Specific Quality Control is essential for Next-Generation Sequencing data usage: case studies in Illumina data from algae, yeasts and plants

Brenda Neves Porto[1], *Andrei Stecca Steindorff*[2], *Lucas Soares de Brito*[3], *Eduardo Fernandes Formighieri*[4]

Abstract

Next-Generation Sequencing technologies have contributed extensively to genomic studies due to the high-throughput data generation. However, errors can occur during the sequencing process and need to be removed before other analyses, otherwise, lead to understated results or wrong conclusions. In this paper we evaluate common errors of Illumina sequencing technology, such as: reads quality, adapter and vector residues, insert size of Long-Jumping Distance libraries. We evaluated data from algae, yeasts and plants, and propose guidelines for Quality Control.

Introduction

Next-generation sequencing (NGS) technologies have expanded the breadth of genomics. Genome data, once restricted to model organisms, can now be generated for any species at remarkable speed and low cost (TRIVEDI et al., 2014). An annotated genome draft provides an *in silico* preview of species metabolism, and it´s a base for other "omics" approaches, such as Transcriptomics, Metabolomics, Exomics etc. However, although cheaper, DNA sequencing technology still generates small fragments that needs to be reassembled.

Usually, millions of reads are generated for a genome assembly, and the number of errors accompanying the increase in the speed of data generation. To ensure quality of the data set, errors such as: i) wrong identification of nucleotide;

[1] Bióloga, doutora em Biotecnologia Vegetal, Universidade de Federal de Lavras, brenda.neves@colaborador.embrapa.br
[2] Biólogo, doutor em Biologia Molecular, Universidade de Brasília, andreistecca@gmail.com
[3] Engenheiro de Redes de Comunicações, mestre em Engenharia de Sistemas Eletrônicos e de Automação, Universidade de Brasília, lucas.brito@colaborador.embrapa.br
[4] Engenheiro-agrônomo, doutor em Biologia Funcional e Molecular, pesquisador da Embrapa Agroenergia, eduardo.formighieri@embrapa.br

ii) contamination and iii) residues of adapters, must also be identified and corrected and/or trimmed.

Data set Quality control (QC) is essential for raw NGS data not to lead to underused data and/or erroneous conclusions. Among the different sequencing platforms, the Illumina (www.illumina.com/systems.html) is the most used (YANG et al., 2013). During the QC analysis for assemble, low-quality reads are removed, as well as sequences from primers, vectors, adapters and contamination. Also, some repetitive content can be marked or separated to not disturb the assembly process (PATEL; JAIN, 2012; ZHOU et al., 2013).

In this study, we evaluated different sets of data and propose Quality Control guidelines to be applied in data sequencing from Illumina, aiming to improve the genome assembly. All work was developed by Bioinformatics Research Group at the Bioinformatics and Bioenergy Laboratory - LBB (lbb.cnpae.embrapa.br). All used software products are free to use and run in Linux OS.

Material and methods

Material

The data set is composed by DNA sequences from two strains of algae (LBA32 and LBA40), three yeast isolates (A1, A5 and A9) and two species of plants (*Attalea speciosa* and *Acrocomia aculeata*). The *Mi-Seq* technology was used to generate *paired-end* (PEs) data from algae and yeasts short insert libraries (SIL). *Hi-Seq* technology was used to generate PEs from plants SIL, and from yeasts and plants Long Jumping Distance Libraries (LJDL).

Reads quality and trimming

The quality visualization of all data sets, before and after filtering and trimming, was performed using the FastQC software, with default parameters (v. 0.11.5, www.bioinformatics.babraham.ac.uk/projects/fastqc). FastQC generates a html formatted result, including general statistics and several graphs for different aspects of reads composition, such as Ns, kmers, adapters, quality and bases proportion. All graphs present an indicative status icon, and must to be evaluated. Usually, the problems are: quality, residues from adapters and contamination.

We evaluated two of the main software's for reads QC: FASTX-Toolkit (v. 0.0.13, http://hannonlab.cshl.edu/fastx_toolkit/) and the NGS QC Toolkit software (v. 2.3.3, www.nipgr.res.in/ngsqctoolkit.html). NGS QC was chosen for next steps because it presents more complete analysis and a graphic interface. We tested different parameters to improve the quality of reads, such as percentage of reads with minimum quality, phred quality threshold, end trimming, and cleavage of adapters and vectors.

