have universal interpretation and high quality data should pass all of them, other checkpoints are specific for different types of sequencing experiments (DNAseq or RNAseq). All of the specific checkpoints

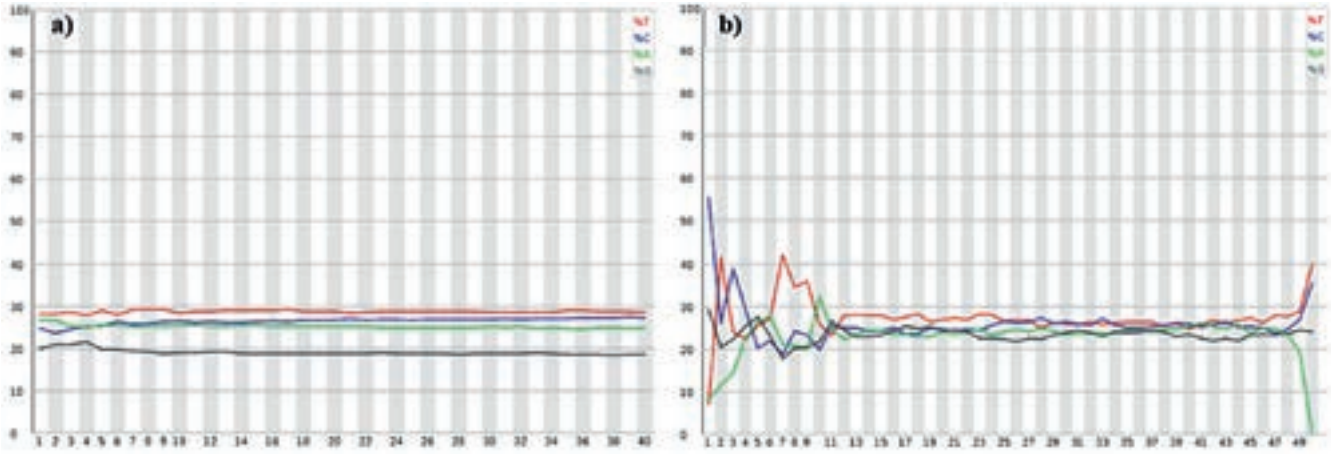

Figure 1. *Per base sequence content graphs showing the read position on X-axis and percentage of each base on Y-axis, a) DNA-seq Illumina data - passed b) total RNAseq Illumina data - failed*

will be discussed briefly in the context of Illumina RNAseq experiments evaluated with FastQC v0.11.5.

Per base sequence content

This checkpoint shows the percentage of each base in each position and it is represented with a graph. According to FastQC interpretation little to no difference is expected between representations of bases, presented as four smooth lines around 25% (Fig. 1a). Warning is issued if the difference between base pairs is greater than 10% in any position, while failure is shown if this difference is greater than 20% (Fig. 1b). Analysis of the most of Illumina RNAseq data issues a failure in this checkpoint, usually not because of any of the common reasons. As stated in help files, but not pointed out in results, the main reason of such failure is the way in which RNAseq libraries are produced (by priming using random hexamers, which is why the bias at the start positions is expected). This is particularly the characteristic of the first 10-13 positions at the 5′-end Illumina RNAseq data and read count reweighting scheme is proposed in order to reduce the impact of this bias [5]. Alternative approach would be to use oligo(dt) priming, but the same study showed that, in this case, data would be highly biased toward 3'-end, and that bias cannot be easily mitigated. Presence of polyA/T tails is another cause of bias at the 3'-end, unrelated with the way of priming. Also, we noted that the abscissa of the representation plot is not equally divided; first nine positions are given separately for each data, and the longer the read the more distant other points become, and sometimes even given as ranges. For long reads this gives a false impression that the data at the start positions are more deviated.

All Illumina RNAseq data will issue a warning or failure in the Per base sequence content module. The choice of further steps depends on the main goal of sequencing and the nature of available referent resources for the organism under investigation. If de novo assembling is in plan, it is advised to remove first 10-13 bases, while in other cases these bases should be retained.

Per sequence GC content

This plot shows the GC content of each sequence for each position compared to the modeled normal distribution, because it is expected that the random library has a nearly normal distribution. It is misleading to interpret the modeled distribution as a curve that shows information from the reference genome or transcriptome of the sequenced organism. This is just a Gaussian distribution parameterized according to the mean and variance of the GC content of the provided reads. Reads will pass this checkpoint if the GC content curve does not deviate too much from the modeled distribution (Fig. 2a). If sequences outside of the normal distribution comprise more than 15% of the total, FastQC will raise a warning, while failure is given if these reads comprise more than 30% of the total (Fig. 2b).

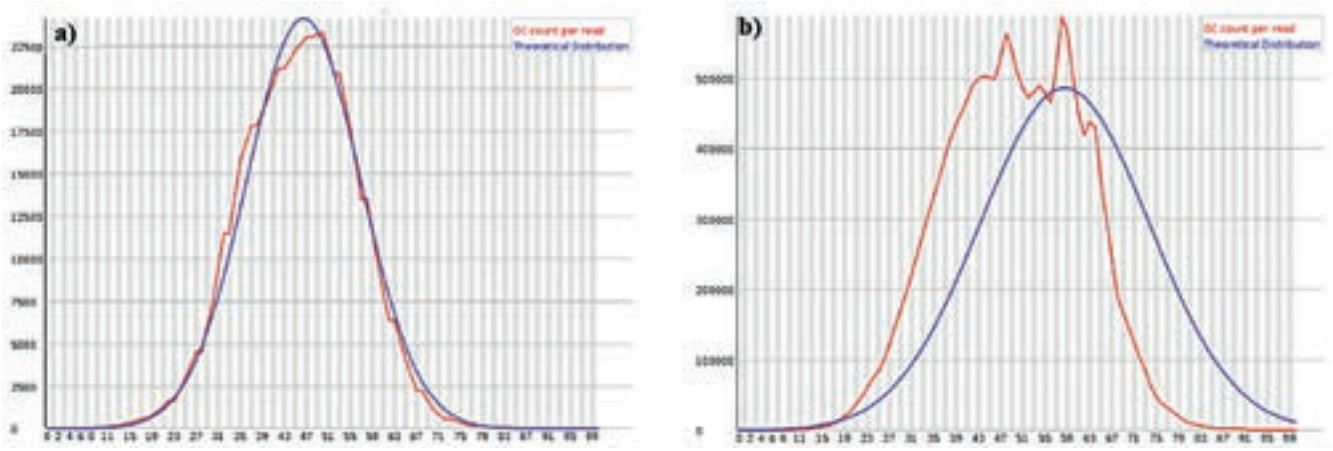

Figure 2. *Per sequence GC content shows the mean GC content on X-axis, while the number of reads is shown on Y-axis; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-failed*

