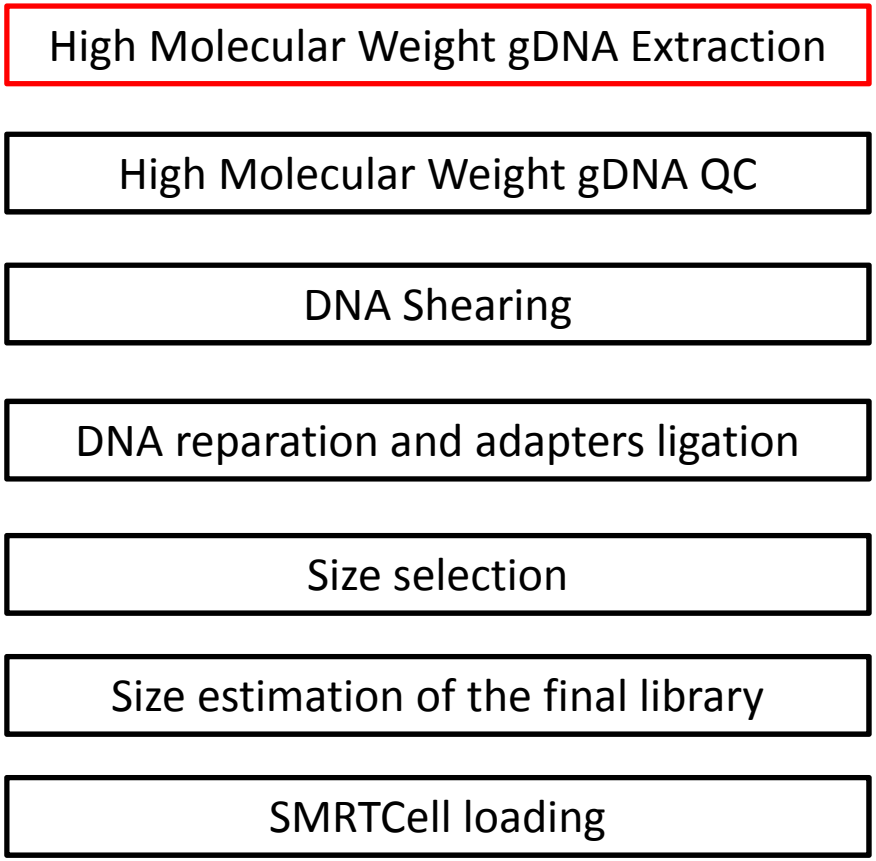
# Large insert library preparation

User Group Meeting

Barcelona 9-10th November

Baptiste Mayjonade - INRA Toulouse (France)

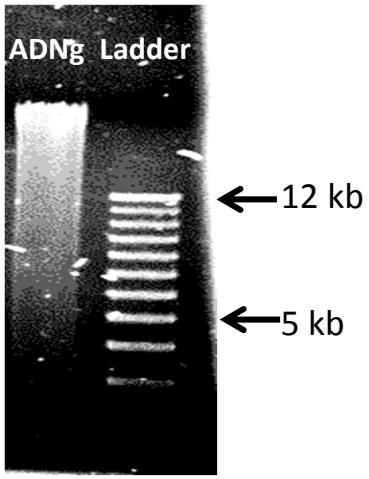
# PacBio large insert library prep workflow



