NGS Goes Automatic :
From library preparation to data quality control

Gaëlle VILCHEZ[1], Claire KUCHLY[1], Jérôme MARIETTE[3], Frédéric ESCUDIE[3], Ibouniyamine NABIHOUDINE[3], Olivier BOUCHEZ[2], Diane ESQUERRE[2], Céline VANDERCASTEELE[2], Johanna BARBIERI[1], Céline JEZIORSKI[1], Sophie VALIERE[1], Clémence GENTHON[1], Marie VIDAL[1], Alain ROULET[1], Sandra FOURRE[2], Catherine ZANCHETTA[1], Adeline CHAUBET[1], David RENGEL[4], Denis MILAN[1], Cécile DONNADIEU[2] and Gérald SALIN[2]

1 DGA, UAR1209 INRA, Plateforme GeT-PlaGe, 31326 Castanet-Tolosan, France
2 GenPhySE, UMR1388 INRA, Plateforme GeT-PlaGe, 31326 Castanet-Tolosan, France
3 MIAT, UR875 INRA, Plateforme bioinformatique, 31326, Castanet-Tolosan, France
4 LIPM, UMR441 INRA, 31326 Castanet-Tolosan, France

Due to the constant increase of second generation sequencers (as HiSeq or MiSeq) throughput and the multiplication of their applications, a challenge is to sequence, at the same time, in a single run, ever more samples of different types, while being able to verify the quality of the produced data. At the INRA GeT-PlaGe facility, we have automated both library production and data quality control steps, in partnership with the Genotoul Bioinformatics facility, for protocols such as whole genome sequencing, Amplicon sequencing (e.g. 16S sequencing on MiSeq for metagenomics studies), stranded RNA-seq, Mate-Pair or whole genome bisulfite sequencing.
Having acquired a solid expertise in library preparation and data quality control of short fragments, our challenge for the coming months will be to integrate data from 3rd generation sequencers in our automated quality control processes, dealing with the specificity of long fragments, in partnership with the bioinformatics community of Toulouse and France Génomique.

Template Quality Control

Picogreen on QuantStudio 6
or QuBit – Nanodrop
ND8000 – Fragment Analyzer

Library preparation

TECAN EVO200
TECAN EVO150
Blue Pippin
Megaruptor

Library Quality Control

QPCR controls
TECAN EVO50 &
QuantStudio 6
Fragment Analyzer

Traceability

eSIToul

Each step of the process is tracked by various modules of our in-house information system. Biological samples are uniquely identified using barcodes.