Analysis of insert size for LJDL data

To evaluate the insert size of LJD libraries from yeast, we used BWA software (v. 0.7.15, https://sourceforge.net/projects/bio-bwa/files/) to map these read pairs against its own genome, then sort results using software SAMtools (v. 1.3.1, https://sourceforge.net/projects/samtools/files), and finally held metric insert size with Picard-Tools software (v. 1.119, https://broadinstitute.github.io/picard/). For *Acrocomia aculeate* data set, we follow the same method, except for genome used as reference for mapping, since we used the know genome of the closest specie, *Elaeis guineensis* EG5 (www.ncbi.nlm.nih.gov/genome/2669?genome_assembly_id=34040).

Results and discussion

Reads QC

Data sets generated by Illumina *Mi-Seq* sequencer (PEs libraries from algae and yeasts) present most of reads with length of 251 bp. They are similar in quality, and after some tests, we choose these parameters for quality filtering, aiming a basic trustable data set for assemblies: 70% reads / phred >= 20 (70/20). At Figure 1, we present some FastQC graphs from LBA32 algae data, being part 'A' a quality plot from Raw data, and 'C' a plot from filtered and trimmed data. The yellow boxes represent data frequency from Percentil 25 to 75, with a red line at median.

The parts 'B' and 'D' of the Figure 1 show nucleotide distribution of all reads along reads position. Considering the pairing of the nucleotides (AT/CG) and that we are showing millions of reads, by chance, we should see just four parallel lines, like the central region of the graphs. The first (5') positions present a typical bias of Illumina library build, but a bias at the final (3´) positions can indicate the

presence of non-trimmed adapters, or, residues from sequencing that must be cut. After filtering and trimming, we recovered 80% of data from algae (LBA32 e LBA40) and 90% of data from yeast (A1, A5 e A9).