# High Molecular Weight DNA Extraction

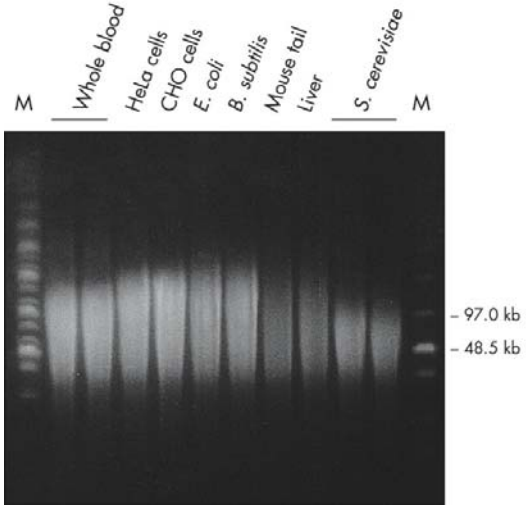
## Spin column

QIAGEN DNesasy Plant Kit



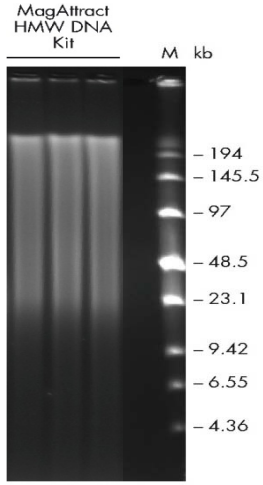

## Flow gravity column

QIAGEN GENOMIC TIPS

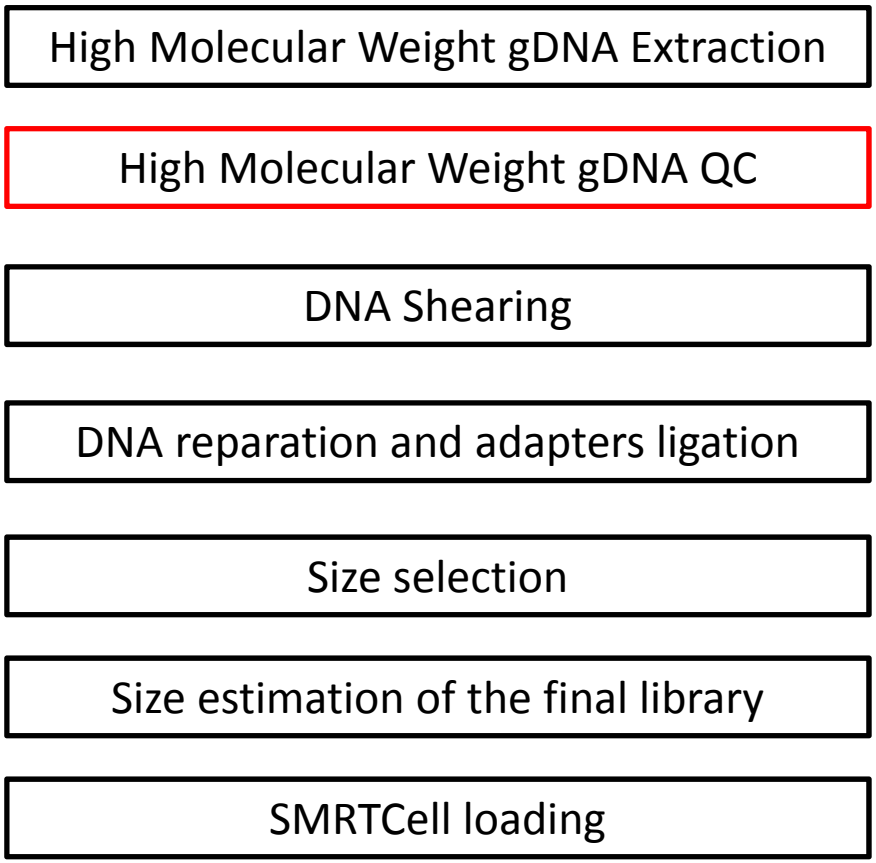



## Magnetic beads

QIAGEN MagAttract HMW



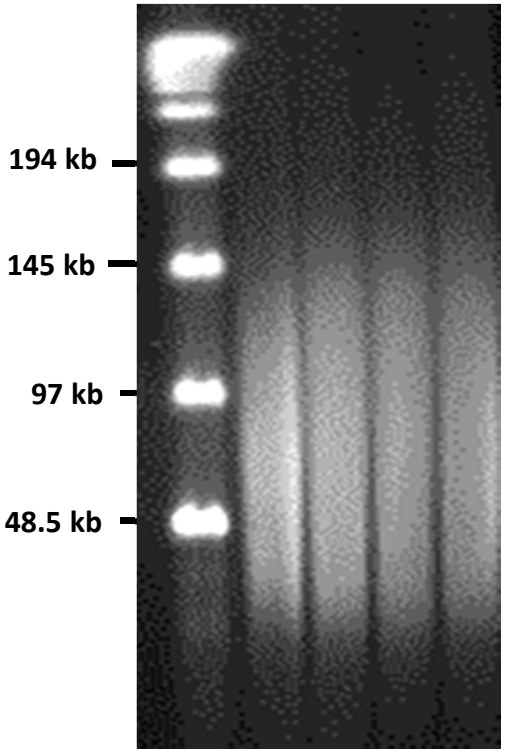
# PacBio large insert library prep workflow



# High Molecular Weight DNA QC

Tools to check DNA integrity

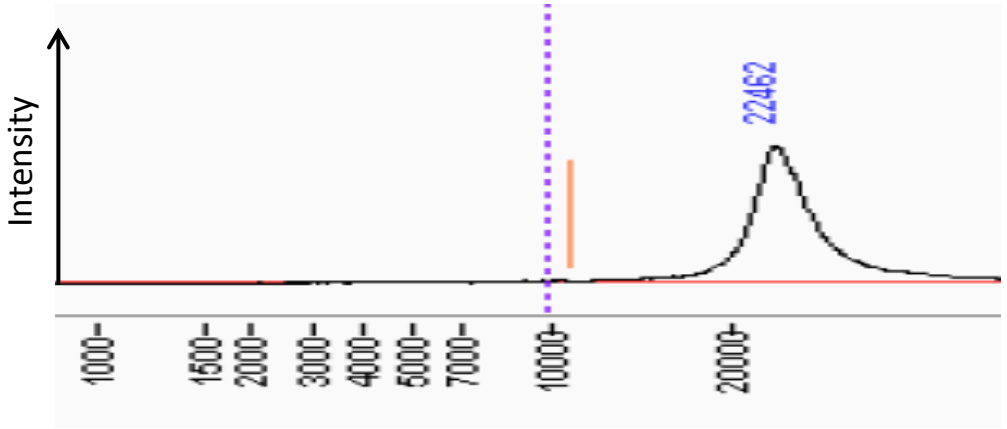
## Pulsed-field electrophoresis




Sunflower DNA on CHEF MAPPER
(Pippin Pulse…)

**Very good DNA size estimation**
**100-200 ng of DNA needed**
**Analysis time ≈ 16h**

## Capillary electrophoresis



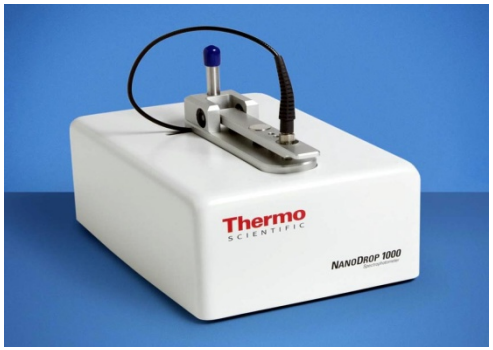

Fragment Analyser
(Tape Station…)

**10 ng of DNA needed**
**Analysis time ≈ 1h**
**False DNA size estimation (good up to 20 kb)**

# High Molecular Weight DNA QC

Tools to check DNA purity

## Spectrophotometric based (Nanodrop…)




$A_{260}$ : DNA concentration

$A_{260/280}$ (>1.8) : protein contamination

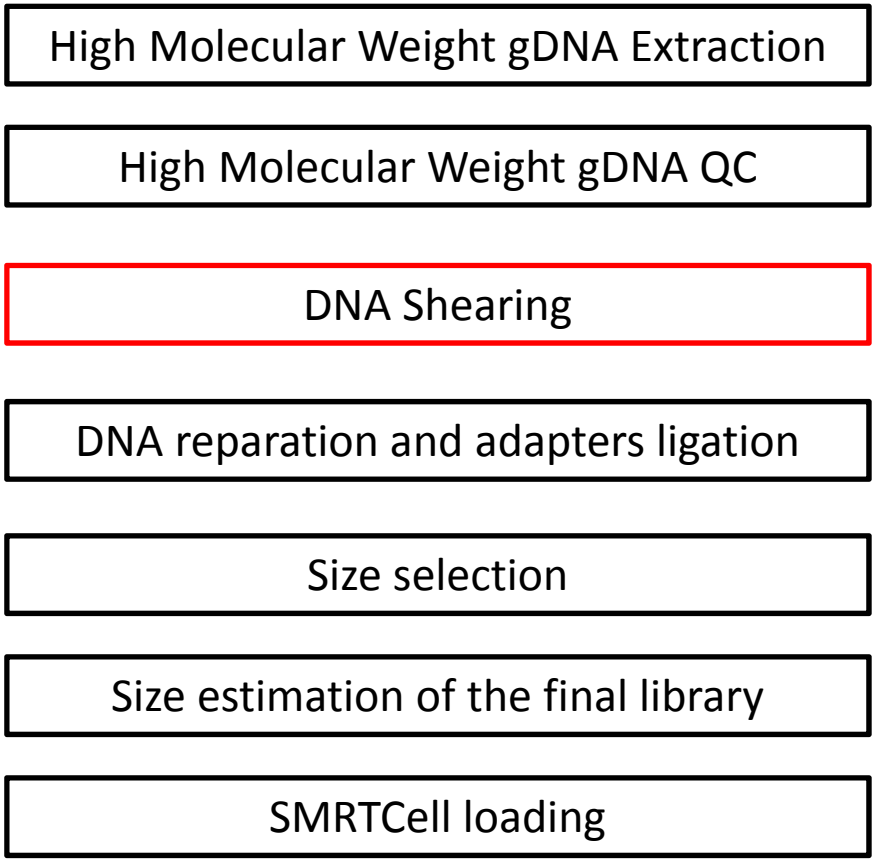
$A_{260/230}$ (>2) : salt, solvent, polyphenol contamination

