LEXOGEN
The RNA Experts

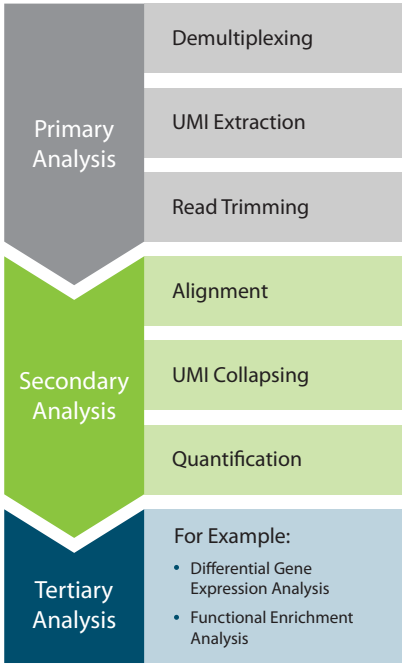
RNA LEXICON

CHAPTER #11

Data Analysis and Quality Control – Primary Analysis

The increase in throughput and popularity of RNA-seq has resulted in an unprecedented need for expertise in bioinformatics and computational resources. Although many steps in RNA-seq analysis have become standardized, others still pose major bottlenecks or pitfalls. As RNA-seq, the de facto standard of transcriptome profiling, continues to push the boundaries of data output, the demands and expectations of processing this data has increased significantly. This is especially true when used for diagnostic or screening purposes where data analysis workflows must be quality controlled to ensure highly reproducible findings. This series of Lexicon chapters outlines the central data processing and quality control steps essential to RNA-seq analysis.

How is RNA-seq data analyzed?



RNA-seq analysis is commonly divided into primary, secondary, and tertiary analysis (Fig. 1). Primary data analysis includes processing the raw sequencing data. This step consists of demultiplexing the samples according to their respective indices (barcodes) and read trimming. Secondary analysis describes the process of aligning and quantifying the pre-processed reads. Tertiary analysis focuses on extracting biologically relevant information from the samples. Often, this step includes differential gene expression (DGE) analysis and gene ontology (GO) term or gene set enrichment analysis, which are described in our upcoming Chapter 13. As tertiary analysis is the final and most extensive analysis step, it also encompasses visualizing the results. Distilling large amounts of data into comprehensive and easily understood figures is an art of its own. However, data visualization and figure creation are beyond the scope of this article and will not be discussed. Data visualization remains one of the most challenging but exciting aspects of RNA-seq analysis and is a process each researcher should further explore on their own.

Figure 1 | Data analysis steps overview. Primary analysis refers to the initial analysis steps in which the reads are prepared for further processing steps. Secondary analysis which we will focus on in Chapter 12 encompasses alignments, UMI analysis and gene / transcript quantification. Tertiary analysis uses the output generated from the primary and secondary steps to analyze the generated data in a physiological context, examples include differential expression analysis comparing multiple conditions or identification of signaling pathways, regulated targets, or interaction partners.

1. Primary Analysis

In this chapter of RNA Lexicon we will focus on pre-processing of the reads during the primary data analysis steps.

Before the Analysis: Sequencing Run Quality Control

Before starting data analysis, sequencing run performance should be evaluated by assessing a set of parameters that are specified for the individual instruments and sequencing modes. These metrics can be analyzed on the sequencers themselves or by using tools like Illumina's Sequencing Analysis viewer. Typically, the total output of the sequencing run is analyzed as well as quality scores. The overall quality score (Q30) is a measure of the run quality, it is defined as threshold for the percentage of bases that should be called with a quality score of 30 or higher.

The Q-score or per base sequencing quality score is a measure of the probability to call a base incorrectly whereby higher Q-scores indicate a lower probability for incorrect base calling. A Q-score of 30, (Q30) indicates a base calling accuracy of 99.9 %. Lower Q-scores can result in a significant percentage of unusable reads and may result in inaccurate conclusions due to a reduction in base calling accuracy.

