README _EPGV_DataTransfer_Illumina® Sequencing

I. Delivered files / Paired-ends (PE) sequences

- 1 Transfer directory : `Project_version_EPGV_QC-Cleaning_dateOfTransfer`
  - 2 files in the root directory (readme and data synthesis)
  - As many sub-directories as PE sequences (14 data files)

(See Figure 1: organization of the directory with one sub-directory)

Nomenclature of sub-directory

FC_ (R or 00x) _s_x (Flowcell barcode, lane (s_x) ... etc)
Where,

- FC = Flowcell barcode
- R = Read (R sometimes replaced by Read1 or absent)
- 00x = Directory id for multiplex
- s_x = Sequences_ lane number of FC

Example:
631UJAXX_R_s_5
FC = 631UJAXX
R = Read
s_5 = Sequences_lane 5

Note: Many different nomenclatures exist and are inherent to the evolution of Illumina® processing.

II. Flowcell (FC) Nomenclature

Barcode: 9 alphanumeric characters

- FC of Genome Analyser II(GA)

					A	A	X	X
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

- FC of HiSeq 2000_ V1 (HS)

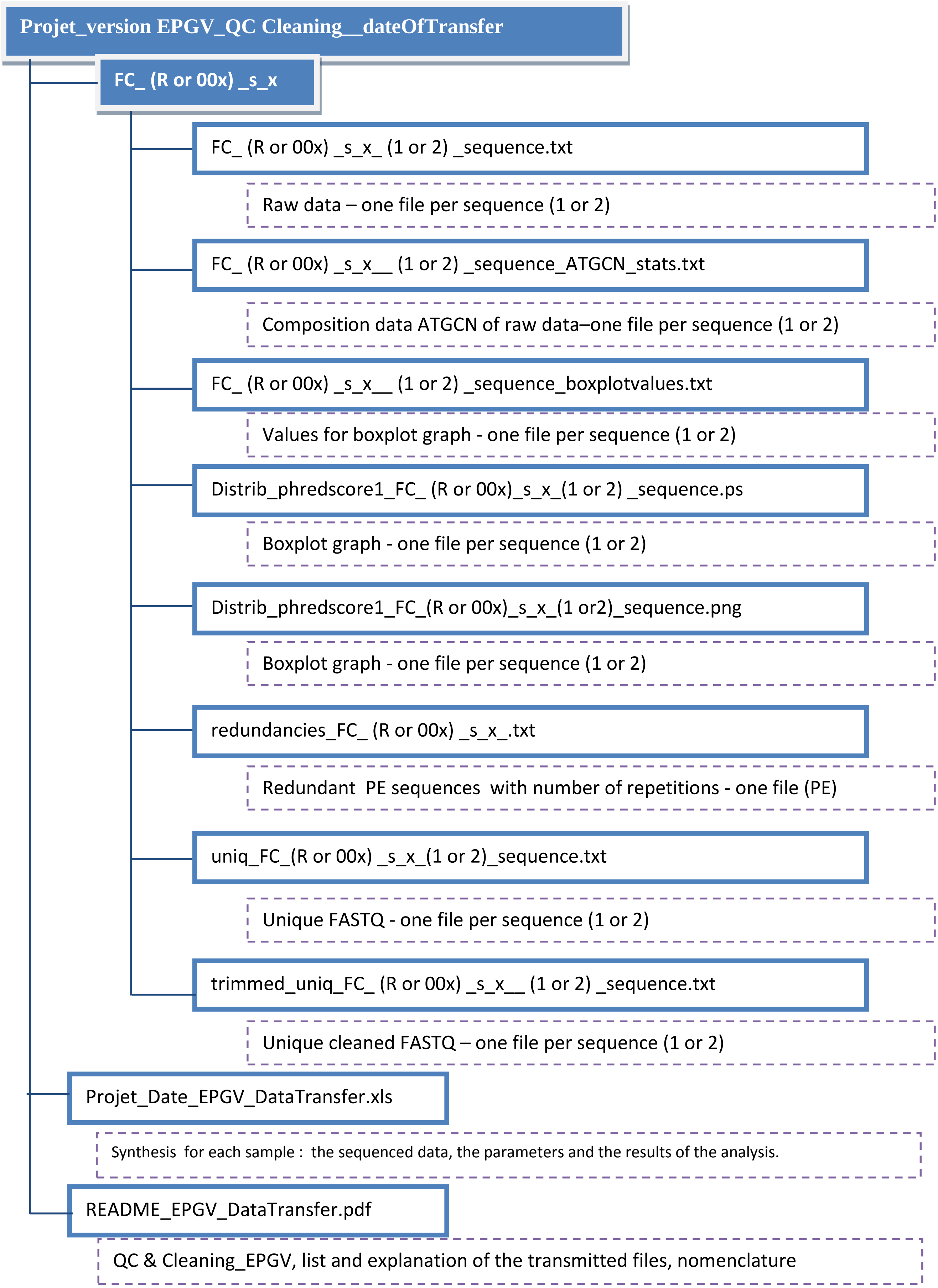
A/B						A	B	X	X
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