**Spectrophotometric DNA concentration ≈ intercalant DNA concentration**

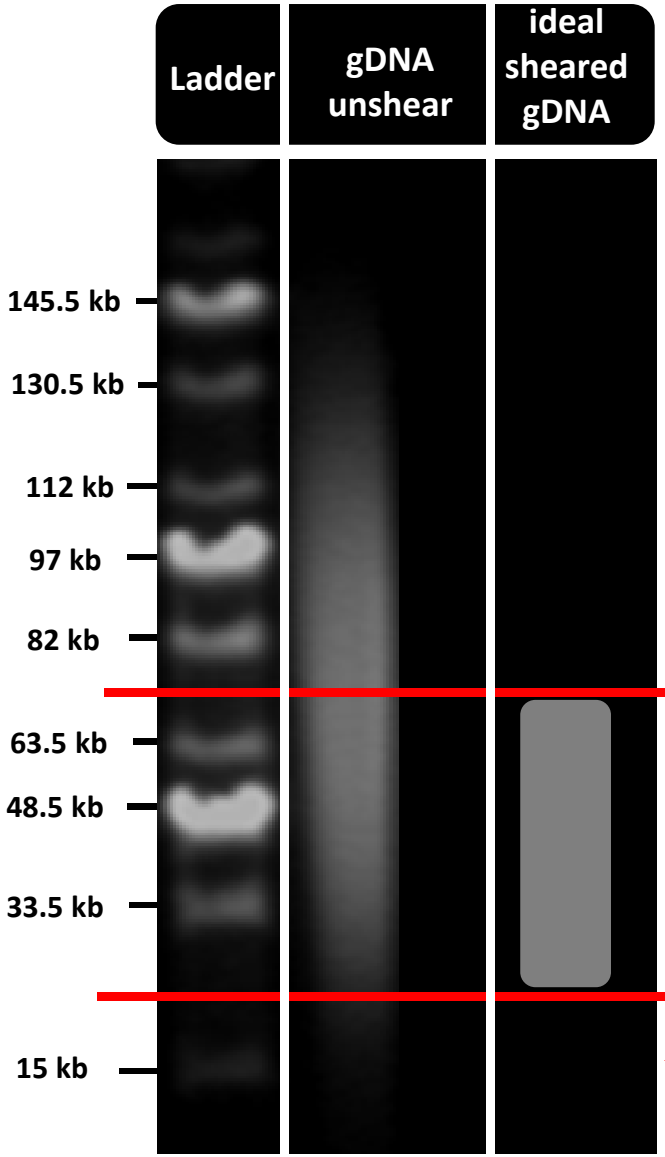
## Intercalant based (Qubit, Picogreen…)


Fluorescence dsDNA specific = DNA concentration

# PacBio large insert library prep workflow



# DNA Shearing




Too large DNA molecules :
- Difficult to load in ZMW
- Won't be fully sequenced (71kb max)
- Better if sheared into shorten molecules (20-71kb) = more molecules = more SMRTCells/library

**71 kb = Max read length** (6 hour movie and polymerase rate=3.3bp/s)

IDEAL DNA MOLECULES SIZE RANGE

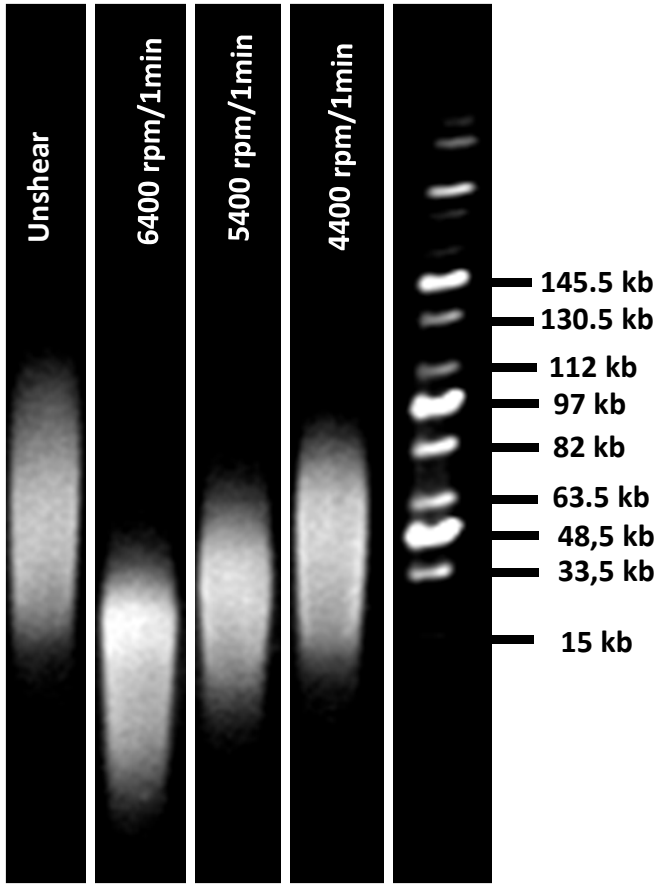
**20 kb = Bluepippin cutoff (10 to 40 kb)**

Too short DNA molecules :
- Will be lost during the size selection step
=> lower the yield during library prep (less SMRTCells/library)

