SARS-CoV-2 NGS Training

NGS Academy for the Africa Pathogen Genomics Initiative

Module 4 - Illumina workflow

SARS-CoV-2 NGS Training

Module 4: Illumina workflow

- Session 1: Concepts, library prep, starting a sequencing run
- Session 2: Library prep QC and sequencing run QC
- Session 3: IDSeq/Data Processing

CHAN ZUCKERBERG BIOHUB

Cristina Tato,
Director
Rapid Response (RR) group

Manu Vanaerschot,
Scientist
RR group

Vida Ahyong,
Scientist
RR group

CHAN ZUCKERBERG INITIATIVE

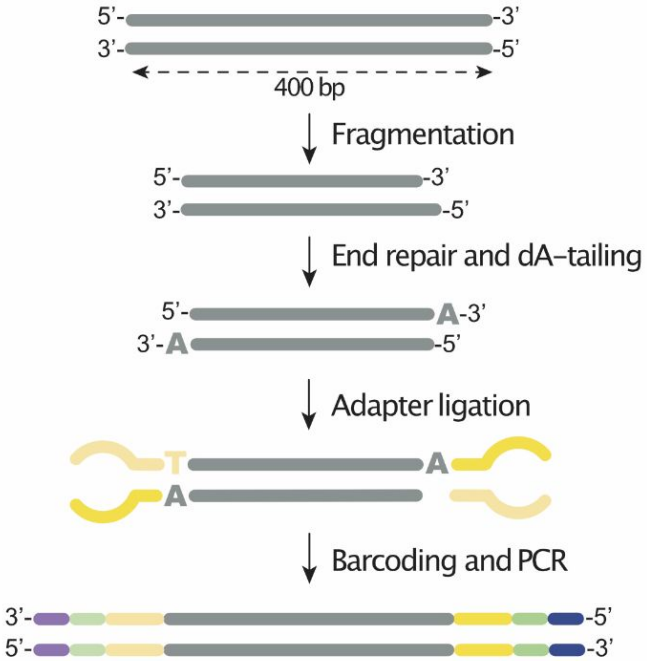
Katrina Kalantar,
Computational Biologist

Liz Fahsbender,
IDseq Application Scientist

Library prep

Ligation protocol

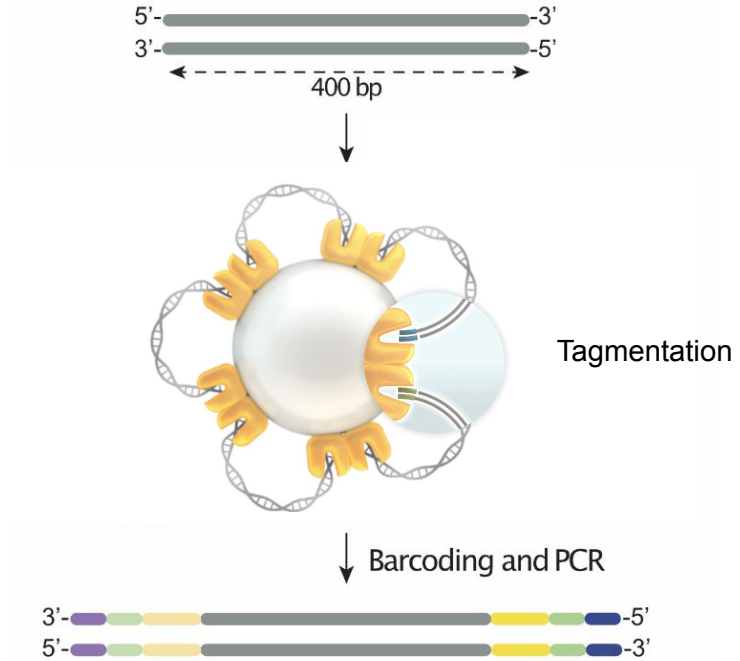
Illumina TruSeq DNA / NEBNext Ultra II DNA