Sequencing

HiSeq 2500, 1To, 3000
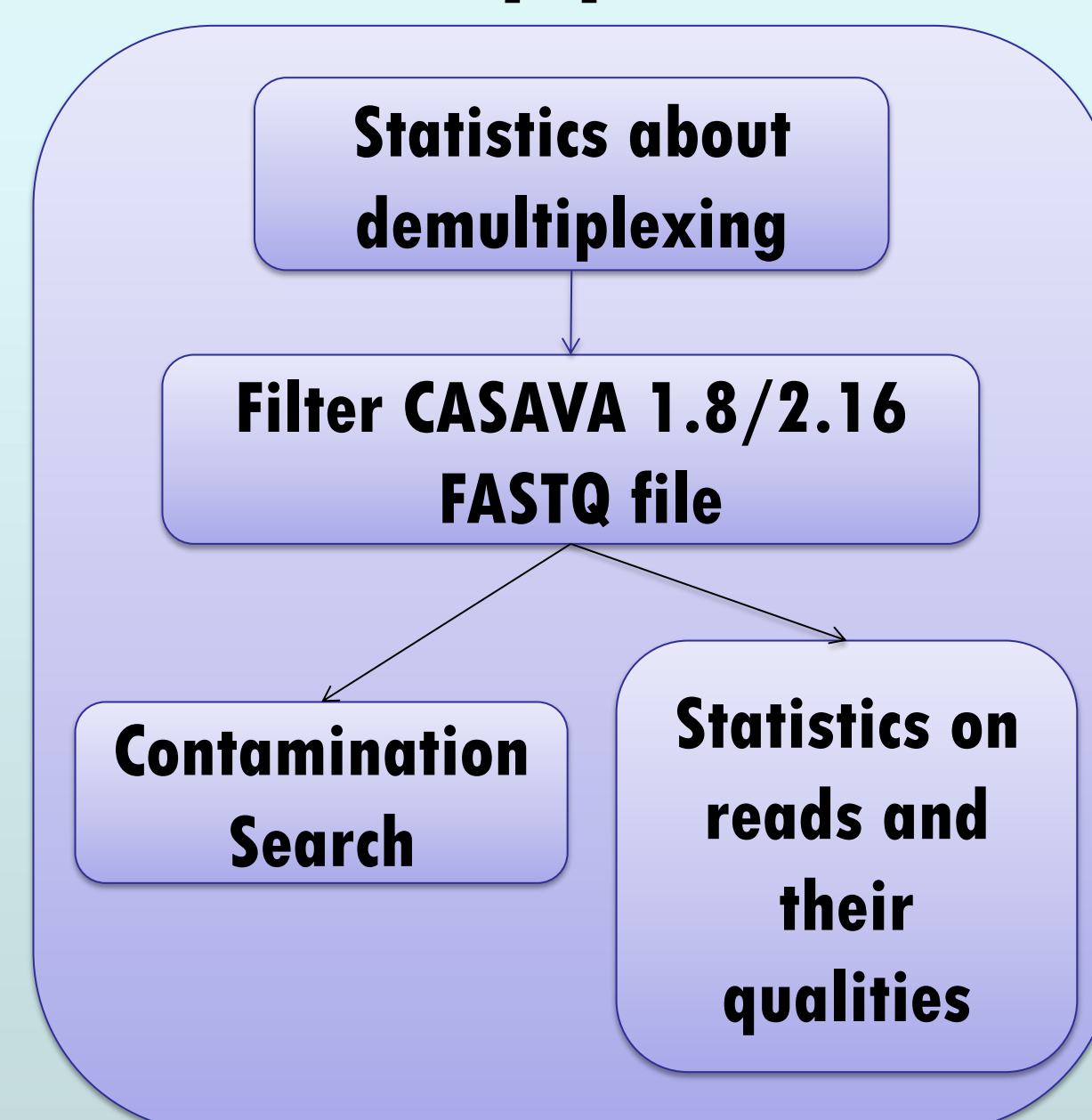MiSeq x2
PacBio RSII
MinION in Early Acces

Data quality analysis
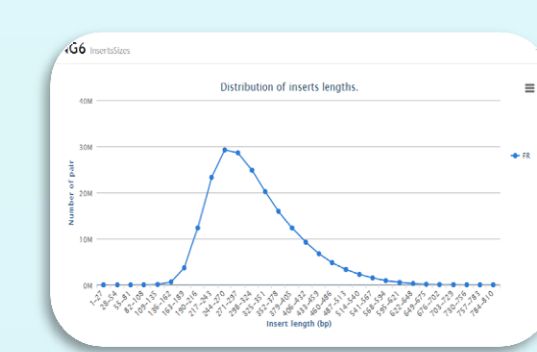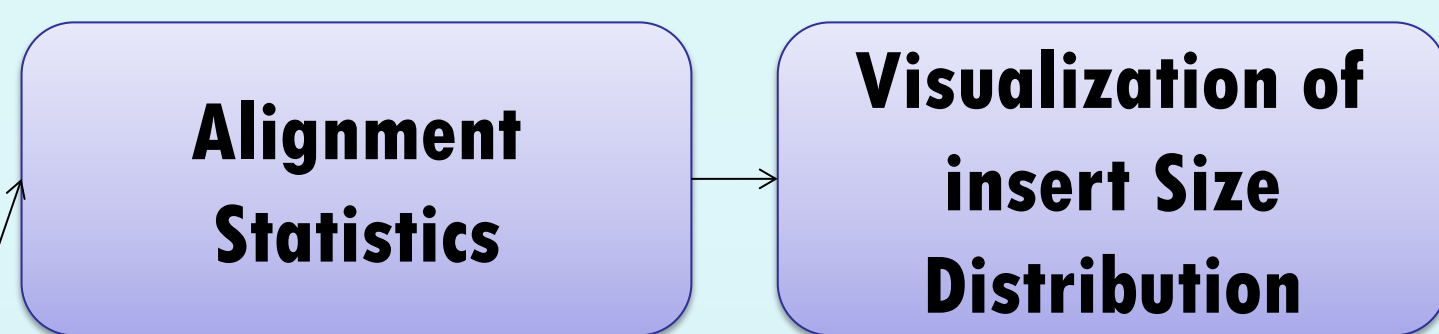## Storage of raw reads and quality control results

