Jan 09, 2020

Sequencing and data quality control

In 1 collection

Adriana Alberti[1], Julie Poulain[1], Stefan Engelen[1], Karine Labadie[1], Sarah Romac[2,3], Isabel Ferrera[4], Guillaume Albini[1], Jean-Marc Aury[1], Caroline Belser[1], Alexis Bertrand[1], Corinne Cruaud[1], Corinne Da Silva[1], Carole Dossat[1], Frédéric Gavory[1], Shahinaz Gas[1], Julie Guy[1], Maud Haquelle[1], E'krame Jacoby[1], Olivier Jaillon[1,5,6], Arnaud Lemainque[1], Eric Pelletier[1], Gaëlle Samson[1], Marc Wessner[1], Genoscope Technical Team[1], Silvia G. Acinas[4], Marta Royo-Llonch[4], Francisco M. Cornejo-Castillo[4], Ramiro Logares[4], Beatriz Fernández-Gómez[4,7,8], Chris Bowler[9], Guy Cochrane[10], Clara Amid[10], Petra Ten Hoopen[10], Colomban De Vargas[2,3], Nigel Grimsley[11,12], Elodie Desgranges[11,12], Stefanie Kandels-Lewis[13,14], Hiroyuki Ogata[15], Nicole Poulton[16], Michael E. Sieracki[16,17], Ramunas Stepanauskas[16], Matthew B. Sullivan[18,19], Jennifer R. Brum[19,20], Melissa B. Duhaime[21], Bonnie T. Poulos[22], Bonnie L. Hurwitz[23], Stéphane Pesant[24,25], Eric Karsenti[9,13,26], Patrick Wincker[1,5,6]

[1]CEA - Institut de Biologie François Jacob, Genoscope, Evry, France, [2]CNRS, UMR 7144, Station Biologique de Roscoff, France, [3]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, France, [4]Departament de Biologia Marina i Oceanografia, Institute of Marine Sciences (ICM), CSIC, Barcelona, Spain, [5]CNRS, UMR 8030, Evry , France, [6]Université d'Evry, UMR 8030, Evry, France, [7]FONDAP Center for Genome Regulation, Santiago, Chile, [8]Laboratorio de Bioinformática y Expresión Génica, Instituto de Nutrición y Tecnología de los Alimentos (INTA), Universidad de Chile, El Libano Macul, Santiago, Chile, [9]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, Paris, France, [10]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genomes Campus, Hinxton, Cambridge , UK, [11]CNRS UMR 7232, BIOM, Banyuls-sur-Mer, France, [12]Sorbonne Universités Paris 06, OOB UPMC, Banyuls-sur-Mer , France, [13]Directors' Research European Molecular Biology Laboratory, Heidelberg, Germany, [14]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany, [15]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, Japan, [16]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine, USA, [17]National Science Foundation, Arlington, Virginia, USA, [18]Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, USA, [19]Department of Microbiology, The Ohio State University, Columbus, Ohio, USA, [20]Present address: Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, Louisiana, USA, [21]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, USA, [22]University of Arizona, Tucson, Arizona, USA, [23]Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, Arizona, USA, [24]MARUM, Center for Marine Environmental Sciences, University of Bremen, Germany, [25]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Germany, [26]Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, Villefranche-sur-mer, France

1 Works for me dx.doi.org/10.17504/protocols.io.qwjdxcn

Adriana Alberti
CEA, Genoscope, France

ABSTRACT

This protocol describes the sequencing and data quality control for the *Tara* Oceans expedition and is part of Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition.