Further, cluster densities and reads passing filter (PF) can be analyzed. This internal "chastity filter" is passed in the first 25 cycles and serves to removes unreliable clusters from the image analysis results.

For example, for a NextSeq500 run in high output paired-end 75 mode, Illumina specifications state that 80 % of bases should have a quality score of ≥ 30 (80 % ≥ Q30). The expected data output should be between 50 – 60 giga bases (Gb) at cluster densities between 129 and 165 k/mm^2 clusters passing filter.

All of these metrics should be analyzed and kept within Illumina specifications for optimal sequencing results as over- and under-clustering during the sequencing run can decrease the data quality. For more information, please refer to the Illumina website and specifications.

Base calling and demultiplexing

During sequencing, Illumina instruments generate raw data files in binary base call (BCL) format. The advantage of using this format is that each base is recorded in the exact moment when it is called ensuring efficient data processing by the sequencer. These BCL files are most commonly converted into FASTQ files for downstream analysis. BCL to FASTQ conversion is achieved using Illumina's proprietary software, bcl2fastq. As many samples are usually multiplexed in a single run and sequenced simultaneously, the data needs to be sorted again to distinguish the samples it originated from in a process called demultiplexing (see Chapter 9 for details on multiplex sequencing). The bcl2fastq software therefore demultiplexes reads into FASTQ files based on sample indices. Optionally, it is possible to attempt to correct index sequence errors during the demultiplexing step. Several alternative tools are available for demultiplexing and index error correction, including Lexogen's iDemux which can maximize sequencing output in combination with a sophisticated Unique Dual Index Set.

For a single-read sequencing run, one FASTQ file per sample is produced. Two FASTQ files per sample are created for a paired-end sequencing run, one file for read 1 and another file for read 2. Because FASTQ files are text-based files, they share some resemblance to FASTA files. However, FASTQ files consist of a four lines-per-sequence format, while FASTA files contain two lines-per sequence. These two additional lines contain information, such as quality scores, which describe the statistical certainty of a specific basecall. The FASTQ format is widely accepted as the standard format for storing unaligned NGS reads and can be used as input for a wide variety of primary and secondary data analysis tools (learn more about the FASTQ format by checking the official Format Specifications). Typically, these files are compressed and stored with the file extension *.fastq.gz.

Demultiplexing and Index Error Correction

Due to the immense data output next generation sequencing produces, various samples are usually mixed in a process called multiplexing and then sequenced as a pool. Each sample is barcoded via short, defined sequences that uniquely identify a given sample. Demultiplexing refers to the process of bioinformatically reversing this pooling step. During this process sequencing reads are associated to the samples they originated from based on these index (barcode) sequence tags and sorted into individual files.

Index sequence errors that have occurred during the sequencing workflow can be corrected when the respective indexing strategy was chosen.

Dual index sequencing offers the best chance to identify errors in the index sequence and salvage the reads for later analysis. Once identified, index sequence errors can be corrected. Sophisticated index sequence designs allow identification and correction of more errors and thus can make more reads accessible for further analysis thereby increasing sequencing data output.

Explore Chapter 9 to read more about the principles of index sequence design. Several tools are available for demultiplexing with and without error correction, e.g., Illumina's bcl2fastq software can also perform the demultiplexing step.

Lexogen has generated an alternative tool, iDemux which is freely available on github. iDemux can demultiplex indices in the i7 and i5 position as well as i1 inline indices that are part of the read. The program was originally designed to demultiplex Quantseq-Pool libraries which can be triple-indexed and thus contain all three index types. By allowing for simultaneous demultiplexing and error correction of all indices, the tool saves valuable processing time and can maximize data output by rescuing reads with index errors.

While error correction performs best with Lexogen UDIs, the tool is highly flexible and can be used for demultiplexing of any index and is also compatible with barcode sequences from other vendors.