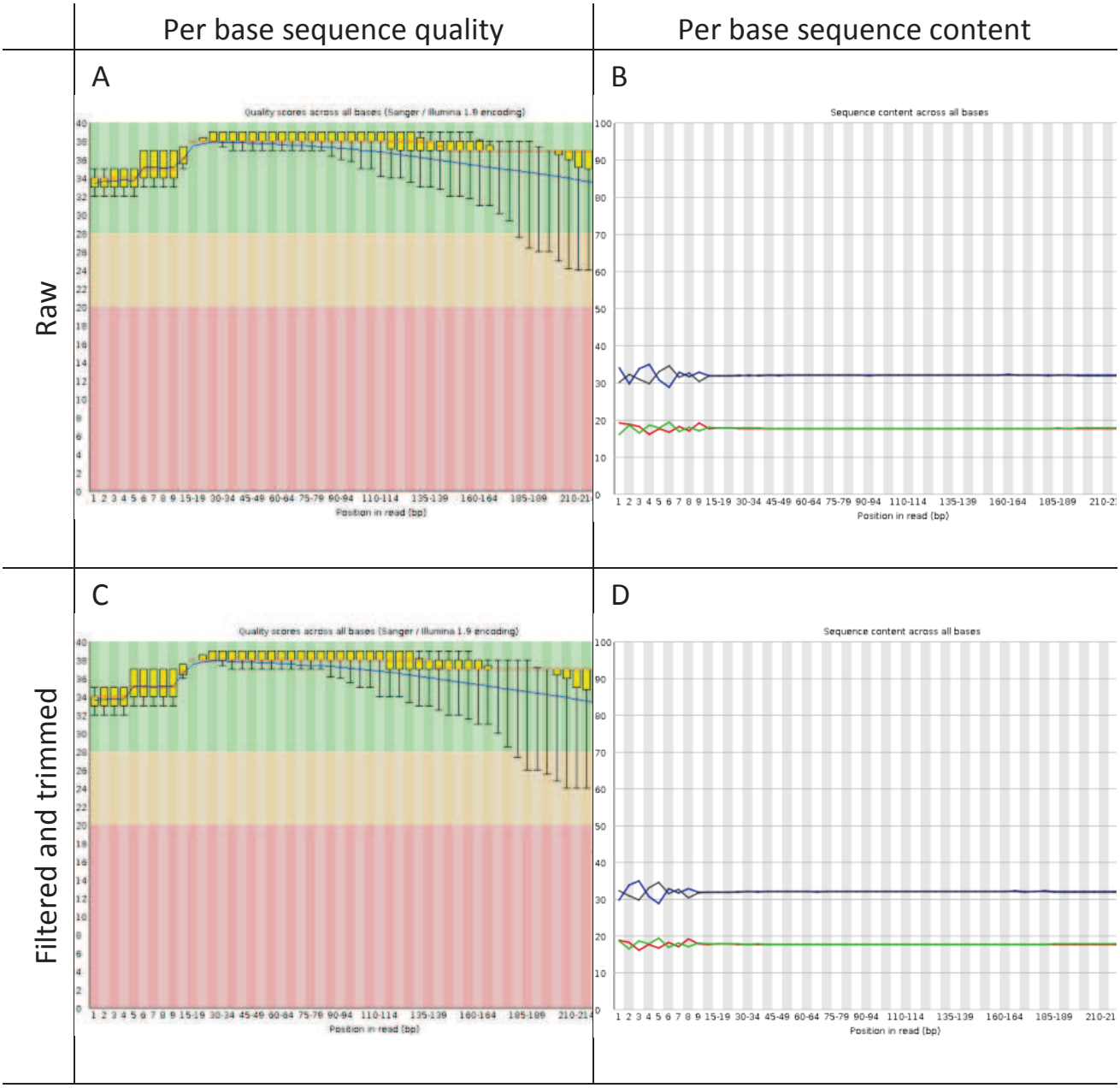


Figure 1. LBA32 algae data set visualization through FastQC. Raw (A and B) versus Filtered/trimmed (C and D) data. Filter parameters: 70% of reads with at least quality phred 20. Trimming parameters: trim 3 bases at 3' end.

Data sets generated by Illumina *Hi-Seq* sequencer present most of reads with length of 125 bp (*High Output Mode*, PEs and LJD libraries from *A. aculeata*) or 150 bp (*Rapid Run Mode*, PEs from *A. speciosa*). The PEs libraries sequencing of

the *A. speciosa* and *A. aculeata* generates discrepant quality: very poor in the first case (Figure 2A) and very high in the second (Figure 2C).