# DNA Shearing

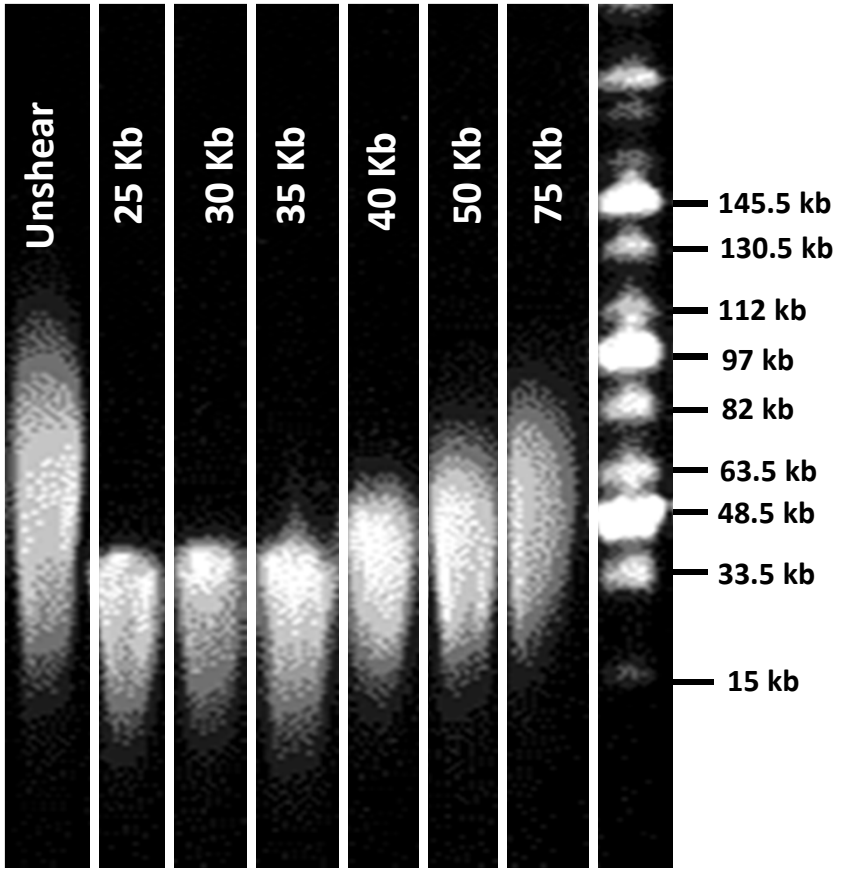
## Devices for DNA shearing

### g-TUBE (1 pass through the g-tube)




**Wide distribution**
**30€ per sample**
**Only centrifuge needed**

### Megaruptor (15 passes through hydropore)

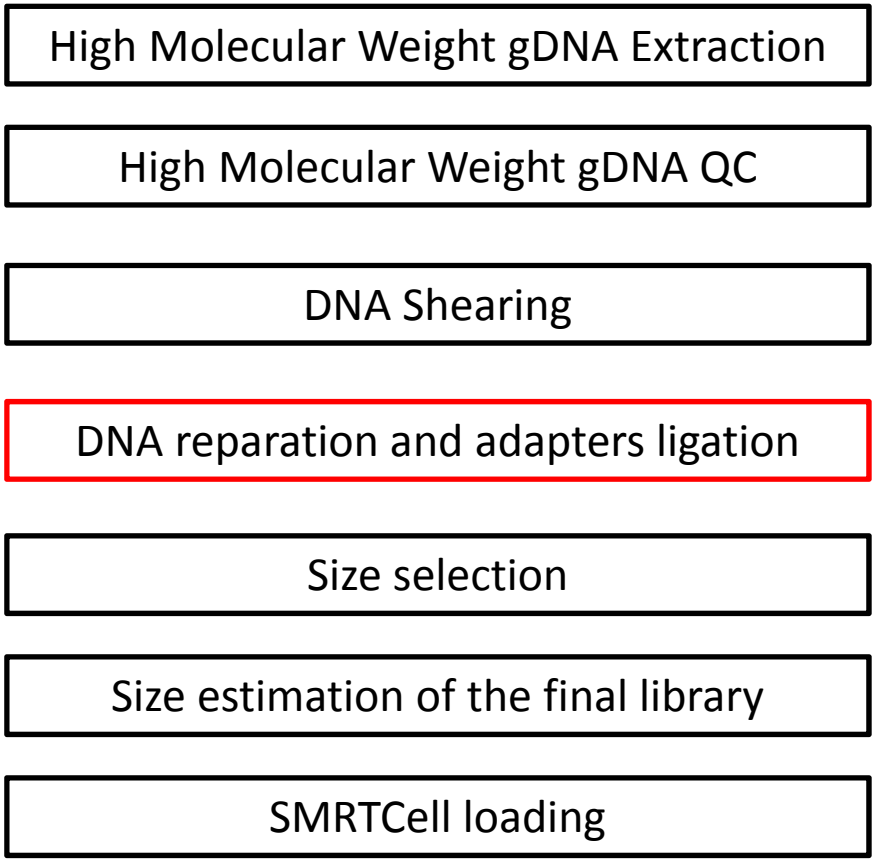



**Tight distribution**
**10€ per sample**
**Megaruptor = 18k€**

### NEEDLE SHEARING

?