UMI extraction

When Unique Molecular Identifiers (UMIs) are incorporated into samples during library preparation, their sequences must be extracted from the FASTQ reads bioinformatically (see also Chapter 8 on UMIs). Failing to remove UMIs from sequencing reads can significantly reduce alignment rates when mapping against a reference genome thus increasing the number of potential mismatches. Typically, this is achieved by "splicing out" the UMI sequence from the read (Fig. 2). Subsequently, the UMI sequence is added into the header of the read. This method retains the UMI sequence of each read without interfering with alignment.

Most bioinformatic tools with UMI functionality are designed to simultaneously extract UMI sequences from FASTQ reads and collapse PCR duplicates post alignment. Commonly, collapsing of UMIs relies on positional information gathered during read mapping, i.e., the mapping coordinates of the read associated with the UMI. We will return to UMI collapsing in Chapter 12 when we focus on secondary data analysis.

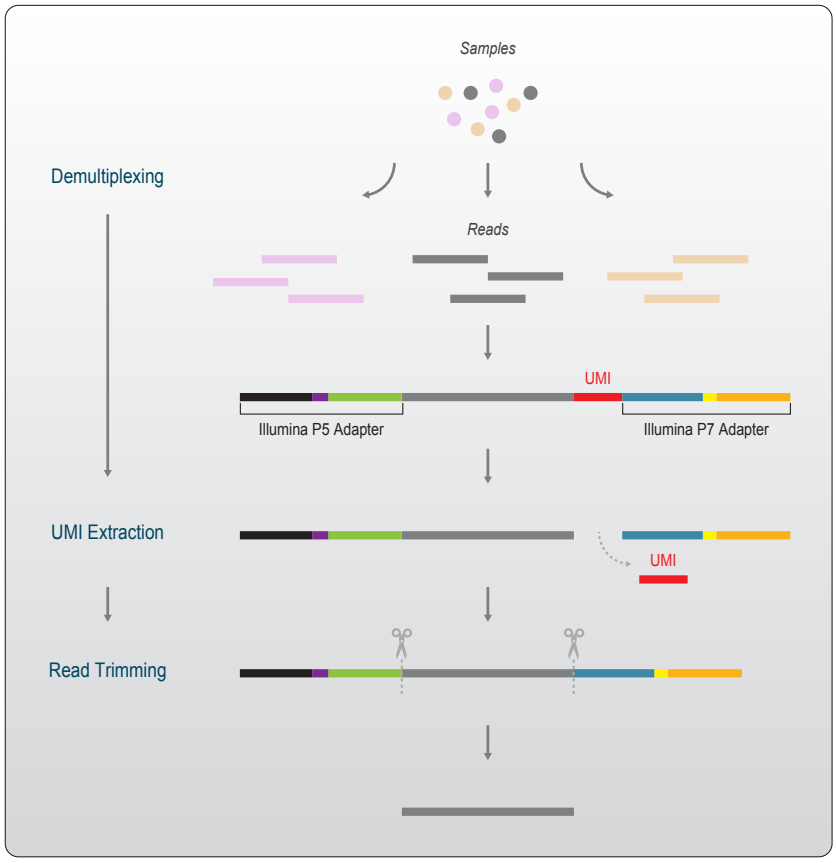


Figure 2 | Primary data analysis. Following the demultiplexing step, UMI sequences are extracted from the read and written into the FASTQ header. UMIs can be located at either end of the sequencing read and their positions is typically specified in the command used for UMI extraction. In the next step, adapter sequences and sequences of low quality, such as homopolymer stretches are trimmed. Failing to remove the UMI or adapter sequences negatively influences read mapping and can lead to low alignment rates.

Read trimming

Prior to mapping reads against a reference genome, it is recommended that the user performs read trimming. Often, Next Generation Sequencing (NGS) reads contain undesirable adapter contamination, poly(A), or poor-quality sequences which should be removed. Failing to remove these problematic sequences may result in reduced alignment rates or false alignments. When utilizing an Illumina sequencer with 2-channel chemistry, it is also advisable to trim poly(G) sequences. These poly(G) sequences result from an absence of signal and will default to G (Fig. 3).