Different types of deviation have different manifestations. If the plot is shifted in comparison with the modeled distribution, the reason is systematic bias and FastQC will not flag this as an error. Another reason for the shifted distribution could be the presence of polyA/T tails (when shifted to the left) and rRNA richness (when shifted to the right) in data [6]. On the other side, the same reasons could manifest like shoulders or peaks in the distribution plot, only in that case it will be reported as error, because any distribution that has peaks and/or is not unimodal is by default addressed as an error. In the distribution plot, different kinds of contamination (adapter dimers and sequences from other organisms) are usually represented with peaks. The shape of the distribution curve is also affected by the short reads, characteristic for total RNAseq reads, and duplicate-rich reads, which are characteristic for all RNAseq experiments. Additionally, in RNAseq data for organisms with highly repetitive genomes GC content varies due to the presence of some classes of transposons which are GC rich [7][8]. As a consequence of aforementioned, GC content plot could take some form of bimodal distribution. These flaws affect many metrics used in QC and a common way of dealing with these flaws of RNAseq data is to perform, after removing contamination, de novo assembling of reads into larger contigs which should mitigate or even completely eliminate the second peak and smooth the main peak.

Sequence Length Distribution

For sequence length distribution it is expected to be uniform (Fig. 3a). If the read length is variable, warning is issued (Fig. 3b), and failure is caused due to the presence of sequences of zero length.

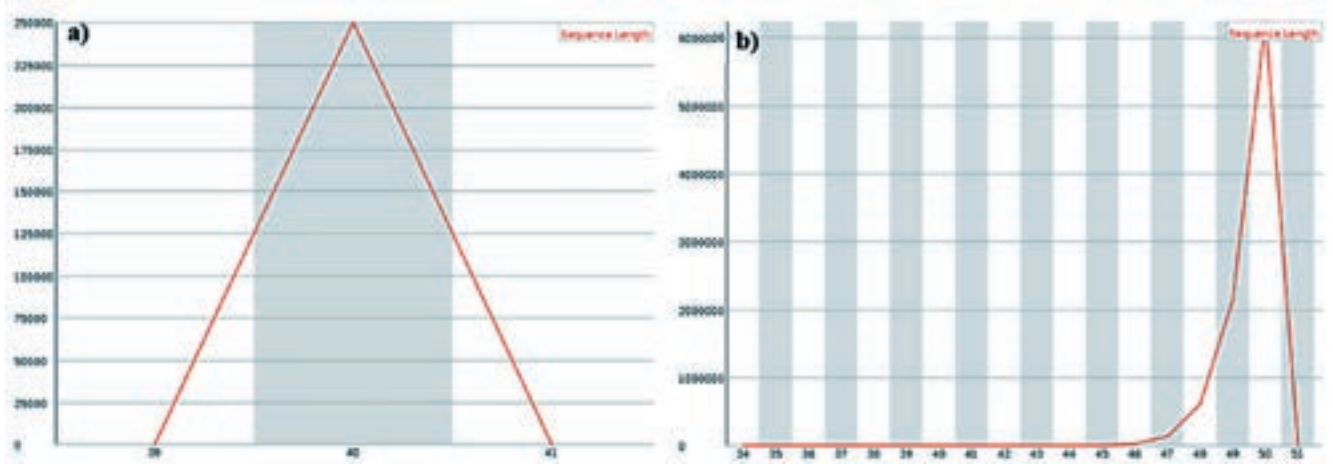

Figure 3. *Sequence length distribution shows different sequence lengths on X-axis and Y-axis presents number of reads; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-warning*

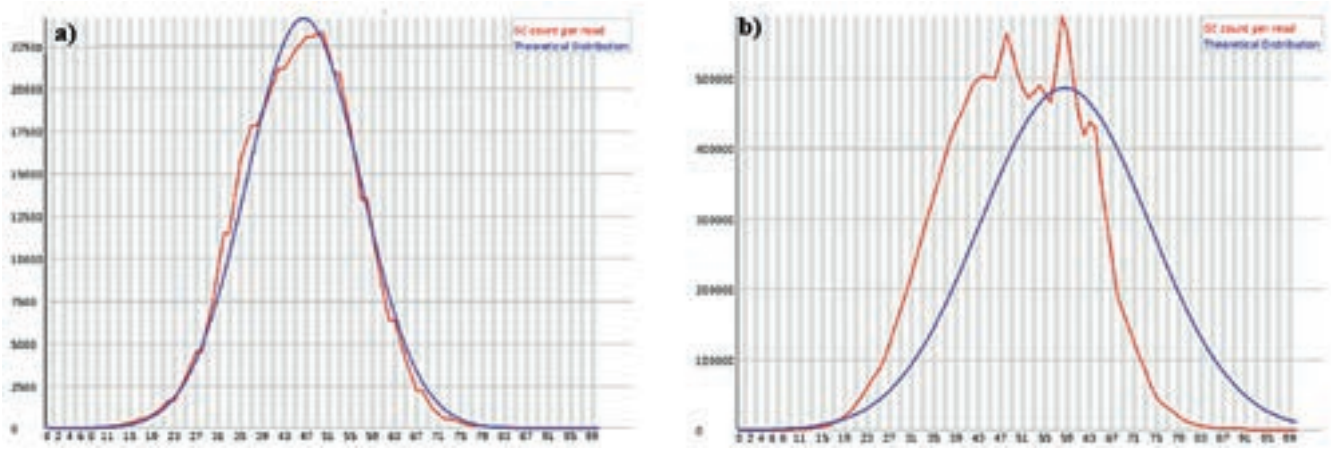

Figure 4. *Sequence duplication level shows different degrees of duplication on X-axis and percentage of the duplication is shown on Y-axis; a) DNA-seq Illumina data-passed; b) total RNAseq Illumina data-failed*

The Help files specify that for some high throughput sequencers it is usual to have different read lengths although this will give a warning. For Illumina it is expected to have the equal length for all reads, but in the case of total RNAseq, some RNA molecules are shorter (like small and microRNA), which is why warnings in this module can be ignored.