# PacBio large insert library prep workflow



## DNA reparation and adapters ligation

AMPurePB purification

### Rotator

More gentle

➡ **Lower recovery?**

➡ **No DNA shearing?**

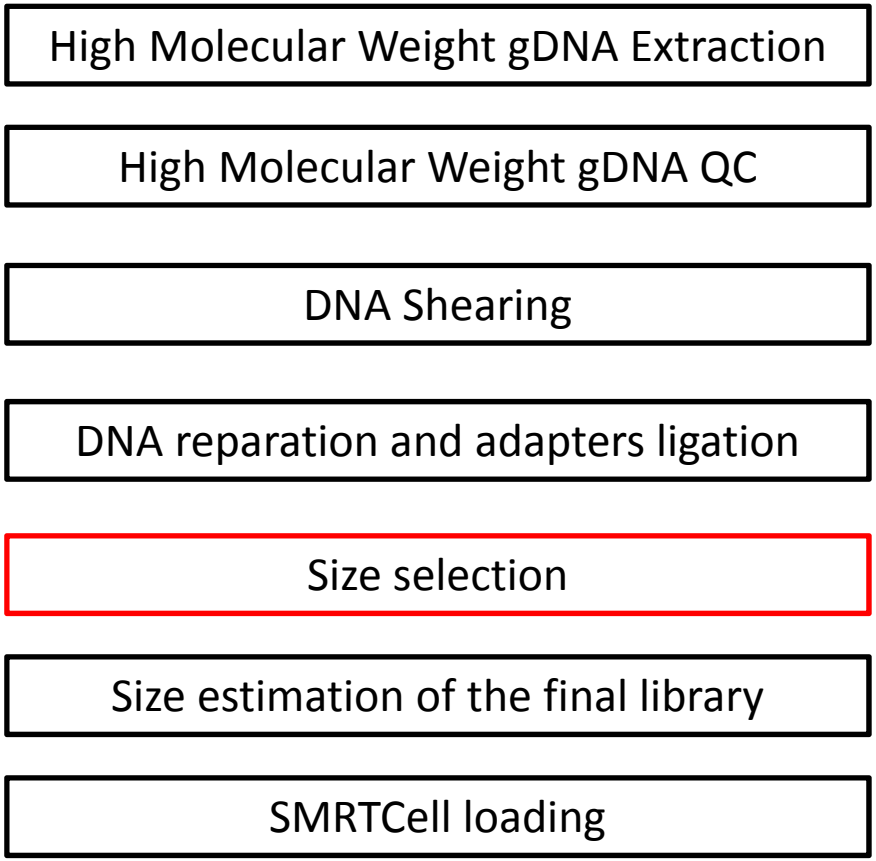
### Vortex


Stronger

➡ **Better recovery?**

➡ **More sheared DNA?**

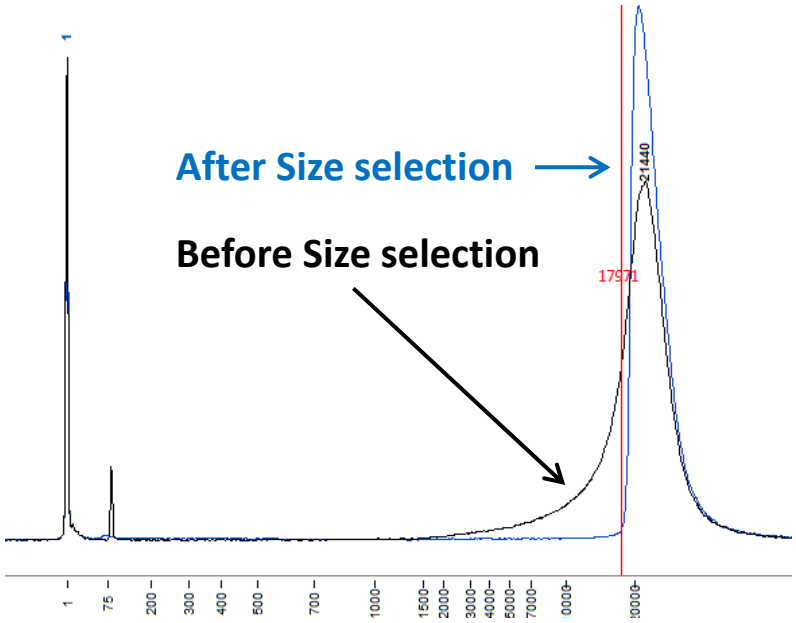
# PacBio large insert library prep workflow



# Size selection

## BluePippin size selection

Able to remove DNA fragment up to 40kb
BluePippin = 15k€




18kb size selected library (Fragment Analyser)

| Size selection cutoff | % Recovery (bluepippin input ≈ 4µg) | Migration time |
|---|---|---|
| 10kb (High pass 6-10kb) | 55-65% | 2h |
| 15kb (High pass 15-20kb) | 40-48% | 3-4h |
| 18kb (High pass 15-20kb) | 30-35% | 4-5h |
| 20kb (High pass 15-20kb) | 18-20% | 4-5h |
| 30-40kb High pass 30-40kb | ???? | ???? |

Alternative : AMPurePB => cost effective and high recovery but removes only fragment below 3-4 kb

## Size selection

BluePippin size selection (g-TUBE vs Megaruptor)

| Shearing Device | Shearing settings | BluePippin cut-off | BluePippin Input | Recovery |
|---|---|---|---|---|
| **g-TUBE** | 4000 rpm | 18 kb | 4000 ng | 1418 ng **(35%)** |
| **Megaruptor** | 40 kb | 18 kb | 4000 ng | 1370 ng **(34%)** |
| **g-TUBE** | 3600 rpm | 20 kb | 3870 ng | 673 ng **(17%)** |
| **Megaruptor** | 50 kb | 20 kb | 3870 ng | 723 ng **(19%)** |

**Same recovery with Megaruptor or g-TUBE**

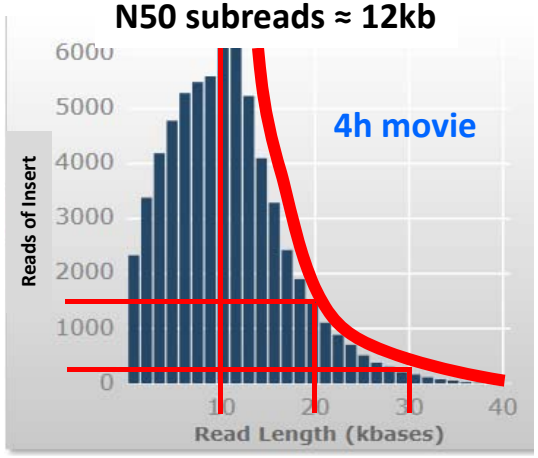
## Size selection

Impact of the size selection on subreads length distribution

**12kb cut off**

(loading 0.15nM= 868Mb)

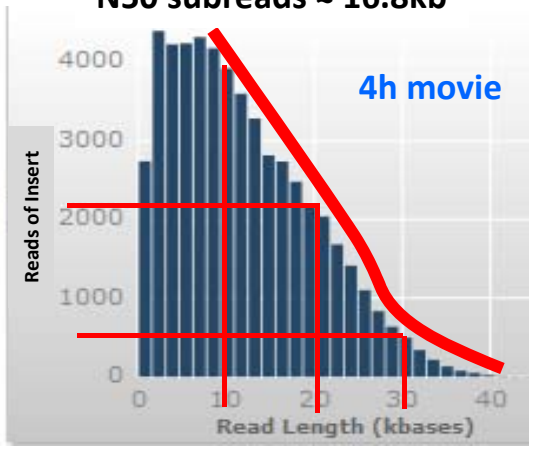
**N50 subreads ≈ 12kb**




**18kb cut off**

(loading 0.30nM= 710Mb)