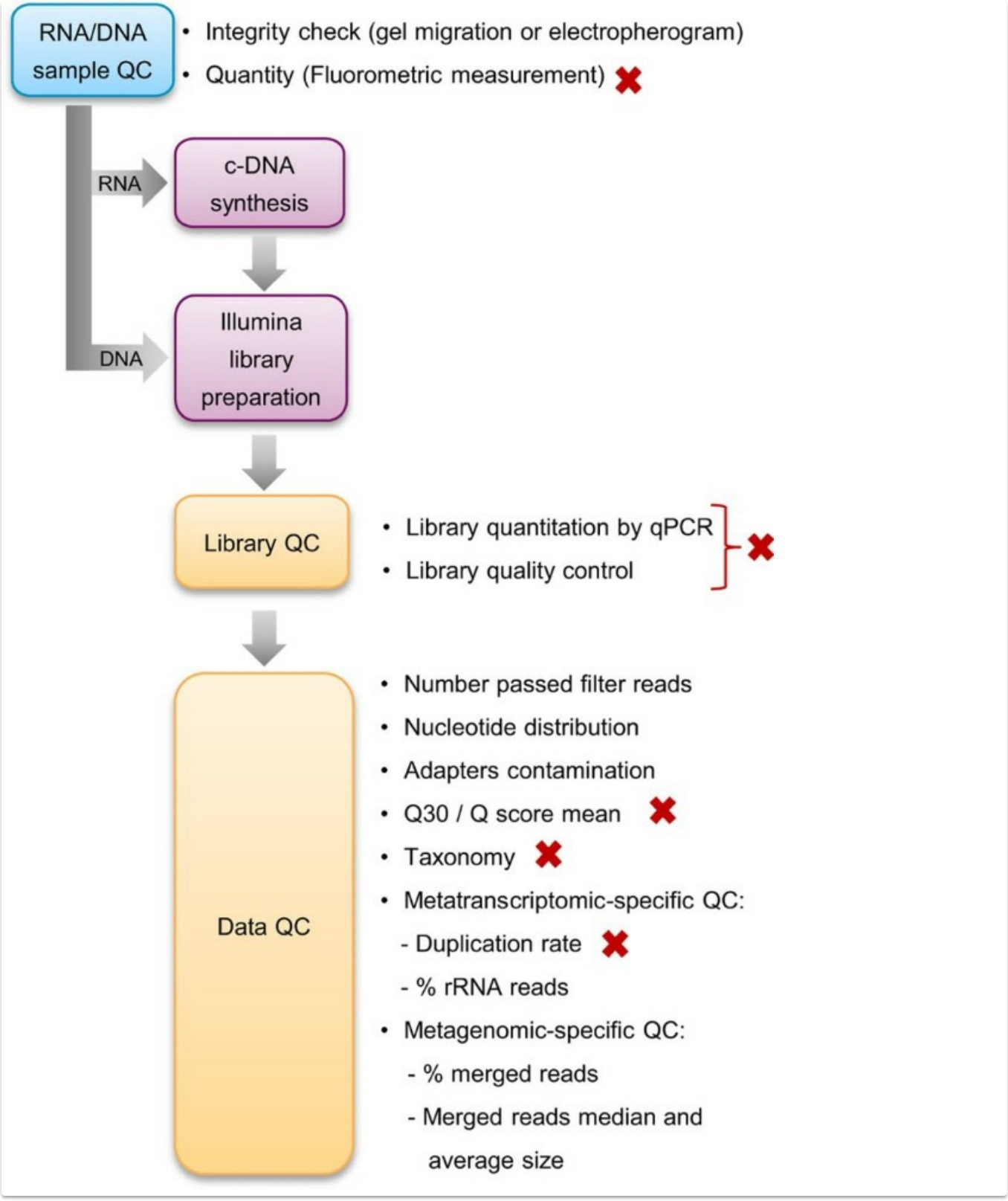

Figure 3: Overview of experimental pipeline from nucleic acids to sequences. (Red crosses highlight QC steps where experiments can be stopped.)

EXTERNAL LINK

https://www.nature.com/articles/sdata201793#methods

THIS PROTOCOL ACCOMPANIES THE FOLLOWING PUBLICATION

Alberti, A. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data* **4**, 170093 (2017)
doi: 10.1038/sdata.2017.93

ATTACHMENTS

Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition.pdf

technote_Q-Scores.pdf

GUIDELINES

1. Sequencing library quality control

All libraries were quantified first by Qubit dsDNA HS Assay measurement and then by qPCR with the KAPA Library Quantification Kit for Illumina Libraries (Kapa Biosystems) on an MXPro instrument (Agilent Technologies). Library profiles were assessed using the DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer. Later on, the quality control step was implemented with quantification by PicoGreen method on 96-well plates and high throughput microfluidic capillary electrophoresis system for library profile analysis (LabChip GX, Perkin Elmer, Waltham, MA).

2. Sequencing procedures

Libraries concentrations were normalized to 10 nM by addition of Tris-Cl 10 mM, pH 8.5 and then applied to cluster generation according to the Illumina Cbot User Guide (Part # 15006165). Libraries were sequenced on Genome Analyzer IIx, HiSeq2000 or HiSeq2500 instruments (Illumina) in a paired-end mode. Read lengths were chosen in order to produce data fitting with bioinformatics analyses needs (Table 2).

Table 2: Summary of libraries generated from *Tara* Oceans DNA and RNA samples and sequencing experiments performed on each type of library.