Sequence Duplication Levels

This checkpoint estimates the degree of duplication using an exact sequence match in a specific way. Common duplicates include PCR and optical duplicates. PCR duplication can arise in transcriptome data because of the specific way of sample preparation and optical duplicates which are the consequence of reading the same cluster twice or more times on sequencer.

Due to reducing memory usage, only sequences that firstly occur in the first 100,000 sequences are taken into account. Additionally, sequences longer than 75 bp are truncated to 50bp. The aforementioned optimizations should provide a good estimate of the data, but this is not the case for RNAseq reads. RNAseq reads tend to be several times longer. Because the transcriptome data include products of alternative splicing, it is possible that different RNAseq reads have some of the first bases the same and the rest of the sequence different. Also, in RNAseq data we expect a plethora of highly expressed sequences and for highly expressed sequences it is expected to have multiple occurrences in the dataset. Regardless, if duplicate sequences make up more than 20% of the sequences taken into account, FastQC will issue a warning, and if there is more than 50% of the duplicate sequences, this checkpoint will raise an error (Fig. 4b).

In section "Common reasons for warnings" of FastQC Help files it is indicated that warnings for RNAseq data could arise because of the nature of the experiment. We state, based on the aforementioned, that if PCR duplicates and contaminants problems are excluded, any kind of error in this module does not affect the overall quality of the RNAseq data even if duplication level is higher than 50%.

Overrepresented sequences

All sequences that represent more than 0.1% of total sequences are labeled as overrepresented and presented in a table with count, percentage and possible source of the sequence (Fig. 5). Warning will be issued if at least one overrepresented sequence is found. A failure will arise if there is at least one sequence that represents more than 1% of the total data.

To conserve memory, this checkpoint looks for candidates for overrepresented sequences in the first 100,000 sequences, but candidates are tracked through the whole dataset, therefore more overrepresented sequences could be found in the dataset. Databases of known contaminants are queried by overrepresented sequences that were found and best hits are being reported. Since Illumina RNAseq library preparation includes adapters (that are usually removed, but sometimes some of them remain in the sequences), this module will label them and they should be removed from the data.

Sequence	Count	Percentage	Possible Source
CTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGGT	57399	0.6247347688043816	No Hit
CCCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATAT	37086	0.40364664255264543	No Hit
CCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGATG	36448	0.3967026055050105	No Hit
CCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	31942	0.34765898334726314	No Hit
CTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACCTT	30483	0.3317791243308065 3	No Hit
CTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTG	29650	0.3227126935146939	No Hit
CTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	28330	0.30834572031269064	No Hit
CCCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATT	28306	0.30808450261810877	No Hit
CCTCCTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTG	24959	0.2716555182945445	No Hit
CAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATC	23543	0.25624367431421374	No Hit
GGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGATCAGC	16550	0.18013136855541934	No Hit
CTGCTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGT	14718	0.16019175120233609	No Hit
GTCTGGAGTCTTGGAAGCTTGACTACCCTACGTTCTCCTACAAATGGACC	14691	0.15989788129593144	No Hit
GCTCCGTTTCCGACCTGGGCCGGTTCACCCCTCCTTAGGCAACCTGGTGG	14422	0.15697006630249294	No Hit
CTTAGGCAACCTGGTGGTCCCCCGCTCCCGGGAGGTCACCATATTGATGC	13736	0.14950359386569428	No Hit
CTCGCTATGTTGCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATC	13654	0.14861110007587286	No Hit
GCTCAGGCTGGAGTGCAGTGGCTATTCACAGGCGCGATCCCACTACTGAT	11675	0.12707182434347557	No Hit

Figure 5 *Overrepresented sequences in totalRNAseq data - warning*

Kmer Content

In total RNA sequencing a lot of different types of molecules are sequenced. Because of the presence of small RNA, mitochondrial and chloroplastic (in case of plants) RNA, residual rRNA (mostly removed during a library preparation, but some rRNA may still remain), overrepresentation of the sequences is expected to some extent. Also, alternative splicing and differential gene expression affect the number of overrepresented sequences. Taking into account the aforementioned, we conclude that warnings and failures can be ignored as long as adapters are trimmed.

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
AAGAACG	2050	0.0	24.986587	43
TTCGGCG	420	0.0	24.305752	44
AGAGCGG	2055	0.0	24.061537	4
GACCGCG	2085	0.0	23.81967	5
GGACCTT	8580	0.0	23.700302	44
ACCGCGT	2110	0.0	23.532833	6
CGCGTTC	2130	0.0	23.516998	8
CCGCGTT	2120	0.0	23.422468	7
CAGACCG	2275	0.0	22.404818	3
TACGGAG	2230	0.0	22.28054	34
ATACGGA	2250	0.0	22.083471	33
AATACGG	2270	0.0	21.984428	32

Figure 6. *List of k-mers*

The K-mer module is similar to the overrepresentation module. This checkpoint seeks for portions of the sequences, 7-mers that show positional bias according to the binomial test. In that way it could identify parts of sequences that can cause problems in further analysis. It is represented dually. For graphical representation it is characteristic that values on the X-axis are unequally represented. In tabular representation count, p-value, frequency and position in the read are shown, enabling to reconstruct the longer kmers (Fig. 6).

Because this type of analysis is slow, only 2% of the data is processed. The sequenced data is considered good if the presence of k-mers is balanced. If a binomial p-value is less than 0.01 for any k-mer, warning will be given (Fig. 7), and if p-value is less than 0.00001 a failure will occur.

This is another module affected by the Illumina bias at the 5'-end of reads and adapters and polyA tails at the 3'-end of the RNAseq reads, but these errors can be ignored as long as adapters are trimmed.