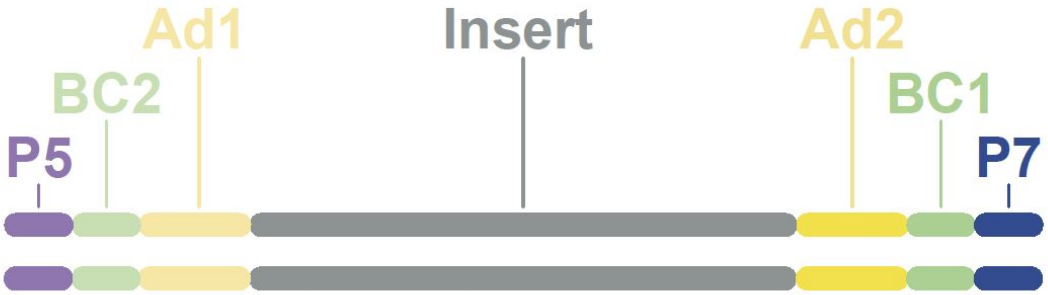
OR

Tagmentation protocol

Illumina DNA prep (= Nextera Flex)



Library prep - final library



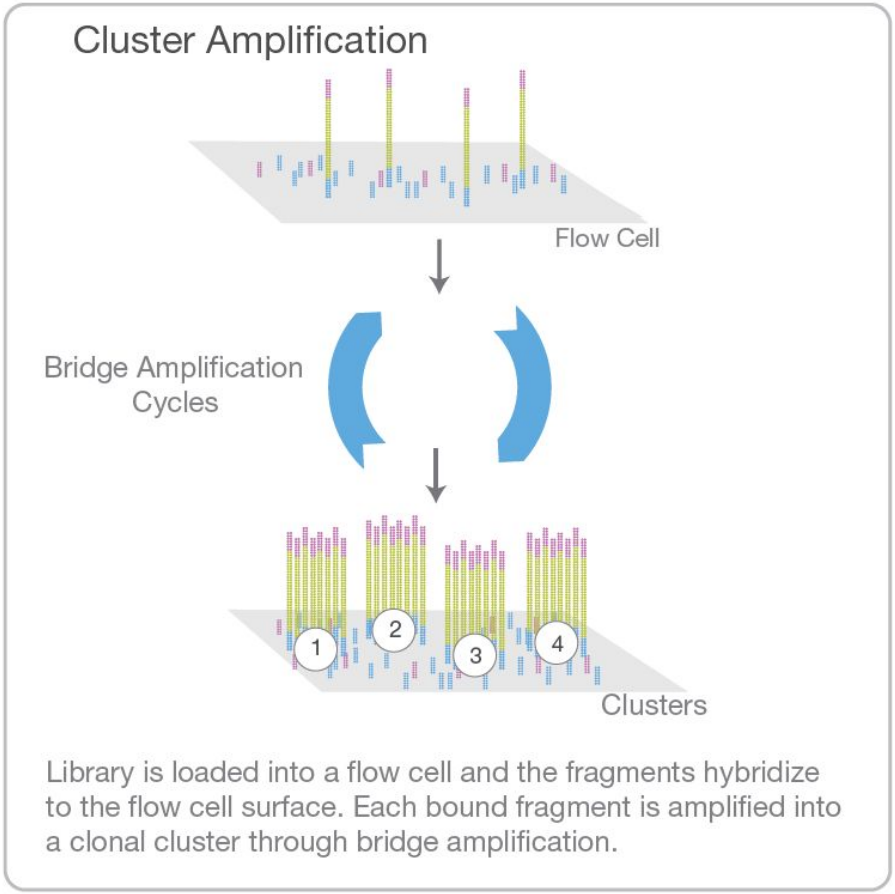
Insert = DNA fragment to be sequenced

Ad1/Ad2 = Adapters that allow initiation of sequencing

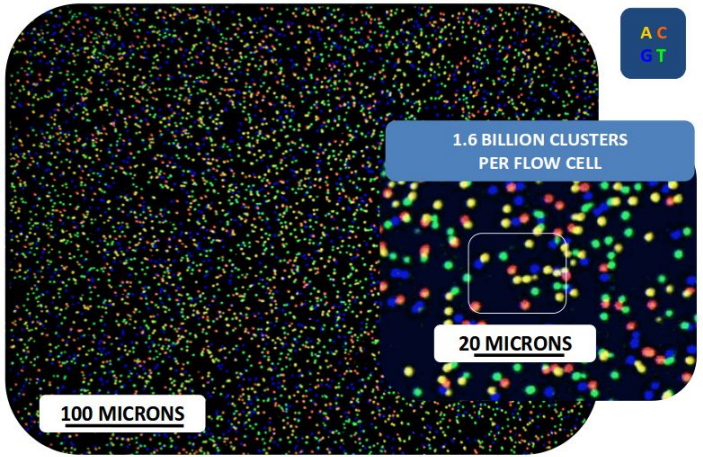
BC1/BC2 = Barcodes specific for each sample (one for each strand)