Sequencing library preparation method	Library insert size (pb)	Sequencing instrument	Read length (PE mode)	Generated libraries*	Mean number of reads per sample (millions of paired reads)
Metagenomics from size fractionated filters (Section 4.1)	180	HS2000	101	855	160
Metagenomics from viral samples (Section 4.2)	150–900	HS2000	101	90	50
SAGs (Section 4.3)	150–900	HS2000	101	49	20
Metatranscriptomic libraries (Section 4.4)	100–600	HS2000	101	467	160
18S metabarcode libraries (Section 4.5)	160	GAIIx	151	884	1.5
16S metabarcode libraries (Section 4.6)	400	HiSeq2500	251	In progress	ND

*
Number of libraries with available readsets in public databases at the date of publication of the paper.

Metabarcoding and metatranscriptomic libraries were characterized by low diversity sequences at the beginning of the reads related respectively to the presence of primer sequence used to amplify 18S and 16S tags and low complexity polynucleotides added during cDNA synthesis. Low-diversity libraries can interfere in correct cluster identification, resulting in drastic loss of data output. Therefore, loading concentrations of these libraries (8–9 pM instead of 12–14 pM for standard libraries) and PhiX DNA spike-in (10% instead of 1%) were adapted in order to minimize the impacts on the run quality.

Sequencing was performed according to the Genome Analyzer IIx User Guide (Part # 15018814), HiSeq2000 System User Guide (Part # 15011190) and HiSeq2500 System User Guide (Part # 15035786).

Data quality control and filtering

A first step in data quality control process was the primary analysis performed during the sequencing run by Illumina Real Time Analysis (RTA) software (Code availability 1). This tool analyses images and clusters intensities and filters them to remove low quality data. Furthermore, it performs basecalling and calculates Phred quality score (Q score), which indicates the probability that a given base is called incorrectly. Q score is the most common metric used to assess the accuracy of the sequencing experiment (http://www.illumina.com/documents/products/technotes/technote_Q-Scores.pdf). After conversion of raw BCL files generated by RTA to fastq demultiplexed data by Illumina bcl2fastq Conversion software (Code availability 2), in-house filtering and quality control treatments developed in Genoscope were applied to reads that passed the Illumina quality filters (named raw reads). The parameters of

these controls are indicated in Fig.2.