QC preprocessing

Because regions of low quality carry less information of interest, the main step in preprocessing

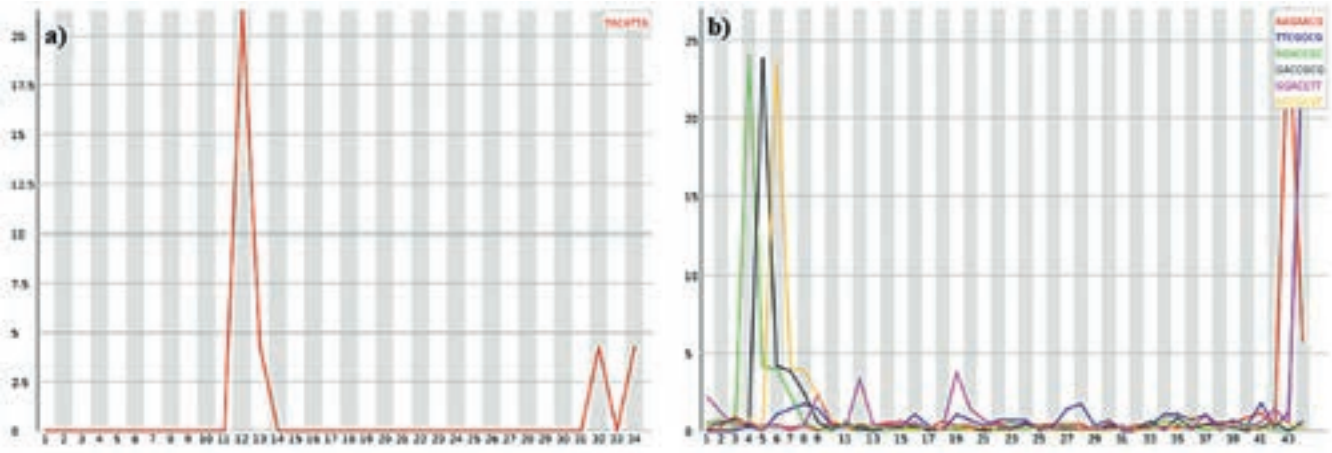

Figure 7. *K-mer content shows position in the reads on X-axis and relative enrichment of a k-mer on Y-axis; a) DNA-seq Illumina data-warning; b) total RNAseq Illumina data-failed*

is to deal with these kinds of sequences. Other steps in the preprocessing of NGS data depend on results obtained by the QC evaluation software. As stated in previous section, Illumina's characteristic 5' nucleotide bias, PCR and optical duplicates, polyA/T tails, rRNA and different kinds of contaminants, like adapters and genomic sequences that belong to different organisms than the sequenced one, are recognized by the QC evaluation tools as errors in data that should be removed in the QC preprocessing step. Many QC preprocessing tools have similar functionalities but they differ in speed, available resources (like adapter files) and effectiveness.

Processing of low quality data

The presence of low quality data is mainly the result of systematic errors and in the case of Illumina platform, common error is substitution. There are three general approaches in the preprocessing of low quality NGS data: error correction, masking and trimming. For error correction, there are three groups of methods applicable to Illumina data: k-spectrum-based methods (Rcorrector [9], Quake, Reptile, Hammer, Musket, Bless, Bloocoo, Lighter, Trowel), multiple sequence alignment (SEECER Karect, Coral, ECHO) and suffix array/tree-based methods (Fiona, SHREC, HSHREC, HiTEC, RACER) with only SEECER and Rcorrector being specially designed for RNAseq data. Brief description and evaluation of all methods can be found in [10], while the evaluation of different k-spectrum-based tools, as the dominant method for Illumina data, is available in [11]. Because k-spectrum-based error correction method uses only local information from sequenced reads, it is not suitable for analyses that include the read mapping task since mapping is done globally. On the other hand, this method is very useful for analyses with de novo assembling in plan because it reduces complexity of created graphs and speeds up the whole process of assembling [12]. Moreover, this error correction method is an integral part of some assembling tools like SOAPdenovo, ALLPATHS_LG, SGA and SPAdes.

When downstream analyses include read mapping, common approaches for dealing with low quality bases are masking and trimming. Masking is a non-invasive process of dealing with low quality data which implies substitution of low quality bases with N's. This is mostly performed with in house scripts, but there is also a tool, FASTQ Masker, included in FASTX-toolkit.

Although masking is a lenient method and it is shown that masking is more effective than trimming in SNP detection [13], trimming is the commonly used technique for this task. Trimming implies removing low-quality parts of sequences and because this is the main task of QC preprocessing, most of the general QC preprocessing tools have this functionality. Window-based (ConDeTri, PRINSEQ, Sickle, SolexaQA, Trimmomatic) and running sum (Cutadapt, SolexaQA with -bwa option, Seqtk) are two types of algorithm families used in trimming software, where the first approach gives better results for common RNAseq tasks like SNP calling, gene expression analysis and de novo assembly [14]. Trimming has different levels of stringency ranging from

mild, with removing bases with quality less than 5 (recommended by [15]), to more aggressive approach with quality threshold between 20 and 30 (recommended by [14]). Because the main goal of RNAseq is to get as much as possible information from the data, we recommend a milder approach for trimming.

Clipping 5' end

To overcome Illumina 5' end sequence content bias caused by random hexamers (presented on Per base sequence content plot), two approaches are suggested. One of them is a correction method specially designed for solving this kind of problem, described in [16] and the other one is clipping. Although clipping is always a loss of information (because it includes removing the correct bases too), this is a widely adopted QC technique for eliminating 5'end Illumina bias. Alternative to overcome this problem is not to eliminate the bias. Rationale behind this is that we can try to conduct downstream analyses without clipping the leading bases, and if we get poor results, we can always go one step back and use one of the aforementioned methods to overcome it.

Adapter removal

In total RNAseq it is common that sequenced fragments are shorter than the read length, which may result in adapter sequences left over. The presence of adapters is checked in QC evaluation, and if adapters are highly abundant, they will be listed in overrepresented sequences. Also, the presence of adapter sequences affects GC content and K-mer content plot. According to the aforementioned it is clear that all adapter sequences should be removed. This can be done with any general QC preprocessing tool.

Removing contaminants

Contamination of RNA sequences may arise from different reasons and it can affect several QC evaluation checkpoints, like Per sequence GC content, Sequence duplication level and Overrepresented sequences. If the goal is just to identify contaminants the appropriate tool is FastQ Screen or VecScreen, while tools like DeconSeq and BBSplit are used for actual decontamination. Another method for removing contaminants includes read mapping to possible contaminants genomes, but the same method can be used in order to check the percentage of contamination. For any kind of contamination detection or removal it is necessary to provide sequences/genomes of possible contaminants. Common contaminants in Illumina total RNAseq is PhiX control viral DNA, Illumina TruSeq primers, vectors like plasmid, phage, cosmid, BAC, PAC, YAC and transposable elements from the cloning host which is usually Escherichia coli or yeast. Other possible contaminants are the result of impurities in the RNA sample and include microbes and other organisms that are being sequenced in the lab. The list of possible contaminants is not finite, and we cannot know in advance which contaminants are expected to be present in the dataset.