- FC of HiSeq 2000_ V3 (HS)

A/B						A	C	X	X
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

The letter A/B in prefix indicates the position of the FC on HiSeq : Position A or Position B.

Figure 1: Organization of the directory with one subdirectory



- **Projet_version EPGV_QC Cleaning__dateOfTransfer**
  - **FC_ (R or 00x) _s_x**
    - FC_ (R or 00x) _s_x_ (1 or 2) _sequence.txt
      - Raw data – one file per sequence (1 or 2)
    - FC_ (R or 00x) _s_x__ (1 or 2) _sequence_ATGCN_stats.txt
      - Composition data ATGCN of raw data–one file per sequence (1 or 2)
    - FC_ (R or 00x) _s_x__ (1 or 2) _sequence_boxplotvalues.txt
      - Values for boxplot graph - one file per sequence (1 or 2)
    - Distrib_phredscore1_FC_ (R or 00x)_s_x_(1 or 2) _sequence.ps
      - Boxplot graph - one file per sequence (1 or 2)
    - Distrib_phredscore1_FC_(R or 00x)_s_x_(1 or2)_sequence.png
      - Boxplot graph - one file per sequence (1 or 2)
    - redundancies_FC_ (R or 00x) _s_x_.txt
      - Redundant PE sequences with number of repetitions - one file (PE)
    - uniq_FC_(R or 00x) _s_x_(1 or 2)_sequence.txt
      - Unique FASTQ - one file per sequence (1 or 2)
    - trimmed_uniq_FC_ (R or 00x) _s_x__ (1 or 2) _sequence.txt
      - Unique cleaned FASTQ – one file per sequence (1 or 2)
  - Projet_Date_EPGV_DataTransfer.xls
    - Synthesis for each sample : the sequenced data, the parameters and the results of the analysis.
  - README_EPGV_DataTransfer.pdf
    - QC & Cleaning_EPGV, list and explanation of the transmitted files, nomenclature

III. Quality Control Process and EPGV Cleaning Version 1.7

A. Quality Control (QC)

Input: 2 FASTQ files PE (paired-ends) of raw data of the reads : Read1 and Read2.

1. Calculating the number of reads in each sequence file

2. Checking the proper structure of each file FASTQ

- Presence of 4 lines per read
- Presence of tags in the read names
- Equality of the number of bases and the number of the Phred scores values for each read

3. Statistics

- Calculation of the ATGCN composition and the total number of bases in each sequence file
- Creation of value file for the boxplot graph in each sequence file
- Creation of file containing the Phred values per cycle in each sequence file

4. Creating boxplot graph for each file

Distribution of "Phred score" values for each cycle (min 1 / 4, median, 3 / 4, max, average)

Phred 40 = 1 error per 10,000
Phred 30 = 1 error for 1000
Phred 20 = 1 error for 100

5. Searching redundancies in PE

- Searching for redundancies pairs Read1_Read2
- Creating a file containing redundant PE
- Cleaning the redundancies from raw files
  - Redundancy of each type with the highest Phred score is kept.