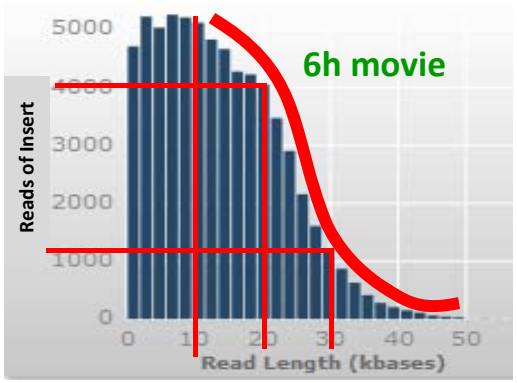
**N50 subreads ≈ 16.8kb**




**18kb cut off**

(loading 0.36nM= 1036Mb)

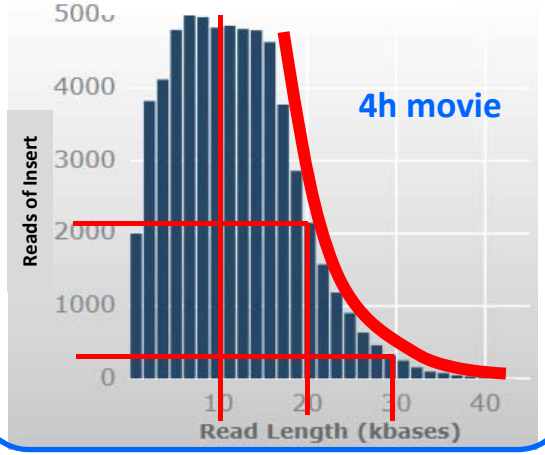
**N50 subreads ≈ 19kb**




**15kb cut off**

(loading 0.20nM= 877Mb)

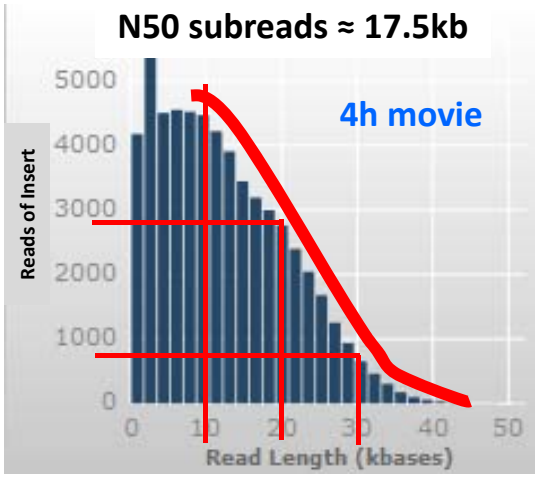
**N50 subreads ≈ 15kb**




**20kb cut off**

(loading 0.45nM= 800Mb)

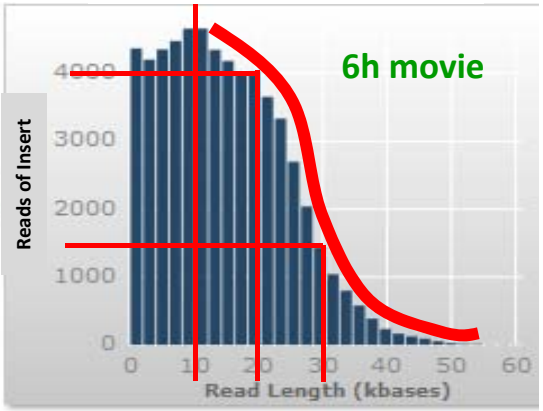
**N50 subreads ≈ 17.5kb**



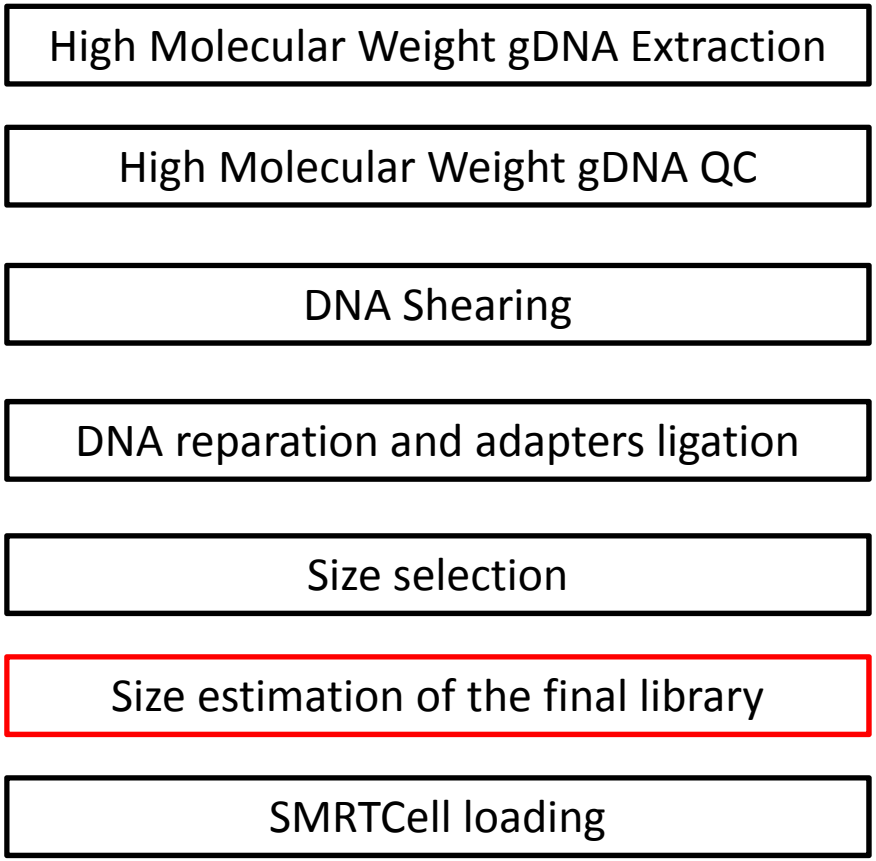

**20kb cut off**

(loading 0.45nM= 1000Mb)

**N50 subreads ≈ 20.5kb**



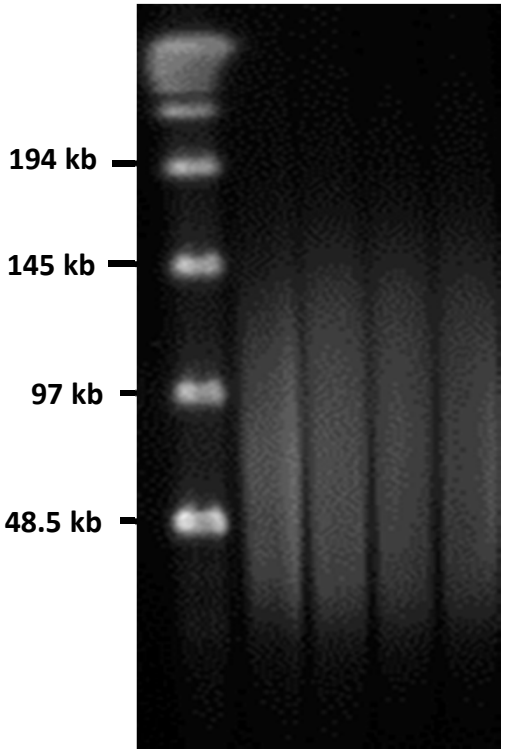
# PacBio large insert library prep workflow