Complete list of contaminants can be obtained using BLAST, but this is a time consuming task because of the huge number of sequences in the RNAseq dataset. Similar, but less time consuming approach would be to randomly select some reasonable number of sequences (eg. 500) from a transcriptome dataset and to blast them. Another approach would be to execute initial mapping and to BLAST unmapped sequences. Anyway, if contaminants do not comprise a significant number of sequences they will not affect the read mapping step, but for assembling they should be removed.

Removing PCR and optical duplicates

As stated in the previous section, although duplicates affect sequence duplication levels plot and GC content plot, removing them in RNAseq data can be more harmful than useful because read counts play a significant role in the downstream analysis of RNAseq data. If there is a good reason for removing the duplicates, one should be aware that, because it is much easier to locate duplicates after then before read mapping, available tools require a BAM file as input. In order to identify duplicates, Picard EstimateLibraryComplexity can be used. Widely used tools are Samtools rmdup (for removing PCR duplicates) and Picard MarkDuplicates (for marking or removing optical duplicates).

Removing polyA/T tails and rRNA

During the process of cleavage and polyadenyl-

ation, mRNA is enriched with polyA/T tails near the 3'end. The presence of polyA/T tails in the RNAseq can be seen in several modules of FastQC report: Per base sequence content, Per sequence GC content and Kmer content. PolyA/T tails can be removed by tools like PRINSEQ (options –trim_tail_right and –trim_tail_left) or with in house scripts. But, one should have a good reason for removing polyA/T tails because although they affect QC evaluation, they have no influence on read mapping (most mappers are not affected by presence of polyA/T tails) and, actually, they are beneficial for de novo assembling (because the polyA/T tail marks the end of transcript).

As part of the transcriptome, rRNA is expected to be present in the dataset to some extent. High abundance of rRNA can affect some checkpoints in QC evaluation, like GC content plot and overrepresented sequences, and they are often flagged as contaminants. Usually, rRNA is not of interest for downstream analysis and, as all other contaminants, it can be removed from the dataset with specific tools (SortMeRNA, BBDuk) or with mapping to the set of rRNA sequences available in sources like SILVA and Rfam database. But, the removal is not necessary because the high level of rRNA presence in RNAseq data does not represent actual contamination of the dataset. It just means that the dataset contains less mRNA than it is expected and because the presence of rRNA doesn't affect the next step of analysis (read mapping or de novo assembling), it is not necessary to remove rRNA from the dataset. Nevertheless, the exact level of rRNA presence in the dataset should be determined in order to resolve the amount of RNA of interest, using tools for identifying the level of contamination like FastQ Screen.

Sometimes rRNA has influence on specific RNAseq analysis and, in that case it can be masked. For example, in order to increase FPKM values (which have influence on differential gene expression) it is desirable to exclude rRNA, tRNA, mitochondrial and chloroplastic RNA and similar elements from analysis. This can be done with Cufflinks tool using –M option.

Use case – evaluation of QC preprocessing methods

According to the aforementioned, we formed two groups of proposed QC preprocessing steps: lenient and stringent, both intended to be used afterwards for read mapping and de novo assembling. Lenient group for mapping included: (1) trimming a read when average quality over a 4bp sliding window drops below 5 – Q5, (2) combination of Q5 with removal of adapters – AQ5, (3) combination of contamination removal with AQ5 – CAQ5, while stringent one comprised: (1) trimming a read when average quality over a 4bp sliding window drops below 25 – Q25, (2) combining adapter removal with Q25 – AQ25, and (3) removal of contamination combined with AQ25 – CAQ25. If the ending dataset contained reads shorter than 25bp, such reads were discarded. For the assembly, we used the dataset with removed adapters and contamination (CA) as a starting point. Lenient group was composed of the CAQ25 dataset and error corrected CAQ5 dataset (EC). Stringent group was formed by removing the first 11 bases from EC and CAQ25 datasets forming ECC and CAQ25C datasets, respectively. With lenient QC preprocessing steps, the resulting dataset will retain more information, and with stringent one the preprocessed dataset will consist of higher quality sequences. Both groups of approaches were compared to the naïve approach which refers to raw data analyses (R).

The dataset

To test the impact of two groups of QC preprocessing steps, we used a publicly available dataset of human breast cancer transcriptome with NCBI SRA accession number SRR2753165. This dataset contains raw Illumina Hiseq 2500 35-50bp single end transcriptome data, with cDNA Library being constructed using TruSeq stranded total RNA with Ribo-Zero Gold (for rRNA removal). Such useful information can guide us in determination of used adapters and primers as well as in the understanding of some QC evaluation results. In this example, the shape of the curve in the sequence length distribution plot is expected to be as in Fig. 3b because sequences are of variable length. Also, because the dataset is composed of transcriptome sequences, the peaks in the right part of the sequence duplication level graph are expected (as in Fig .4b).

Other FastQC results of interest are presented in Figures 1b, 2b, 5, and 7b. Deviations in 3'end in distribution of bases in sequences (Fig. 1b), in k-mer

content plot (Fig. 7b) and shift to the right in the GC content plot (Fig. 2b) imply the high abundance of polyA/T tails. We used NCBI BLAST [17] to determine origins of overrepresented sequences (Fig. 5). In this case, they were all small nuclear RNA (snRNA): 16 of them were 7SL RNA (type of SRP RNA and part of Alu transposable element), 3 were 7SK RNA, 3 were uncharacterized SRP RNA and 1 was uncharacterized snRNA. Both 7SL RNA and 7SK RNA are highly abundant GC-rich sequences. SRP RNA defines perinucleoar compartment [18] which is known to be present in breast cancer genomic data [19]. This explains the peculiar shape of GC content plot (Fig. 2b) to some extent. Other reasons include higher number of duplication sequences and abundance of rRNA sequences in the dataset.