Output:

- Created and delivered files (see § I I. Delivered files / Paired-ends (PE) sequences)

One per sequence file : Read1 and Read2

- `ATGCN_stat.txt`
- `boxplotvalues.txt`
- `Distrib_phredscore1.ps`
- `uniq_sequence.txt`

One PE file:

- `redondances.txt`

- *Files Created not provided (for EPGV use):*
  - *`hashtable.txt` – Phred values per cycle*
  - *`QC_.log` – 1 report with the scripts output*

B. Cleaning

Input:

- 2 FASTQ files PE (paired-ends) of non-redundant data: `uniq_ (1 or 2) _sequence.txt`
- 1 file containing the cleaning parameters (that are available in the file "`Projet_Date_EPGV_DataTransfer`")

1. Calculating the number of sequences in incoming file

2. Cleaning

Cleaning sequences relative to quality (trimming) and length parameters defined in input:

Parameters:

- "qualité mini" : minimum quality score (Phred) allowed by base
- "limit score" : lower limit value of the quality score (Phred) of the base, knowing that a score = 2 is an indeterminate score for Illumina®
- "moyenne mini" : average quality scores (Phred) of the sequence
- "taille mini" : minimum size of the sequence allowed after trimming

Workflow (see. Figure 2)

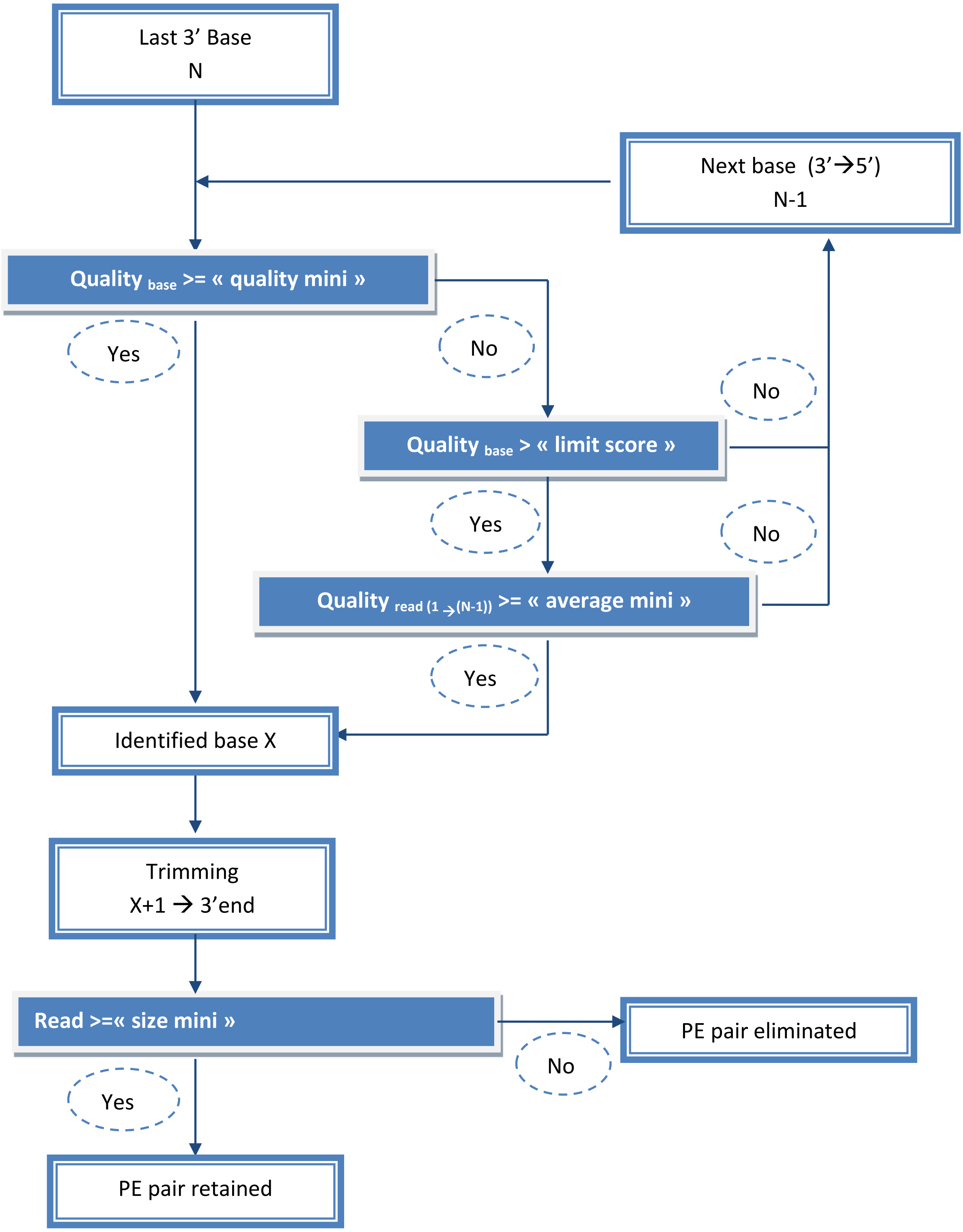
The sequences of the two reads are processed independently from its last base at 3' end screening the sequence base by base (3'-> 5 '). The goal is to define the position "base X" beyond which trimming at 3' side will be made (X +1 position until the last 3' base). Each base "N" is analyzed according to the different parameter values determined by the quality analysis, including the boxplots.