Main quality control workflows for Illumina's data

Core pipeline

- Statistics about demultiplexing
- Filter CASAVA 1.8/2.16 FASTQ file
- Contamination Search
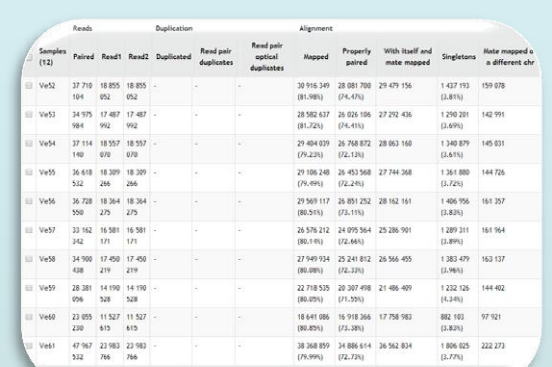- Statistics on reads and their qualities

These NGS pipelines are provided by NG6[1], an information system built upon the jflow workflow management system.

DNA or RNA Resequencing
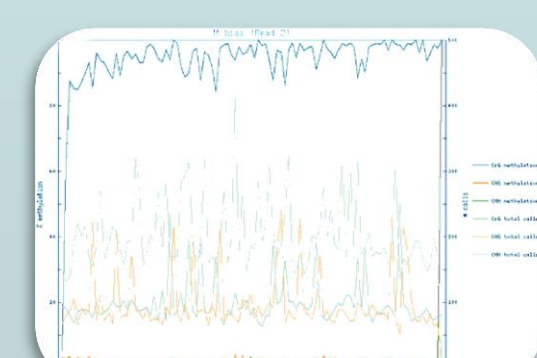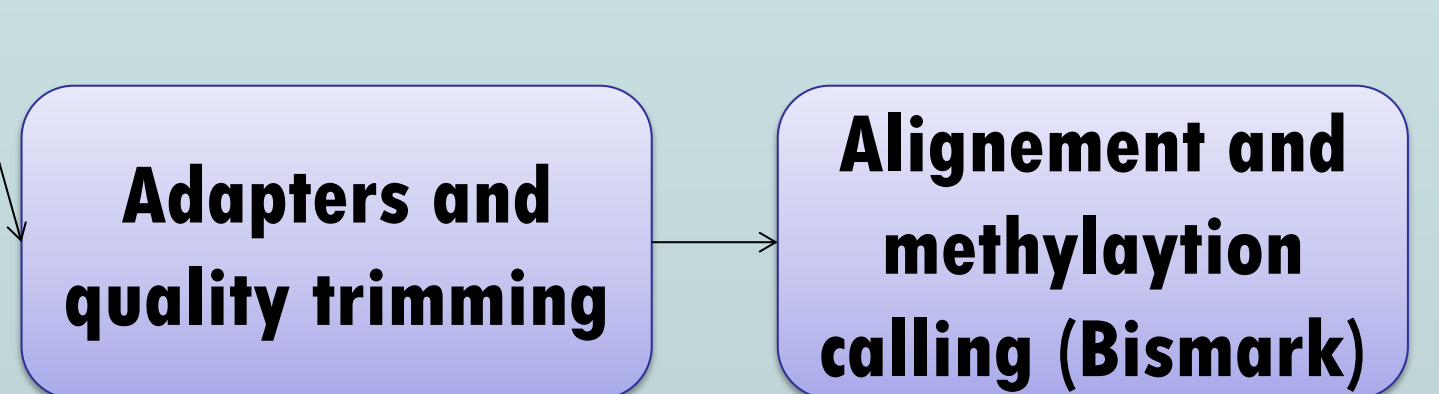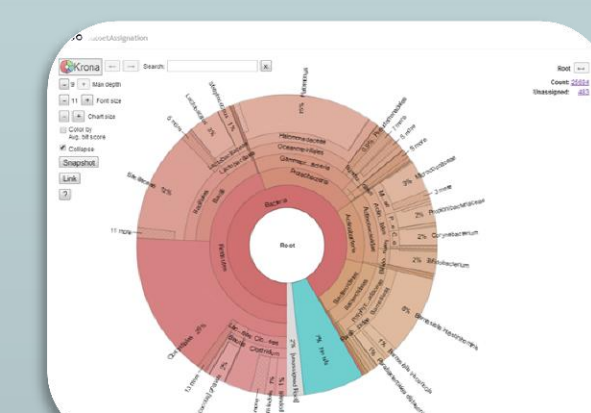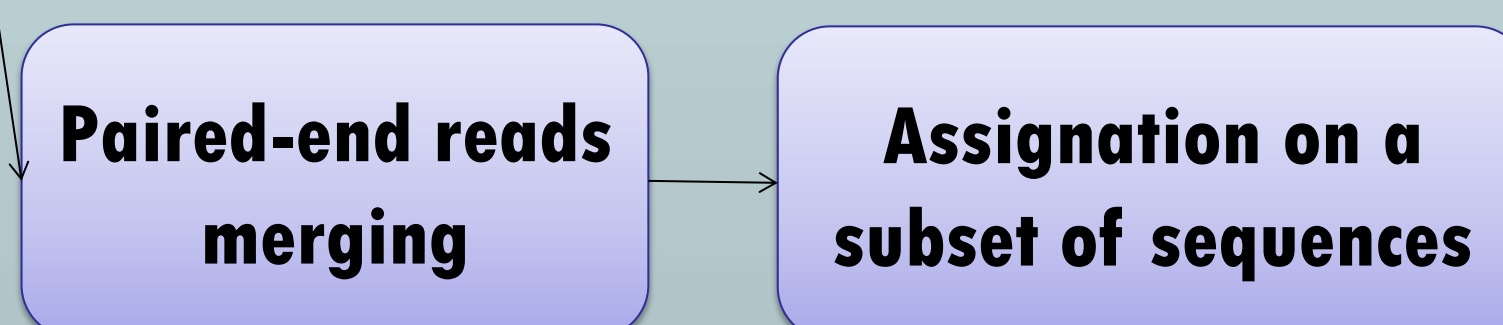
- Alignment Statistics
- Visualization of insert Size Distribution

Mate Pair Sequencing

- Nextera MatePair Adapters trimming
- Alignment Statistics
- Distribution of Insert size visualization

Whole Genome Bisulfite Sequencing

- Adapters and quality trimming
- Alignement and methylaytion calling (Bismark)