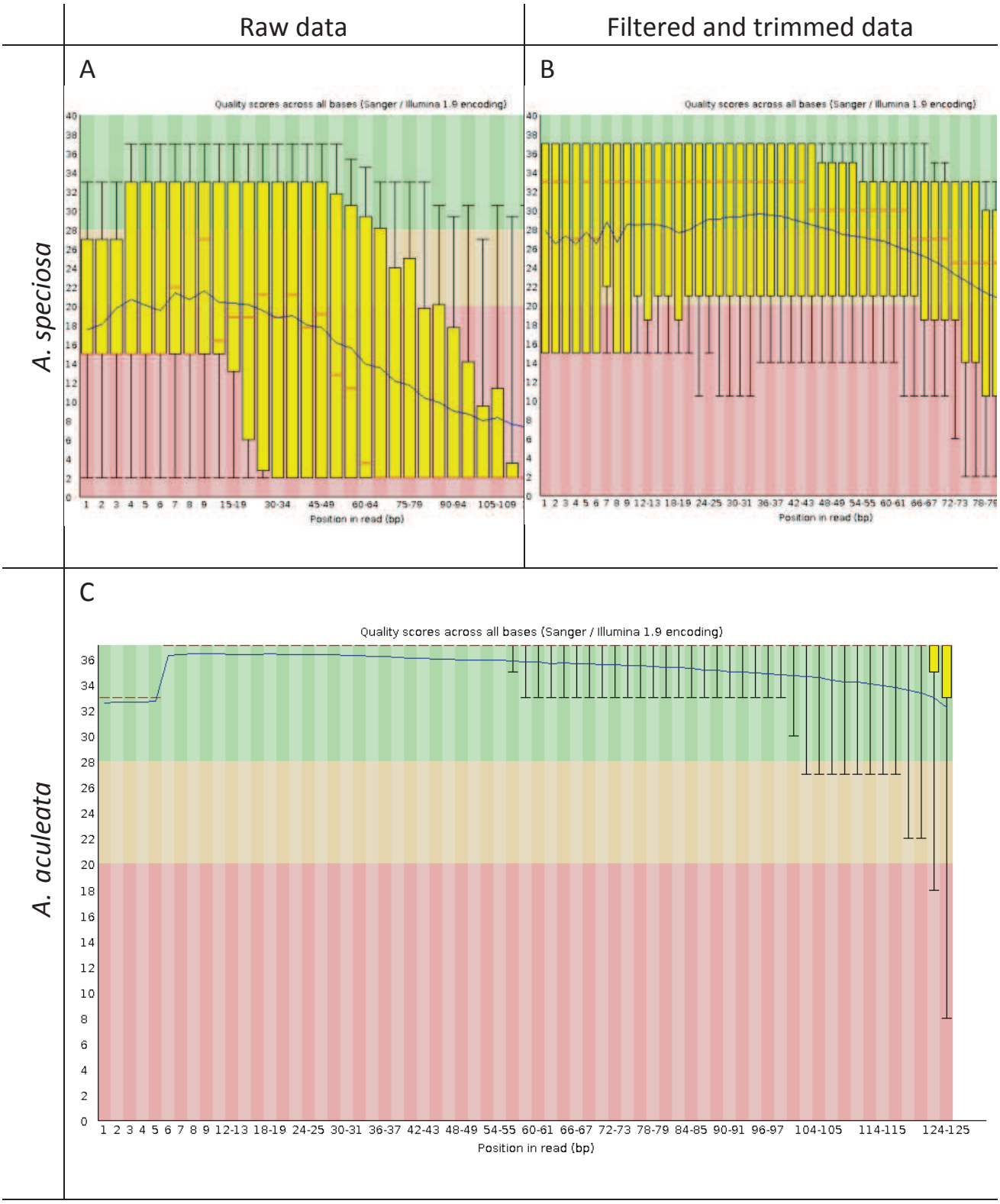


Figure 2. Plants Hi-Seq results: *A. speciosa* ('A' and 'B') and *A. aculeata* ('C') data sets visualization through FastQC. Raw ('A' and 'C') and Filtered/trimmed ('B'). Filter parameters: 50% of reads with at least quality phred 20. Trimming parameters: 10 nt 5'-end, 40 nt 3'-end.

DNA samples degradation during transportation and the mode of run are the most likely causes of the poor quality. We tested some QC approaches, and choose by cost/benefit: cut 10 bp at 5'-end and 40 bp 3'-end, filtering reads with at least 50/20, resulting only 40% of original data (Figure 2B), even with low quality (with 70/20, remain only 10%). Moreover, it was not necessary to trim or filter the *A. aculeata* data (Figure 2C).