Two popular tools for read trimming are cutadapt[1] and Trimmomatic[2] .

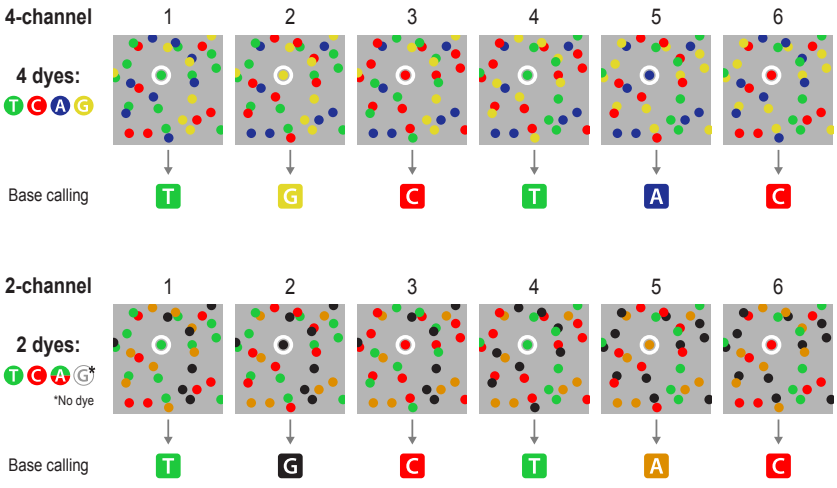


Figure 3 | Base calling for Illumina sequencers using 4-channel chemistry (top) or 2-channel chemistry (bottom). While all four nucleotides are labeled with a fluorescent dye when 4-channel chemistry is used, only two dyes are used to distinguish the four nucleotides when 2-channel chemistry is used. Here, T is labeled green, C is labeled red, and A is labeled both green and red. As red and green overlay for A, a signal can be detected in both channels clearly identifying the base. In contrast, G is unlabeled and therefore, does not generate a signal in any of the channels. Sequences of low quality that fail to generate a signal and remain dark will be called as "G" per default. Poly(G) sequences are therefore often seen in sequencing data from instruments using 2-channel chemistry. Trimming of these sequences improves the quality of the data for subsequent analysis steps.

2. Quality Control

When analyzing sequencing data, it is essential that issues are detected early on. Detecting errors prior to analysis saves valuable time and resources and ensures sound biological conclusions are made. Concordant to the concept "garbage in, garbage out"(Fig. 4), one cannot expect to generate biological meaningful results in tertiary analysis when processing fundamentally flawed data during primary and secondary analysis.

Therefore, strict quality control at each analysis step must be performed to thoroughly understand the strengths and weaknesses of a data set and ensure conclusions are made in good scientific practice.

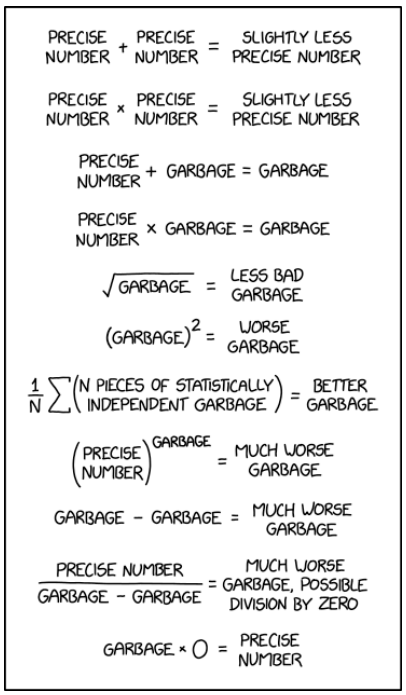


Figure 4 | Input / Output Concept. The quality of the data and control of the processing steps is fundamental for drawing sound conclusions. As the "garbage in, garbage out" concept signifies, flawed data will lead to flawed conclusions. Image curtesy of xkcd comics.