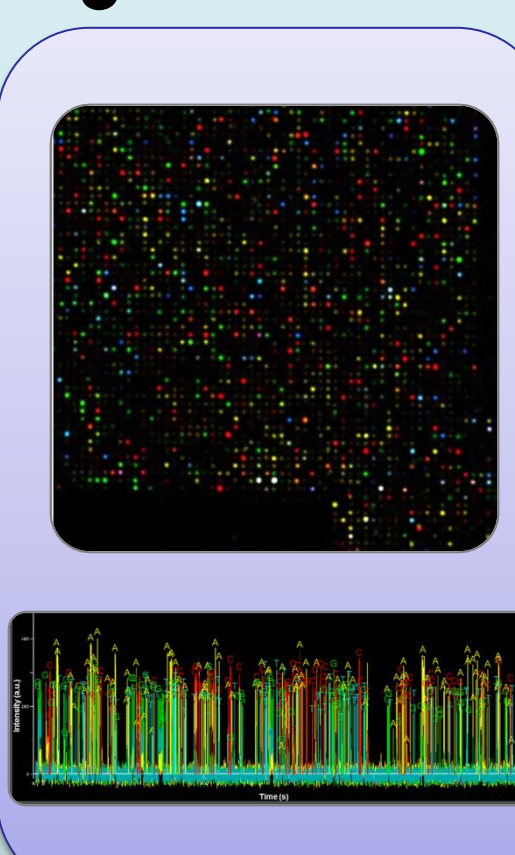### Amplicon/16S sequencing

- Paired-end reads merging
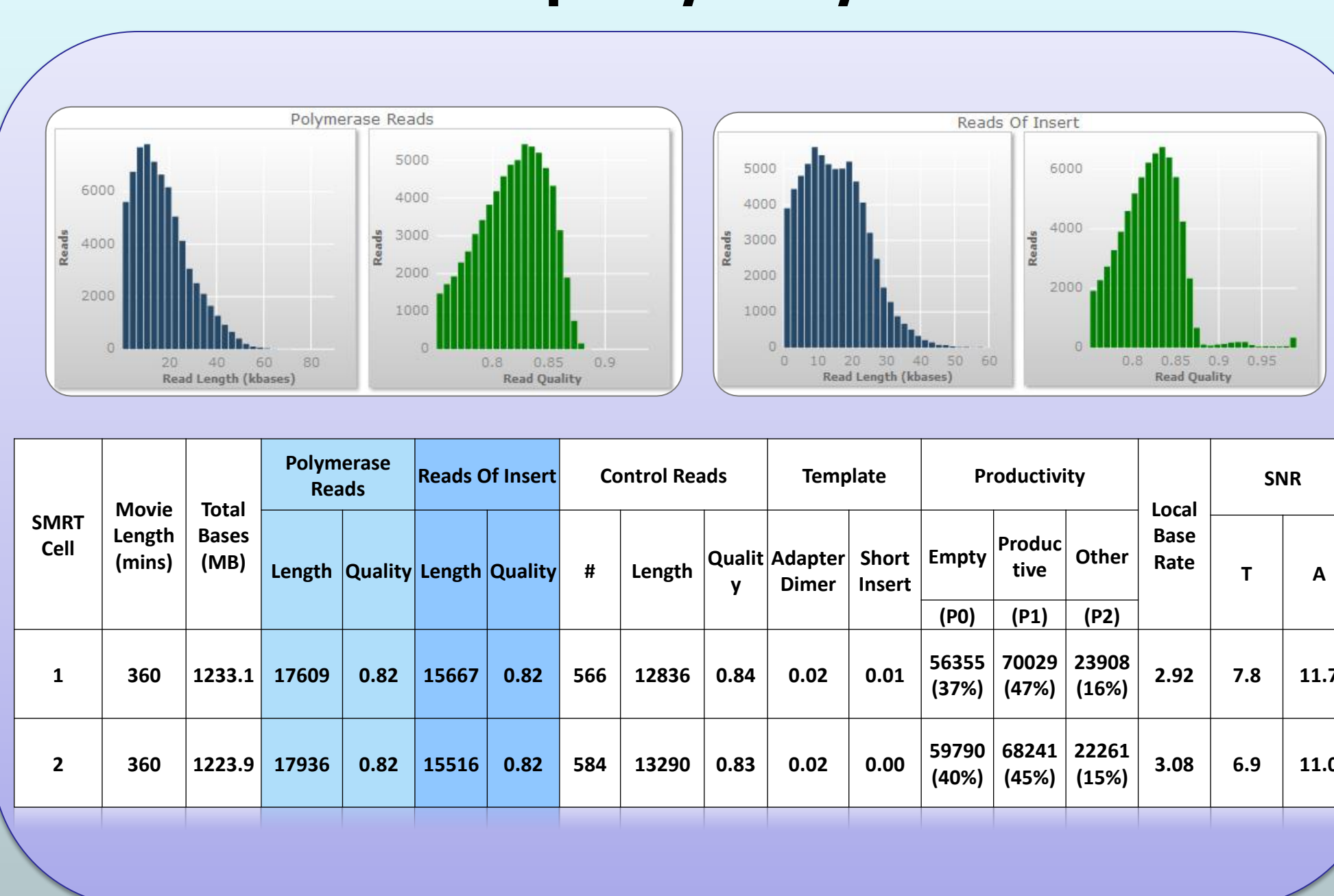- Assignation on a subset of sequences