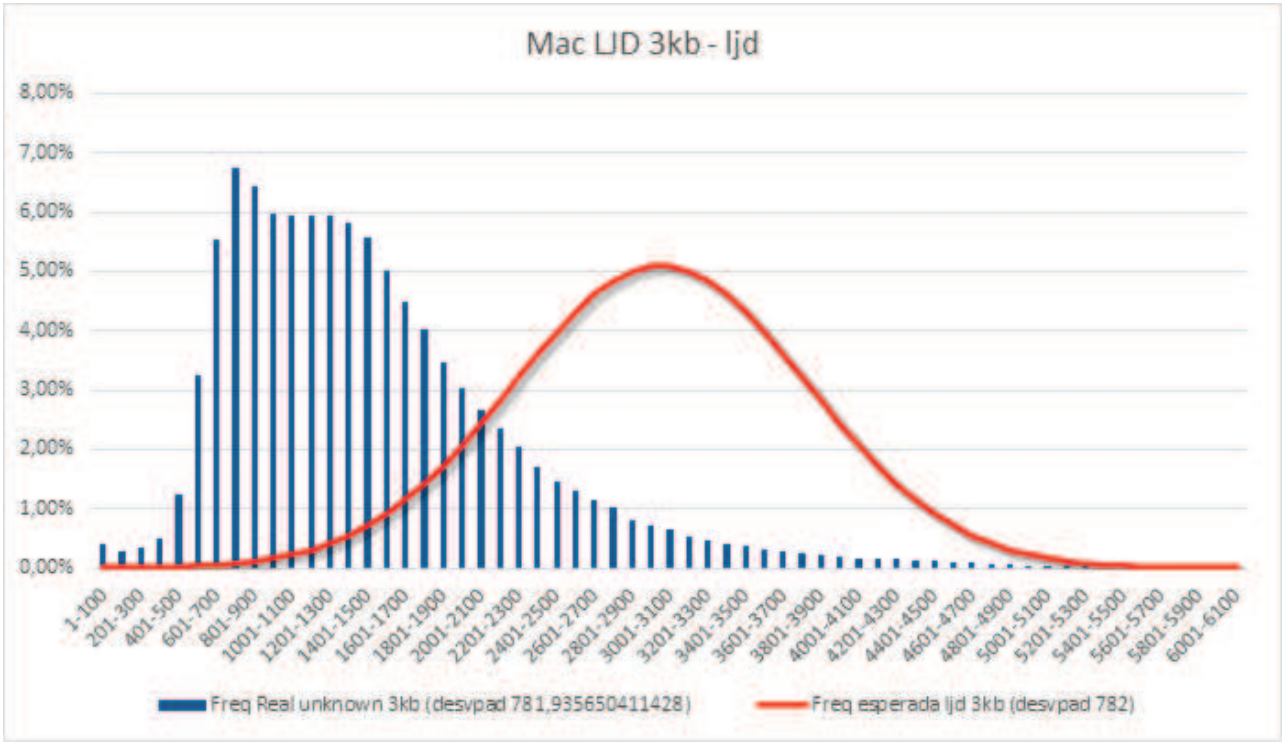
LJD libraries insert sizes

LJD libraries usually generate less reads than PEs, but although most of the quality assessment is similar, an additional step is required to evaluate insert size (distance between pairs at the original DNA sequence). To evaluate the data set from *A. aculeate*, presented at Figure 3, we used a draft genome from the closest specie with genomic sequences public available (*Elaeis guineensis* EG5) as reference, based in the premise that most part of mapped pairs will be in relatively conserved regions. Around 30% of the reads mapped, what means millions of matches, and a representative sample to support our discussion. Yeasts analysis presented similar patterns.

Figure 3 patterns shows that most part of the data generated from LJD libraries present sizes smaller than expected. Based in this analysis, it is needed to regroup read pairs according to the insert size, considering the limit standard deviation of used assembler, to allow better assembly.