## Size estimation of the final library

Tools to evaluate the final size of the library

### Pulsed-field electrophoresis




CHEF MAPPER

(Pippin Pulse…)

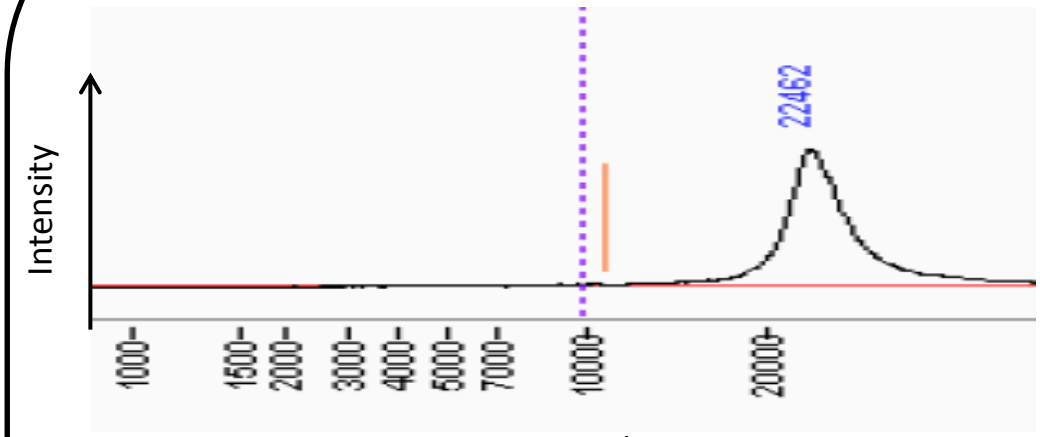
**Very good DNA size estimation**

**100-200 ng of DNA needed**

**→ (loss of library)**

**Analysis time ≈ 16h**

### Capillary electrophoresis




Fragment Analyser

(Tape Station…)

**10 ng of DNA needed**

**Analysis time ≈ 1h**

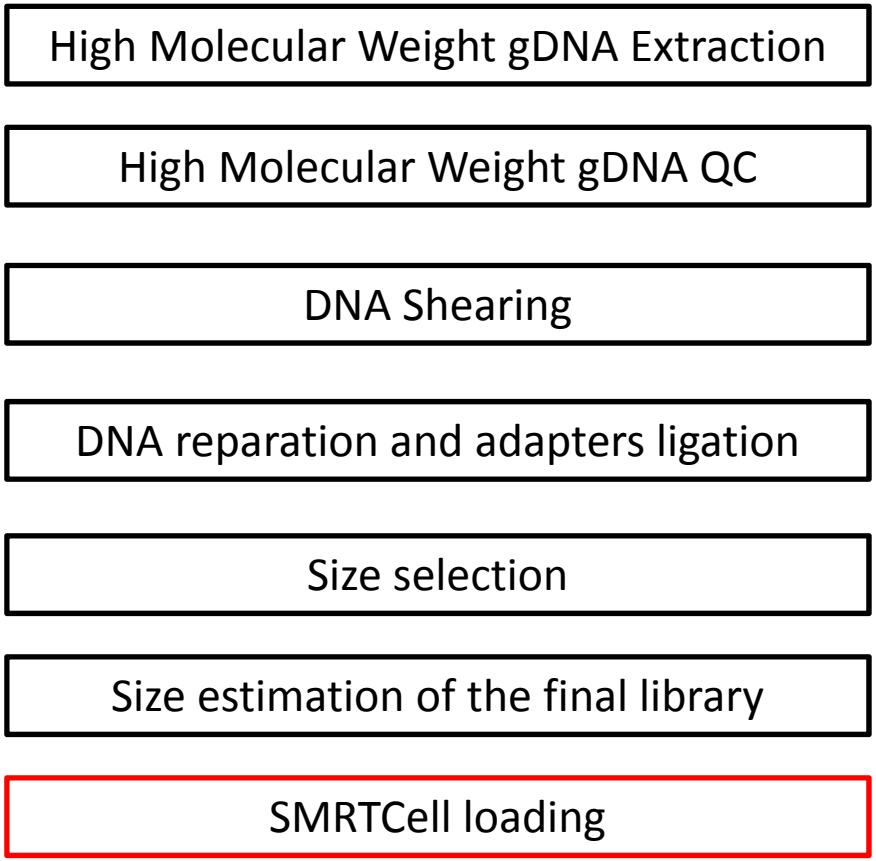
**False DNA size estimation (good up to 20 kb)**

**Underestimation of the size of the DNA molecules**

**Bad conversion of mass to molar concentration**
**(larger molecules => less DNA molecules)**

**Underloading with binding calculator**

# PacBio large insert library prep workflow



# SMRTCell loading

## Impact of the incubation time for polymerase binding

| Binding polymerase | Rotator (MagBeads) | Loading | Movie Length (mins) | Total Bases (MB) | Polymerase Reads | | Reads Of Insert | | Productivity | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Length | Quality | Length | Quality | Empty (P0) | Productive (P1) | Other (P2) |
| **30 minutes at 30°C** | **1h** | **0,26nM** | 360 | 872.70 | 15504 | 0.83 | 13505 | 0.84 | 51% | 37% | 12% |
| | | | 360 | 833.08 | 15261 | 0.84 | 13277 | 0.84 | 53% | 36% | 11% |
| | | | 360 | 802.99 | 15208 | 0.84 | 13100 | 0.84 | 55% | 35% | 10% |
| | | | 360 | 846.70 | 15348 | 0.84 | 13193 | 0.84 | 47% | 37% | 16% |
| | | | | | | | | **MEAN** | **52%** | **36%** | **12%** |
| **4 hours at 30°C** | **1h** | **0,26nM** | 360 | 1379.21 | 16738 | 0.85 | 14126 | 0.85 | 35% | 55% | 11% |
| | | | 360 | 1365.96 | 16472 | 0.85 | 14069 | 0.85 | 35% | 55% | 10% |
| | | | 360 | 1627.99 | 17235 | 0.84 | 14689 | 0.85 | 25% | 63% | 12% |
| | | | 360 | 1277.37 | 15978 | 0.84 | 13894 | 0.84 | 35% | 53% | 12% |
| | | | | | | | | **MEAN** | **33%** | **57%** | **11%** |