For every specific protocol, a new workflow is built with jflow components, based on open source software or our in-house tools.
Upcoming : Sequencing RNA/DNA from a unique cell obtained with the C1 technology (Fluidigm)

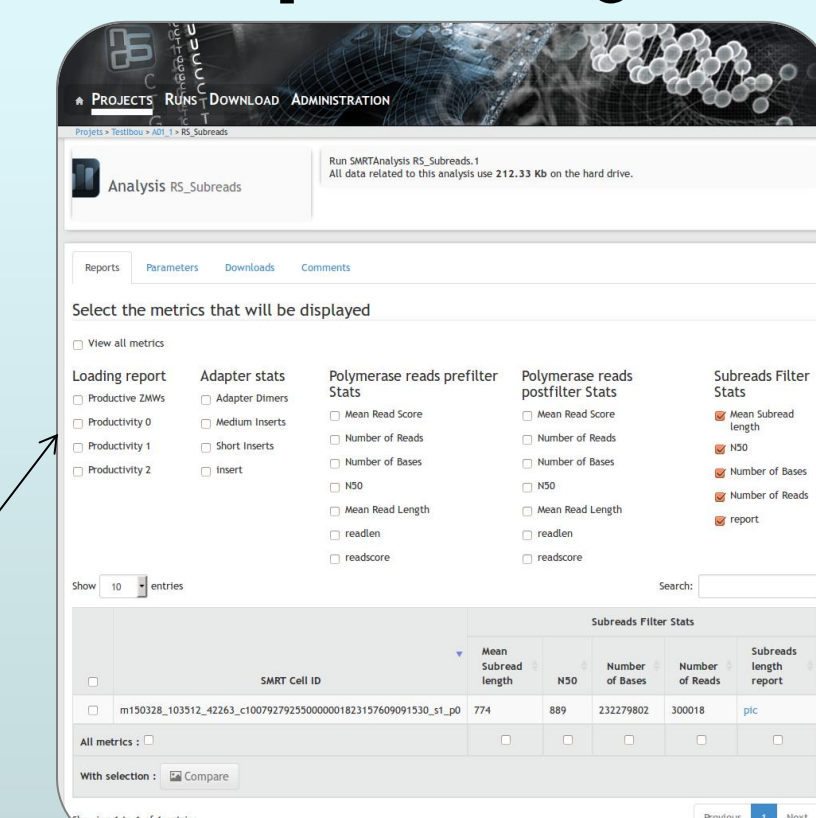## Under integration in NG6 : Main quality control workflows for PacBio's data
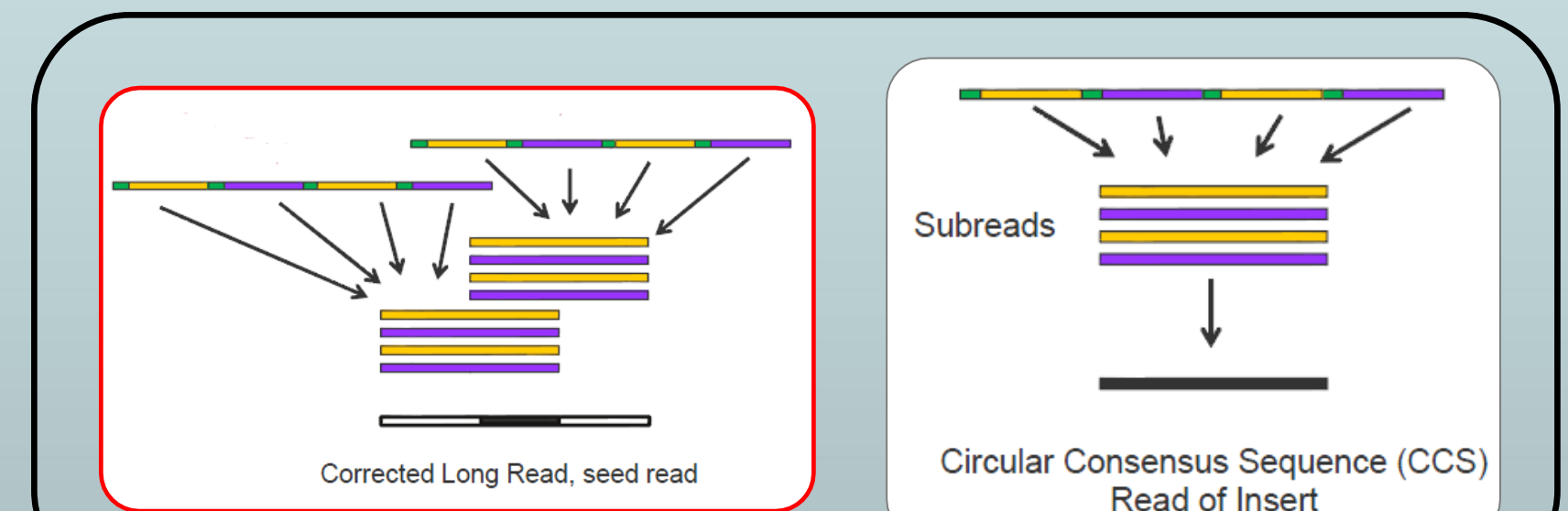
Sequencing / H5 Files generation