To get deeper insight in the dataset content, we checked the level of contamination using Fastq Screen (Fig. 8). In order to determine the list of possible contaminants from other species, we sampled 500 reads from unmapped reads and blasted them against the human reference genome. Sampled sequences that remained uncharacterized were blasted against BLAST nt database. We detected contamination with Enterobacteria phage phiX174, so we used a standard list of contaminants. With less than 1% of PhiX and vectors from the UniVec [20] database, we may state that the dataset contamination is low. Further steps depend on the choice of the following step in downstream analyses. For the mapping step there is no need to clean the dataset because the contaminants will not map to the reference sequence. On the other hand, the assembly step requires as much clean data as possible and removal of any amount of contaminants is required. Also, it should be noticed that nearly 1% of the dataset consists of rRNA, which is not severe, but it surely influences the shape of GC content graph.

Presence of polyA/T tails can easily be checked by using simple grep command. In this dataset, around 7% of sequences have some form of polyadenilation. When the reads are short, the tails can be sequenced through and that causes the occurrence of other nucleotides after polyA/T tail. But, this is not a problem, because many read mappers and de novo assemblers perform well with polyadenilation, so it is not necessary to remove it from the dataset in any case.

As it is shown in Fig. 5 and Fig. 8, there is no significant abundance of adapters. Nevertheless, all adapters should be removed from the dataset, independently of the choice of the following steps.

For quality trimming, to remove adapters and to discard reads shorter than 25bp (which can arise after quality trimming and adapter removal), we used Trimmomatic v0.36 because it is easy-to-use, fast, lightweight window-based QC preprocessing tool, specially designed for Illumina data. To remove PhiX and vector sequence contamination we used read mapping with TopHat v2.1.1 [21].

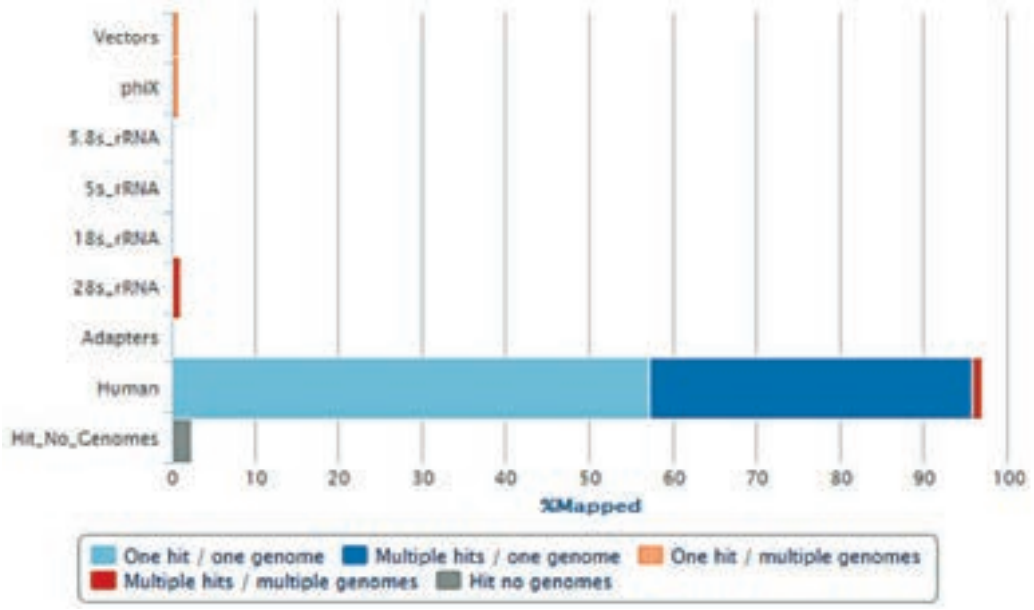

Figure 8. *Level of contamination using Fastq Screen*

Mapping

In order to compare different approaches and their benefit to the results of mapping step, we chose widely used read mapper Tophat2 and compared its performance when using default parameters, as is done in many studies, with its performance when using chosen set of parameters that contribute to the higher accuracy and confidence of results. As a reference genome we used GRCh38.p11 with appropriate GFF files. For comparison we used statistics from Tophat2 (Table 1a) taking into account not just read mapping rates but also the number of useful sequences. Additionally, we used Qualimap v.2.2.1 to count the number of reads mapping that belong to exonic (90.53%), intronic (7.61%) and intergenic (1.87%) regions as well as to determine duplication rates (29.09% for raw data and lenient approaches and 28.71% for stringent approaches) and uniformity of coverage through 5' and 3' bias (5' bias is 0.28, 3' bias is 0.56 and ratio between 5' and 3' bias is 0.84).

Similar studies, taking into account only the quality trimming, showed that mapping rate increases with the stringency of quality trimming [14], while the absolute number of aligned reads decreases [22]. We can state the similar - QC preprocessing approaches that include lenient quality trimming are beneficial in absolute number of aligned reads and in that way they contribute to overall amount of usable information in final dataset, but only one approach, CAQ5, provide the increase in mapping rates as well. The reason for this is the presence of 73583 contaminant sequences in AQ5 dataset, which increases the number of sequences in the dataset, but not the number of mapped sequences. On the other hand, the absolute number of multi aligned reads in a lenient QC preprocessing approach is also increased when compared to a stringent method, but the overall percentage is lower in lenient methods.

When we conducted a similar analysis with Tophat2, only this time we used setting of additional parameters as more appropriate for the nature of the set and the sequenced data themselves, we obtained the higher mapping rate for all approaches, but the conclusion was the same. Since we expected a variability in the dataset, we increased the number of mismatches (-N 4) and number of mutations expressed through edit distance metric (--read-edit-dist 5). Also, we demanded realignment of reads with edit distance higher than 2 (--read-realign-edit-dist 2) and no multihits (-g 1). Results are given in Table 1b. Again, we used Qualimap to determine quality of mapping (duplication rate: 27.35%-27.73%; mapping to specific regions: exonic - 91.16%, intronic - 6.96%, intergenic - 1.88%; 5' bias: 0.16, 3' bias: 0.61, 5'-3' bias: 0.3) and it also revealed the overall better quality of mapping when appropriate parameters in Tophat 2 were used.

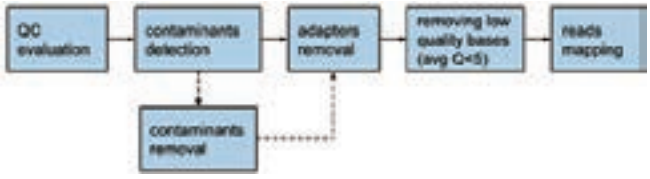

Figure 9. *Proposed workflow for reads mapping task*

Appropriate parameters contributed to the improvement of results in all approaches, with CAQ5 giving the best result. The proposed workflow is given in Fig. 9.