P5/P7 = Flow cell binding sites

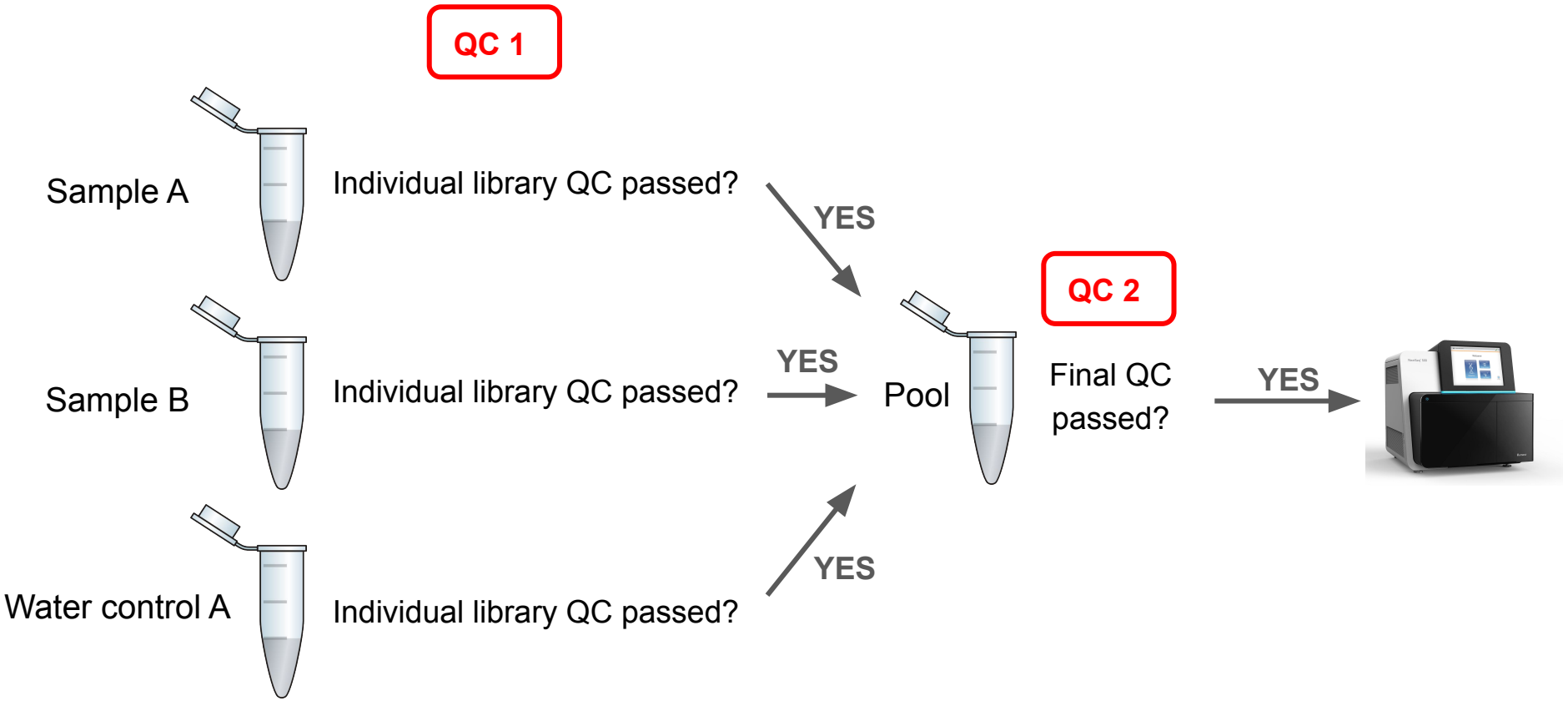
Cluster amplification is essential for Illumina sequencing