Quality controlling FASTQ reads with FastQC

The most commonly used tool to quality control FASTQ reads is FastQC[3]. FastQC presents the user with a report containing a variety of relevant summary statistics. Each statistic on the FastQC report is scored according to a 'traffic light' system. Green indicates normal data, orange is borderline or slightly abnormal, and red represents unusual or poor-quality data.

At Lexogen, FastQC is run before and after each primary analysis step to determine the impact of the analysis step performed, as well as the effect that data quality will have on downstream analysis. While acceptable and often unavoidable to have some yellow or red warnings during or at the end of primary analysis, the workflow should be optimized to maximize the number of "green lights". Due to the unique library preparation chemistry inherent to Lexogen products, certain FastQC summary statistics may be flagged as unusual in the report. However, this is normal and expected.

For example, when using a library preparation protocol with UMIs and running FastQC before trimming or UMI extraction, the "Per Base Sequence Content" will typically display a red light. Adapter or UMI sequences naturally bias the "Per Base Sequence Content" and will be flagged as unexpected. Products from the QuantSeq family may also display a warning for the "Sequence Duplication Levels" statistic. This is because FastQC was originally designed as a quality control method for whole-genome shotgun sequencing data where high sequence diversity is key. For 3' transcriptome sequencing methods such as QuantSeq, reads are concentrated at the 3' UTR, which results in an overall lower sequence diversity and overestimation of potential PCR duplicates.

Software specific diagnostic output

In addition to examining FastQC results, it is also advisable to study the diagnostic output of the different bioinformatic tools used.

For example, Bcl2fastq will output the number of reads demultiplexed for each sample. Unexpected ratios, or a very low number of reads for all samples can indicate incorrect sample barcoding or that the wrong barcodes have been supplied to the software. Similarly, trimming software such as Cutadapt or Trimmomatic also provides diagnostic output on the percentage of total reads and bases trimmed, as well as the average length of the remaining reads.

After successful pre-processing and read quality control, the reads are prepared for the next analysis steps. Stay tuned and read up on Secondary Data Analysis in Chapter 12 of our RNA Lexicon.

Literature:

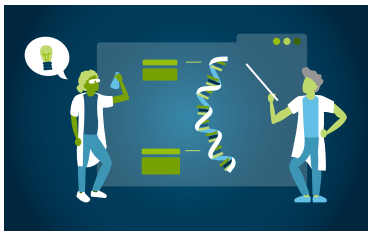
1. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17: 10-12. DOI: 10.14806/ej.17.1.200

2. Bolger AM, Lohse M, Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114-2120. DOI: 10.1093/bioinformatics/btu170, PMID: 24695404; PMCID: PMC4103590.

3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

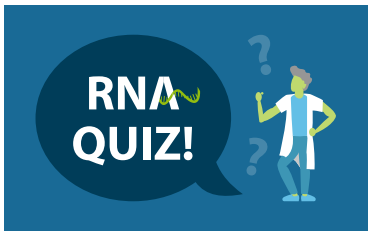
LEXOGEN
The RNA Experts

Curious to learn more?



Explore more chapters in our RNA LEXICON:
www.lexogen.com/rna-lexicon

Watch our RNA EXPERTise Videos:
www.lexogen.com/rna-expertise-videos



Show your RNA expertise and master all questions of our RNA Quiz:
www.lexogen.com/lexicon-quiz-5

Lexogen GmbH
Campus Vienna Biocenter 5
1030 Vienna, Austria
Telephone: +43 (0) 1 345 1212
Fax: +43 (0) 1 345 1212-99
info@lexogen.com
www.lexogen.com

Lexogen, Inc.
51 Autumn Pond Park
Greenland, NH 03840, USA
Telephone: +1-603-431-4300
Fax: +1-603-431-4333