Table 1. *Comparison of Tophat2 mapping results. R- raw data, C- removed contamination, A – removed adapters, Q5 – trimmed reads using sliding window with quality threshold set to 5, Q25 – trimmed reads using sliding window with quality threshold set to 25. Reads shorter than 25bp were discarded.*

a) mapping performed with default parameters							
Data	**R**	**Q5**	**Q25**	**AQ5**	**AQ25**	**CAQ5**	**CAQ25**
No. of mapped seq	9049533	9051992	8913954	9051962	8913946	9051626	8913640
Mapping rate	98.5%	98.5%	98.7%	98.5%	98.7%	99.3%	99.5%
% of multi aligned	19.8%	19.8%	19.9%	19.8%	19.9%	19.8%	19.9%
b) mapping performed with parameters –g 1 –N 4 --read-edit-dist 5 --read-realign-edit-dist 2							
No. of mapped seq	9072621	9075046	8929607	9075035	8929538	9074742	8929260
Mapping rate	98.7%	98.8%	98.9%	98.8%	98.9%	99.6%	99.6%

De novo assembly

An accurate assembly requires a high quality clean data, which can be provided in one of two ways: by using stringent QC preprocessing approach on cleaned data (CAQ25) or by using lenient approach on clean data (CAQ5) combined with error correction (EC) performed using Rcorector (a tool specially designed for this task in Illumina RNAseq data), with maximum number of correction set to 5. Additional steps in QC preprocessing, like elimination of Illumina specific bias, can contribute to the quality of assembly. According to Fig. 1b we determined that the first 11 bases are affected with Illumina bias so we used Trimmomatic to remove them from EC and CAQ25 datasets. In that way we formed two additional datasets, ECC and CAQ25C, in order to test the impact of clipping of biased bases on de novo assembly. All five datasets (R, CAQ25, EC, CAQ25c and ECc) were assembled using Trinity v2.4.0 [23] with minimal contig length set to 30 (because initial reads are short, ranging from 25 to 50bp). We didn't change the values of other default parameters because in Trinity they are set to give the best results for standard RNAseq data. For the evaluation of the assembly we mutually compared values of several standard metrics: N50, average contig length (both measured using in house perl script), reads mapping rate to transcriptome, transcriptome mapping rate to reference transcriptome and the number of proteins that matched the assembled transcriptome more than 80% (Table 2). All read mappings to transcriptome were conducted by using bowtie2 [24] with parameter --local, while the number of proteins was obtained by using BLASTX [25] and by the utility perl script provided within Trinity installation.

A desirable resulting assembly has the greater assembly length and the smaller number of contigs. That is why we observed the ratio between these two values in different approaches, with an aim to obtain the larger number represented as average contig length. Because contigs do not have uniform length, for mapping of reads to assembly and for mapping of assembly to reference transcriptome we observed only mapping rates and not the number of mapped contigs. Datasets used in stringent approaches included the clipping of the 10 bases from the 5' end, which made the sequences shorter and led to significant loss of information. This was reflected in the results by larger number of contigs and smaller assembly length, which is why the assembly had low average contig length. Another affected metric is the number of proteins which matched transcriptome in high percentage, used to reflect the number of almost full transcripts, showing that, although the mapping rate to reference transcriptome was high, obtained transcripts were not complete. On the other hand, it was beneficial for the mapping task, which was expected because it is easier to fit in the shorter than longer sequence in the reference sequence, but the longer ones are more reliable. Aforementioned is even more noticeable in the used dataset because the initial dataset consisted of very short sequences (35-50 bp) and clipping made the great amount of extremely short sequences which were not suited for further analysis. Methods in the lenient group gave better results, with similar results in most of the metrics, but with EC approach giving the highest number of >80% complete transcripts. According to the aforementioned, we recommend using EC approach.

Table 2. *Comparison of Trinity results for differently preprocessed data with parameter --min_contig_length set to 30. R – raw data, C – removed contamination, A – removed adapters, Q5 – trimmed reads using sliding window with quality threshold set to 5, Q25 – trimmed data using sliding window with quality threshold set to 25, EC – CAQ5 with mostly 5 error corrected bases, c – clipped 10 bases from 5' end. Reads shorter than 25bp were discarded.*

Data	R	CAQ25	EC	CAQ25c	ECc
N50	69183	69101	68014	120637	120571
average contig length	100.19	99.56	100.81	63.83	64.35
% of reads mapped to transcriptome	92.39	92.23	92.38	90.24	90.50
% of transcriptome mapped to reference transcriptome	86.49	86.71	86.31	90.21	89.95
No. of proteins matched the transcriptome >80%	2784	2793	2939	922	979