~ loading concentration
~ fragment length



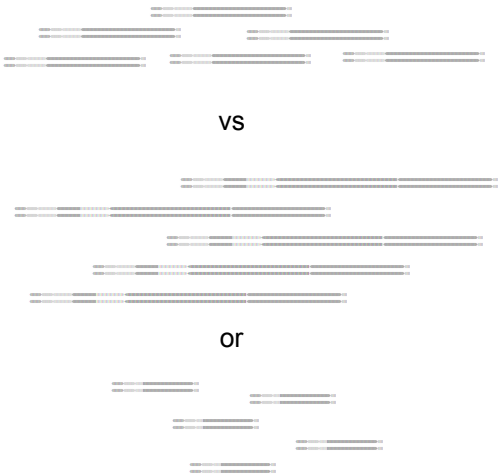
Library prep QC



Library prep QC

Concentration

Length



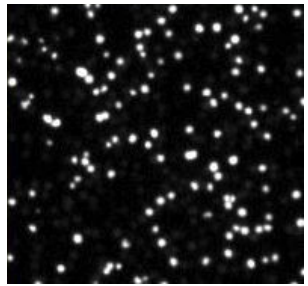
Remove adapter dimers

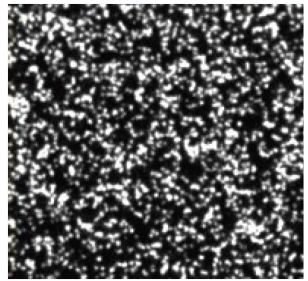
Library prep QC - Concentration

Why is an accurate concentration important?

- Individual samples: to pool equal amounts of different samples together
  - Aim for an even amount of sequencing data per sample

- Final library: sequencer requires loading with a very precise concentration
  - If concentration is overestimated (actual conc. is lower than what you think)
    - Underloading and underclustering of sequencing run
    - Sequencing run will yield less sequencing data than usual
  - If concentration is underestimated (actual conc. is higher than what you think)
    - Overloading and overclustering of sequencing run
    - Sequencing run will suffer from higher error rates and reduced data output





Library prep QC - Concentration

DO NOT USE to quantify libraries

Spectrophotometry
(e.g. Nanodrop)

- DNA absorbs light most strongly at 260 nm
- Level of absorbance ~ concentration of DNA

BUT other molecules (RNA, proteins) & contaminants also absorb light at 260 nm, causing up to 10-fold **overestimation** of the DNA concentration!

Fluorometry
(e.g. Qubit, qPCR)

- Detects fluorescent dyes that only bind DNA
- Most reliable method for NGS

Electrophoresis
(e.g. Tapestation, Bioanalyzer)

- Compare intensity of DNA bands to standards
- Especially useful to determine concentration in certain fragment length range

Library prep QC - Concentration

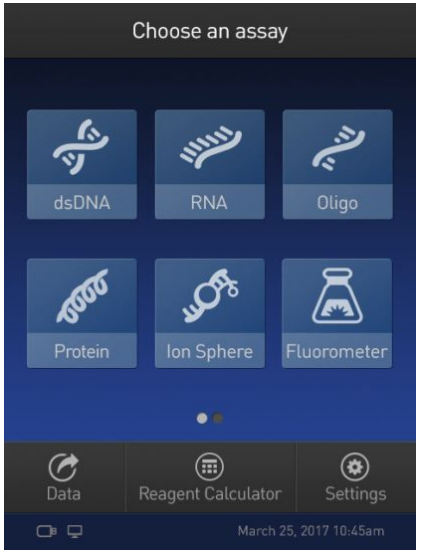
Qubit dsDNA High Sensitivity (HS) Assay kit

- dsDNA HS Reagent (200x concentrate)
- dsDNA HS Buffer
- dsDNA HS Standard #1
- dsDNA HS Standard #2

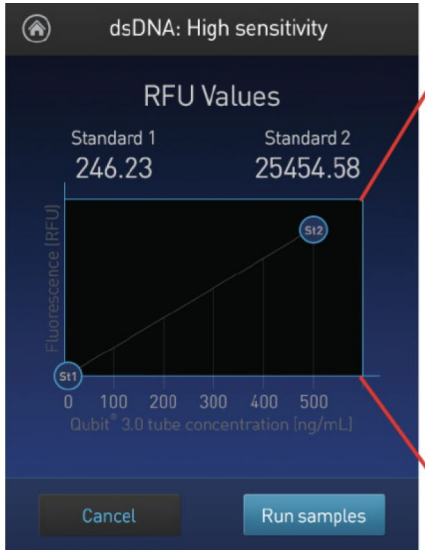
1. Prepare Working Solution:
   - Diluting HS reagent 1:200 in HS buffer
2. Dilute in Working Solution:
   - Standards: 10 µl standard + 190 µl Working Solution
   - Samples: 1 µl sample + 199 µl Working Solution
3. Vortex 2-3 seconds before running standards/samples