The search for the base X is as follows:

- The base "N" has a value greater than or equal to the value of "quality mini" defined (eg. qualité mini > = 30)
    - ➔ This base corresponds to the "base X": the sequence of the position "1" at 5' until this "base X" (included) will be retained after trimming.
- The base "N" has a value lower than the value of "quality mini" defined (eg. qualité mini <30):
    - o If the base "N" has a value lower than the value of "limit score" (eg. limit score < 3)
        - ➔ The review is reset on the base "N-1"
    - o If the base "N" has a value greater than the value of "limit score" and lower than the value of "qualité mini" (eg: respectively > = 3 and <30) then the average quality scores of the sequence is calculated from the position "1" at 5' until this base "N" (Not included).
        - ▪ If the average quality score is lower than the value of "average mini" (eg: moyenne mini <30)
            - ➔ The review is reset on the base "N-1"
        - ▪ If the average quality scores is greater than or equal to the value of parameter "average mini" (eg: moyenne mini > = 30)
            - ➔ This base corresponds to the "base X": the sequence from the position "1" at 5 ' until this "base X" (included) will be retained after trimming.

At the end of the trimming, if the remaining size of the read is lower than the value of the "size mini" (eg: "taille mini "<30), the associated PE pair is eliminated.

Figure 2 : Cleaning



3. Cleaning reads with N compared to the parameter ”Max Number of allowed N” (eg. “max nb N accepté”)

- The number of N is calculated separately for each Read 1 and Read 2
- If the number of N in one of the reads is greater than the "max no. N accepted " (eg: > 0), the PE pair is eliminated

4. Calculation of ATGCN composition and total number of bases for each cleaned file

Output:

- Created and delivered files (see § I I. Delivered files / Paired-ends (PE) sequences.)
  One per sequence file : Read1 and Read2
  - `trimmed_uniq.txt`
- Created files not provided (use EPGV):
  - `Trim_ . log` – report file of the process

IV. Additional information

A. Illumina® control sequences in the raw data

Séquencer	Flowcell	Library Type	PhiX174	CTRL Librairies
HiSeq	Multiplex *	TruSeq	-	+
		Autre	-	-
	Simplex **	TruSeq	+	+
		Autre	+	-
GA	Multiplex	TruSeq	-	+
		Autre	-	-
	Simplex	TruSeq	-	+
		Autre	-	-

Table 1: Presence of control sequences or control samples (CTRL) Illumina®.
+ Corresponds to the presence of the sequence and - corresponds to the absence of the sequence.
* Multiplex Flowcell: At least one lane has a multiplex of samples. Reading the index is programmed.
** Simplex Flowcell : Each lane has only one sample. Reading the index is not programmed.