LJD file data set

A



PE file data set

B

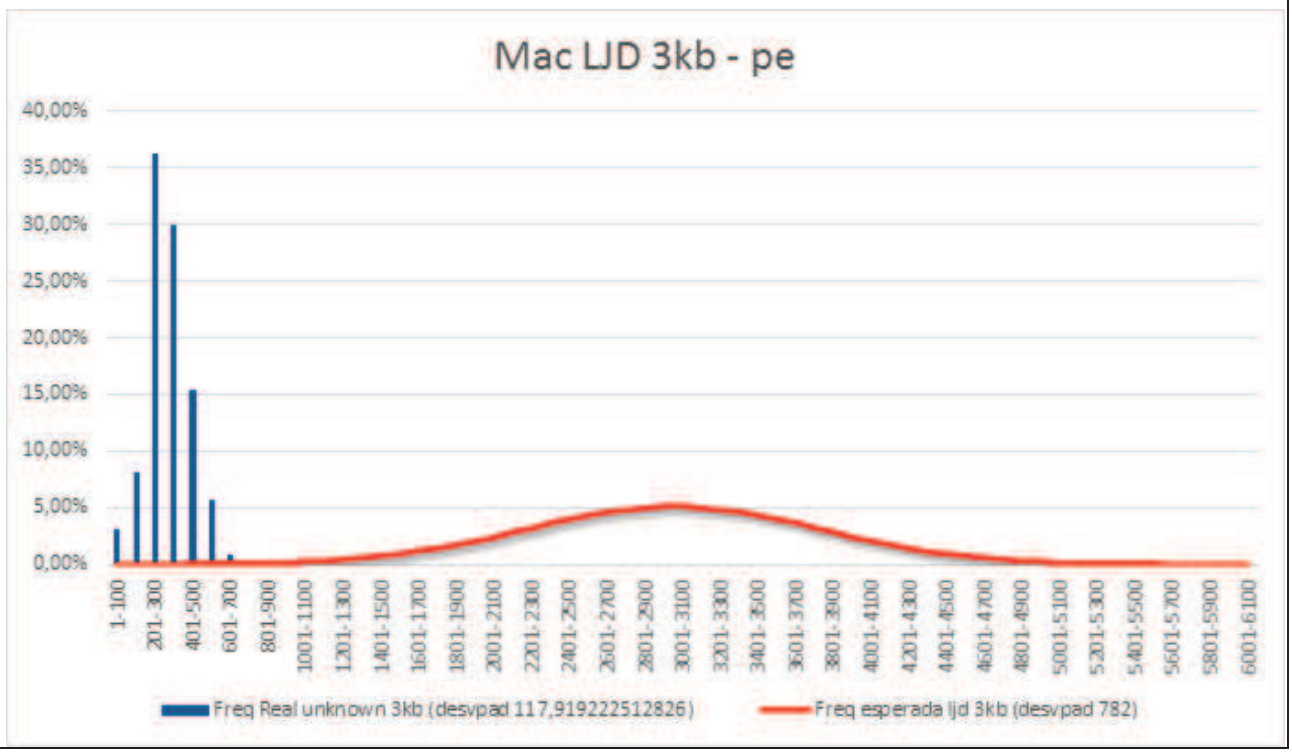


Figure 3. *Acrocomia aculeata* - insert size evaluation from LJD sequencing. The normal distribution curves (as reference, in red) was built with the parameters: *mean* = 3,000; *standard deviation* = 782 (same value found in real mapped data from LJD file).

Quality Control guidelines for Mi-Seq and Hi-Seq data

As a final result, summarizing and systematizing the analysis, we present guidelines to be considered in Quality Control of Illumina sequencing data:

a. Run FastQC into each raw data file and take a look at the graphs, mainly:
   - Per base sequence quality (Figure 1 and 2) – check data quality and define quality filter initial parameters to test.
   - Per base sequence content (Figure 1) – check lines, compare to quality, define ends trimming.

b. Run NGS QC Toolkit for each pair of files (R1 and R2), varying filter parameters according to previous evaluation, assembler to be used, genome size and complexity, and amount of data or genome coverage (e.g. 70/20, 70/30, 80/20), and run again FastQC for each filtered file. The NGS QC can: i) remove low quality reads; ii) cut adapters and vectors; and iii) cut under demand 5'-end and 3'-end of all reads.

c. Alternatively, and/or complementary, you can run Trimmomatic (www.usadellab.org/cms/?page=trimmomatic) using similar parameters.

d. After running these tests and define the better parameters, remember to delete all other results (keep just the FastQC results). Compact original files and work from now with filtered ones. If there is low disk space, develop a script to run filtering, FastQC, and then delete filtered files before run next test.

e. For LJD data, also do the insert size evaluation (download the closest genome, map with BWA, sort (SAMtools), measure (Picard-Tools), regroup mapped data according to insert size).

f. Additional tips to help assemblies: i) find and separate reads from organelles; ii) verify other contamination (e.g. Blast); iii) check organism identification (phylogeny of ribosomal and mitochondrial genes); iv) filter repetitive content; v) talk to a bioinformatician before hiring sequencing – seek collaboration during project planning.

g. Check other works of this event that complement the information of this work: Lucas S. de Brito – Repeats; Andrei S. Steindorff – Yeasts assembly.

Conclusions

Quality Control of sequencing data is not a simple task. It demands robust infrastructure of hardware and software, specific software and several tests with different parameters. However, some guidelines can make the job easier.

Insert size of LJD data is an undervalued issue. Use the correct values of insert size during is definitely an important aspect for genome assembly.

Financial support

This study was supported by the Brazilian Ministry of Science, Innovation and Technology (MCTI) through a grant provided by Conselho Nacional de Desenvolvimento Científico (CNPq).

References

PATEL, R. K.; JAIN, M.. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. **PloS One**, San Francisco, v. 7, n. 2, p. e30619, 2012.

TRIVEDI, U. H.; CÉZARD, T.; BRIDGETT, S.; MONTAZAM, A.; NICHOLS, J.; BLAXTER, M.; GHARBI, K. Quality control of next-generation sequencing data without a reference. **Frontiers in Genetics**, Lausanne, v. 5, p. 111, 2014.

YANG, X.; LIU, D.; LIU, F.; WU, J.; ZOU, J.; XIAO, X.; ZHAO, F.; ZHU, B. HTQC: a fast quality control toolkit for Illumina sequencing data. **BMC Bioinformatics**, London, v. 14, n. 33, p. 1, 2013.

ZHOU, Q.; SU, X.; WANG, A.; XU, J.; NING, K. QC-Chain: fast and holistic quality control method for next-generation sequencing data. **PLoS One**, San Francisco, v. 8, n. 4, p. e60234, 2013.