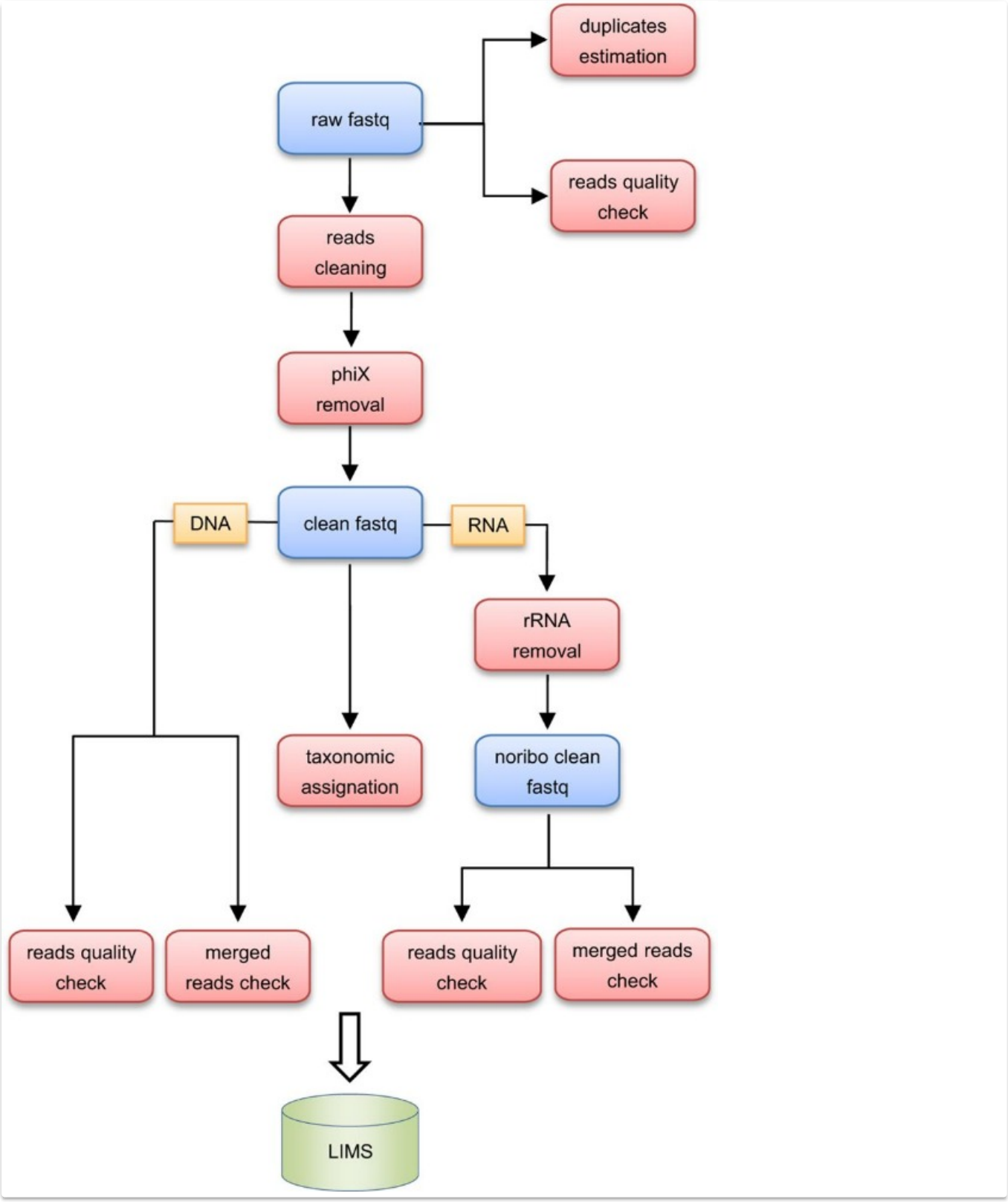

Figure 2: Data processing flowchart.

This processing allows obtaining high quality data and improves subsequent analyses.

Filtering steps were applied on whole raw reads as shown in Steps 1-3.

Data quality control was performed on random subsets of 20,000 reads before (raw reads) and/or after filtering steps (clean reads) as shown in Steps 4-8.

Code availability

1. Real Time Analysis software: http://support.illumina.com/sequencing/sequencing_software/real-time_analysis_rta/downloads.html

2. Conversion: http://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html

3. Fastx_clean software, http://www.genoscope.cns.fr/fastxtend

4. *FASTX-Toolkit*, http://hannonlab.cshl.edu/fastx_toolkit/index.html

5. fastx_estimate_duplicate software, http://www.genoscope.cns.fr/fastxtend

6. fastx_mergepairs software, http://www.genoscope.cns.fr/fastxtend

Filtering

1 Remove the sequences of the Illumina adapters and primers used during library construction from the whole reads. Remove low quality nucleotides with quality value<20 from both ends. Keep the longest sequence without adapters and low quality bases. Trim sequences between the second unknown nucleotide (N) and the end of the read. Discard reads shorter than 30 nucleotides after trimming.

These trimming steps are achieved using fastx_clean (Code availability 3), a software based on the FASTX library (Code availability 4).

2 Remove the reads and their mates that map onto run quality control sequences (Enterobacteria phage PhiX174 genome, Data Citation 1: GenBank NC_001422.1) using SOAP aligner.

3 Apply a specific filter aiming to remove ribosomal reads to data generated from metatranscriptomic libraries sequencing. In *Tara* Oceans project, the reads and their mates that map onto a ribosomal sequences database are filtered using SortMeRNA v 1.0 (ref.), a biological sequence analysis tool for filtering, mapping and OTU-picking NGS reads. It contains different rRNA databases and we use it to split the data into two files: rRNA reads in a file (ribo_clean) and other reads in another file (noribo_clean).

Data quality control

4 Estimate duplicated sequences rates from single and paired sequences on raw reads and cleaned reads (after filtering steps), using fastx_estimate_duplicate (Code availability 5), a software based on the FASTX library.

The following steps (4, 5 and 6) are performed on a subset of randomized 20,000 reads

5 Perform taxonomic assignation by aligning with Mega BLAST (Blast 2.2.15 suite) a subset of 20,000 reads against the nt database (http://www.ncbi.nlm.nih.gov/nucleotide), and using Megan software (version 3.9).

6 Do the merging step with fastx_mergepairs (Code availability 6), a software based on the fastx library. Extract the first 36 nucleotides of read2 and perform alignment between that seed and read1. Launch merging if the alignment was at least of 15 nucleotides, with less than 4 mismatches and an identity percent of at least 90%. For each overlapping position, retain the nucleotide of higher quality.

Final data quality report

7 Calculate read size, quality values, N positions, base composition and known adapters sequences detection before (raw reads) and after filtering the reads (cleaned reads). Evaluate each dataset using specific toolboxes generated from this pipeline (see Technical validation paragraph in paper).