Data quality analysis*

Import in ng6

SMRT Cell	Movie Length (mins)	Total Bases (MB)	Polymerase Reads		Reads Of Insert		Control Reads		Template			Productivity			Local Base Rate	SNR	
			Length	Quality	Length	Quality	#	Quality	Length	Adapter Dimer	Short Insert	Empty (P0)	Productive (P1)	Other (P2)		T	A
1	360	1233.1	17609	0.82	15667	0.82	566	12836	0.84	0.02	0.01	56355 (37%)	70029 (47%)	23908 (16%)	2.92	7.8	11.7
2	360	1223.9	17936	0.82	15516	0.82	584	13290	0.83	0.02	0.00	59790 (40%)	68241 (45%)	22261 (15%)	3.08	6.9	11.0

Possibility of Reads Correction*

Corrected Long Read, seed read

Subreads

Circular Consensus Sequence (CCS) Read of Insert

* Executed by SMRT-Analysis, a Pacific Biosciences's software

Upcoming : A new LIMS for NGS samples, sequencing and analysis tracking

Data available on a common website with Genotoul Bioinformatics Facility : http://ng6.toulouse.inra.fr/

1. Mariette J, Escudie F, Allias N, Salin G, Noirot C, Thomas S, Klopp C. NG6: Integrated next generation sequencing storage and processing environment. BMC Genomics 2012, 13:462.

http://get.genotoul.fr