B. Adapter sequences in the raw data: Library sizing-Length of the reads

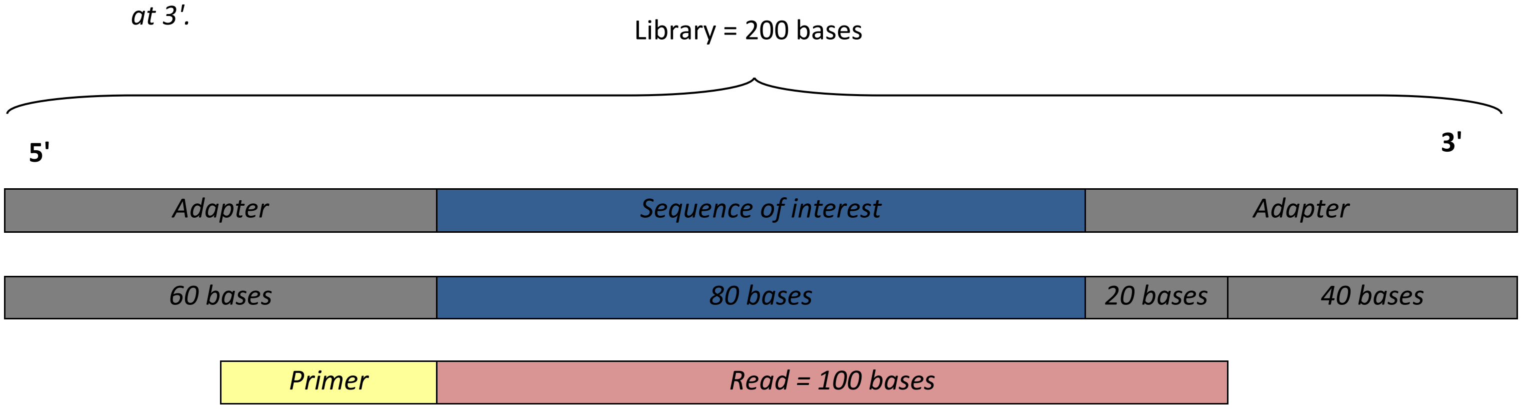
Depending on the selected size during the library sizing and on the length of reads performed during the sequencing, the obtained sequences at 3' may contain a part of the sequence of adapters (variable length). This part is variable because the sizing generates fragments of different size around an average value selected to achieve it.

The elimination of these adapters sequences is not included in the QC-Cleaning_EPGV_1.7.X.

In the present case of Illumina TruSeq® libraries conducted since January 2011 *, all fragments of a library includes 120 bases (2 * 60) which correspond to the adapters.

Thus, for an actual sequence length of 100 bases, the reads of all TruSeq libraries corresponding to fragments smaller than 220 bases include a portion of the adapter sequences because the supplementary sequence joined to the sequence of interest corresponds to the adapter sequence.

For instance take a TruSeq DNA library "sized" around 200bp. For this library,
- The size of fragments generated vary between approximately 180 and 220 bases
- The sequences corresponding to the fragments size <220 bases include X bases of adapter at 3'. Example: for fragments = 200 bases, 80 bases correspond to the sequence of interest (200-120) and 20 bases at 3'corresponds to the adapter sequence (See.Figure below)
- The sequences corresponding to fragments size > = 220 bases do not contain the adapter sequence at 3'.



C. Useful documents

- The Illumina sequences are described in the paper `[date]-Illumina-Customer-Sequence-Letter.pdf`. To receive it please contact us at support-epgv@cng.fr

- The "Validation" part of the data synthesis `Projet_Date_EPGV_DataTransfer.xls` contains the necessary information to judge the relevance of these points on the concerned samples: type of library, average size of the librairies, presumed presence of these sequences