Note : All theses data are produced with the same library and SMRTCell lot

# SMRTCell loading

## Impact of the SMRTCell lot

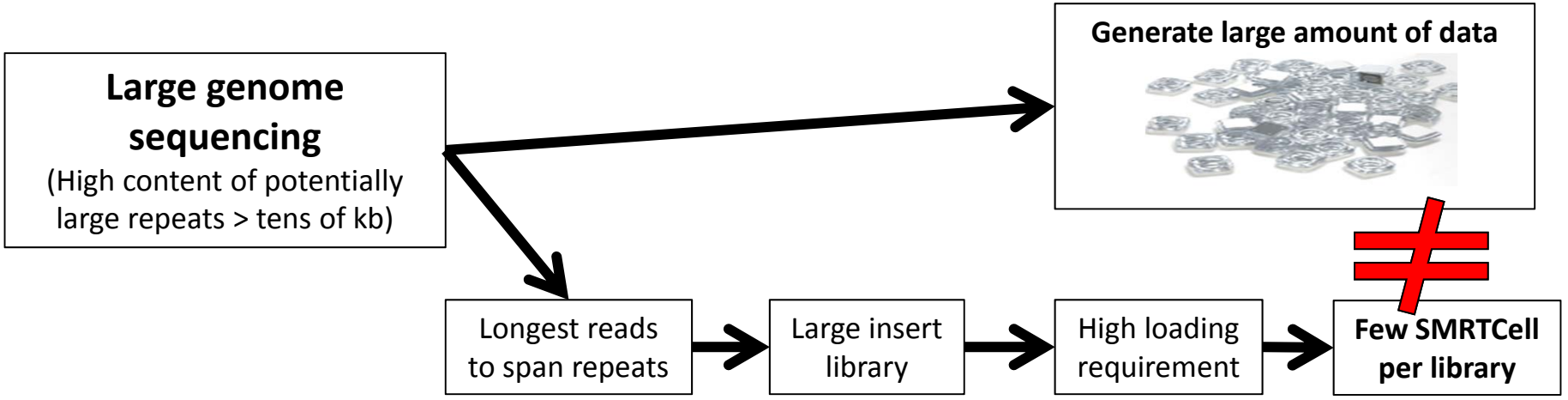
| | Sample Name | Movie Length (mins) | Total Bases (MB) | Polymerase Reads | | Reads Of Insert | | Productivity | | | SNR | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Length | Quality | Length | Quality | P0 | P1 | P2 | T | A |
| SMRTCell Lot1 | Library g-tube 5000 rpm BluePippin 15Kb<br>Loading = 0.25nM | 240 | 993.69 | 13940 | 0.83 | 11624 | 0.84 | (36%) | (47%) | (16%) | 8.3 | 13.3 |
| | | 240 | 874.48 | 14511 | 0.84 | 11914 | 0.84 | (49%) | (40%) | (11%) | 8.2 | 13.7 |
| | | 240 | 1201.26 | 14218 | 0.84 | 11904 | 0.84 | (30%) | (56%) | (14%) | 8.7 | 13.2 |
| | | 240 | 1085.38 | 14208 | 0.84 | 11901 | 0.84 | (36%) | (51%) | (14%) | 7.4 | 14.2 |
| SMRTCell Lot2 | Library g-tube 5000 rpm BluePippin 15Kb<br>Loading = 0.25nMb | 240 | 681.44 | 14943 | 0.84 | 12101 | 0.84 | (27%) | (30%) | (42%) | 8.0 | 8.2 |
| | | 240 | 625.57 | 15157 | 0.84 | 12174 | 0.84 | (26%) | (27%) | (47%) | 7.8 | 7.0 |
| | | 240 | 606.16 | 15364 | 0.84 | 12319 | 0.84 | (26%) | (26%) | (48%) | 7.4 | 7.6 |
| | | 240 | 514.58 | 15336 | 0.83 | 12437 | 0.84 | (23%) | (22%) | (55%) | 6.7 | 7.2 |

**Titration needed for each SMRTCell lot?**

# SMRTCell loading

## Impact of the size of the library

| Size selection cut-off | Nb of SMRTCells per library (Library input = 5µg) (Loading for ≈1gb/SMRTCell) | Amount of data per library |
|---|---|---|
| 12kb | 40 SMRTCells | ≈40 Gb |
| 15kb | 30 SMRTCells | ≈30 Gb |
| 18kb | 15 SMRTCells | ≈15 Gb |
| 20kb | 8 SMRTCells | ≈8 Gb |
| 30kb | ??? | ??? |
| 40kb | ??? | ??? |



# Outlook

How to improve the library preparation?

- For high size selection cut off (30 to 40kb) : add a size selection step after shearing?
  - All DNA extraction and DNA shearing protocols generate DNA molecules between 10 and 40Kb
  - Useless to put these DNA molecules into expensive repair steps (can represent a significant part of the library and will be lost in the final size selection step)

- DNA damage repair after size selection (BluePippin)? (repair damaging DNA during electrophoresis)
  - **Requirement of AMPure PB purification => library losses**
  - **Less damaged DNA molecules => Increase polymerase read length?**

# Outlook

How to increase subreads length

| Top 10 of our longest subreads |
| --- |
| 80974 bp |
| 79860 bp |
| 79834 bp |
| 78105 bp |
| 77481 bp |
| 76881 bp |
| 76558 bp |
| 76355 bp |
| 75569 bp |
| 75559 bp |

- Max read length with **6 hour movie** : 6 h x 60 min x 60 sec x 3.3 bp/s ≈ **71kb**

- Max read length with **8 hour movie** : 8 h x 60 min x 60 sec x 3.3 bp/s ≈**95kb !**
- Max read length with **10 hour movie** : 10 x 60 min x 60 sec x 3.3 bp/s ≈ **120kb !!!**

- Better way = faster polymerase (P7?)

# Thanks

LIPM:
Jérôme Gouzy
Nicolas Langlade
Munos Stéphane
Chris Grassa
Sébastien Carrere
Erika Sallet
Ludovic Legrand
Marie-Claude Boniface
Nicolas Pouilly

Get-PlaGe:
Cécile Donadieu
Gérald Salin
Denis Milan

CNRGV:
Hélène Bergès
William Marande