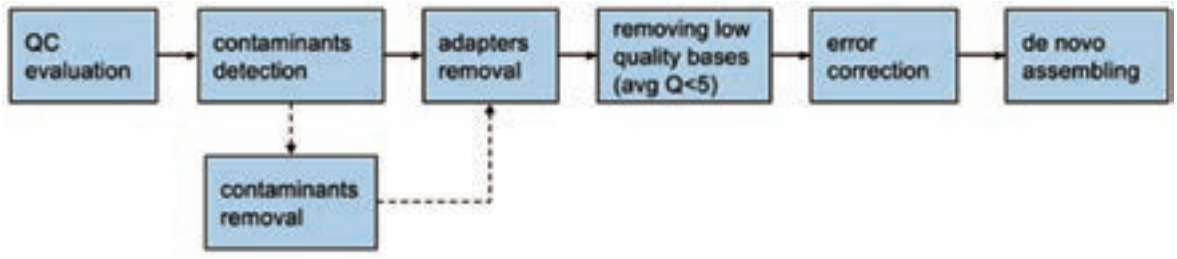

Figure 10. *Proposed workflow for de novo assembly task*

Conclusion

Specifics of RNA sequencing and RNAseq datasets have to be taken into account when choosing the best approach in the evaluation, preprocessing and analysis steps. Widely used software for quality control FastQC is not adjusted to RNAseq data which is why an adequate interpretation of obtained results is necessary, but the ultimate choice of appropriate QC preprocessing steps depends on the following step of downstream analysis - mapping or de novo assembling. Generally, all levels of QC preprocessing can be grouped into three approaches: naive approach which presumes that raw data is a high quality data, lenient approach which preserves more information, and stringent approach which provides cleanest dataset. Evaluation of approaches demands using significant and reliable metrics like mapping rate, number of mapped sequences and multi-mapped reads ratio for mapping task and N50, average contig length, percentage of assembly mapped to reference transcriptome, ratio of unique reads mapped to transcriptome and number of proteins matched the transcriptome in more than 80% for assembly task. For the comparison of QC results between different approaches and as a use case for studying the influence of QC on further analyses steps, we used publicly available RNAseq raw data from well researched and reference resources rich genome to provide reliable results. Our results showed that when downstream analysis included mapping, lenient methods with adapter removal included gave the best results. Moreover, results revealed that an influence of parameter selection on the mapping task exists, which is why parameters should be selected according to the nature and purpose of the NGS experiment. For the assembling task, the best results were obtained by using a lenient approach with error correction.

References

[1] Sheng, Q., Vickers, K., Zhao, S. *et al*. Multi-perspective quality control of Illumina RNA sequencing data analysis. Briefings in Functional Genomics 2017; 16(4): 194–204.

[2] Bolger, A. M., Lohse, M., Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 2014; 30(15): 2114–20.

[3] FastQC tool. Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [cited 24 August 2021]

[4] Okonechnikov, K., Conesa, A., García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data, Bioinformatics 2016; 32(2): 292–94.

[5] Hansen, K. D., Brenner, S. E., Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 2010; 38(12).

[6] Delhomme, N., Mähler, N., Schiffthaler, B., *et al.* Guidelines for RNA-Seq data analysis. Epigenesys; 17 November 2014.

[7] Meyers, B. C., Tingey, S. V., Morgante, M. Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. Genome Research 2001; 11(10):1660–76.

[8] Gu, Z., Wang, H., Nekrutenko, A. Li, W.H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. Gene 2000; 259(1-2), pp. 81-8.

[9] Song, L. Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience 2015; 4(1),

[10] Yang, X., Chockalingam, S. Aluru, S. A survey of error-correction methods for next-generation sequencing. Briefings in Bioinformatics 2013; 14(1), 56–66.

[11] Akogwu, I., Wang, N., Zhang, C., Gong, P. A comparative study of k-spectrum-based error correction methods for next-generation sequencing data analysis. Human Genomics 2016;10(20).

[12] Gabaldón, T. Alioto, T. Whole-Genome Sequencing Recommendations. In A.M. Aransay, J.L. Lavín Trueba Eds. Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing, Springer International Publishing 2016; pp.13-41.

[13] Yun, S., Yun, S. Masking as an effective quality control method for next-generation sequencing data analysis. BMC Bioinformatics 2014; 15(1).

[14] 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).

[15] MacManes, M. D. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics

2014; 5(13).
[16] Hansen, K. D., Brenner, S. E., Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Research 2010; 38 (12).
[17] Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden T.L. NCBI BLAST: a better web interface. Nucleic Acids Research 2008; 36:W5-9.
[18] Wang, C., Politz, J. C., Pederson, T., Huang, S. RNA Polymerase III Transcripts and the PTB Protein Are Essential for the Integrity of the Perinucleolar Compartment. Molecular Biology of the Cell 2003,14(6):2425–35.
[19] Kamath, R., Thor, A., Wang, C. *et al.* Perinucleolar Compartment Prevalence Has an Independent Prognostic Value for Breast Cancer. Cancer Research 2005; 85(1):246-53.
[20] UniVec database. Available at: ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/ [cited 24 August 2021]
[21] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R, Salzberg, S.L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 2013, 14(4
[22] Williams, C. R., Baccarella, A., Parrish, J. Z., Kim, C. C. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 2016;17.
[23] Grabherr, M. G., Haas, B. J., Yassour, M., et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 2011; 29(7):644–52.
[24] Langmead, B., Salzberg, S. L. Fast gapped-read alignment with Bowtie 2, Nature Methods 2012; 9(4):357–59.
[25] Gish, W. States, D.J. Identification of protein coding regions by database similarity search. Nature Genetics 1993, 3:266-72.

Received: September 10, 2021
Accepted: November 19, 2021

ABOUT THE AUTHORS

Dragana Dudić is an Assistant at the Faculty of Informatics and Computer Science at the University Union Nikola Tesla for several Computer Science subjects. She obtained BSc and MSc in Computer Science at the Faculty of Mathematics, University of Belgrade where she is finishing her PhD in the field of Bioinformatics. Dragana is interested in research of repetitive sequences, especially human and plant transcriptomics.

Bojana Banović Đeri is a Research Associate in the Laboratory for Plant Molecular Biology in the Institute of Molecular Genetics and Genetic Engineering, University of Belgrade. In 2013. obtained PhD in Molecular biology at the Faculty of Biology, University of Belgrade in the field of plants' molecular biology, genetics and proteomics (characterization of buckwheat heteromorphic self-incompatibility).

Vesna Pajić is a Bioinformatician at Seven Briges Genomics. She has a strong education professional with a PhD. focused in Computer Science. Vesna is experienced in both academia and the biotechnology industry. She is interested in applications of mathematics and CS in life sciences and skilled in computer applications, bioinformatics, and management.

Gordana Pavlović-Lazetić is a Professor of Computer Science at Faculty of Mathematics, University of Belgrade. Gordana has forty years of teaching experience at the University of Belgrade in different programming languages and database courses. Her recent research interests are in bioinformatics and natural language processing. She spent two years as a researcher at the University of California, Berkley, and the Relational Technology Inc, as a consultant for extending the Relational Database System Ingres to manage text.

FOR CITATION

Dragana Dudić, Bojana Banović Đeri, Vesna Pajić and Gordana Pavlović-Lažetić, Fotis Lazarinis, Demystification of RNAseq Quality Control, *JITA – Journal of Information Technology and Applications Banja Luka*, PanEuropien University APEIRON, Banja Luka, Republika Srpska, Bosna i Hercegovina, JITA 11(2021) 2:73-86, (004.62.032.26:577.216.3), (DOI: 10.7251/JIT2102073D), Volume 11, Number 2, Banja Luka, December 2021 (69-140), ISSN 2232-9625 (print), ISSN 2233-0194 (online), UDC 004