Library prep QC - Concentration

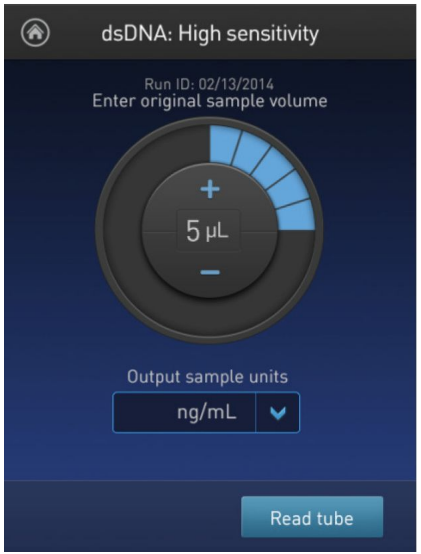
Qubit dsDNA High Sensitivity (HS) Assay kit



Select dsDNA HS workflow



Run the 2 standards



Enter sample volume added to tube

Run sample
Top value = original conc.
Bottom value = diluted conc.

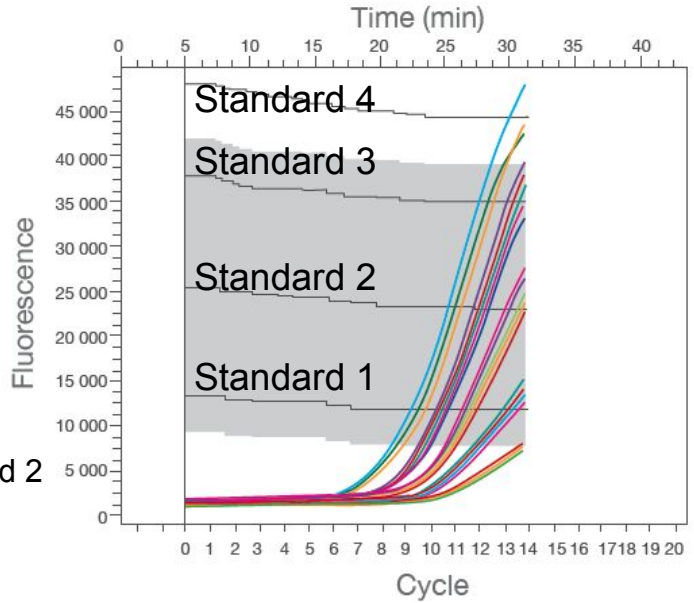
Library prep QC - Concentration

- **What if the concentration is too low?**
  - In future runs, consider adding more cycles to the indexing PCR step
  - Amplify library using P5/P7 primers and qPCR-based library amplification kit
    - e.g. Roche KK2702

1. Set up qPCR reaction: library + mastermix + P5/P7 primers
2. Treat standards as separate samples (do not add mastermix)
3. Start run

Temperature	Time	Cycles
98°C	45 sec	1
98°C	15 sec	25
63°C	30 sec	
72°C	1m 45 sec	
Plate read		
72°C	20 sec	

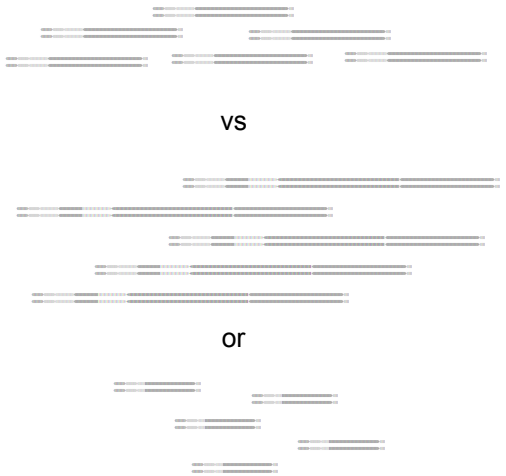
4. Pause cycler during the 20s at 72°C step when a sample crossed the Standard 2 line, remove the sample and continue amplification for other samples
5. After amplification, perform bead cleanup on amplified libraries and QC again



Library prep QC

Concentration

Length



Remove adapter dimers

Library prep QC - Length

Why is it important to assess fragment length in a library?

- Long fragments (>1.2 kb) do not cluster on Illumina flow cells and are therefore not sequenced
- Short fragments are preferentially amplified (better clustering) and will thus be preferentially sequenced
- Fragment length distribution impacts calculations to go from ng/µl (Qubit) to nM for loading concentrations

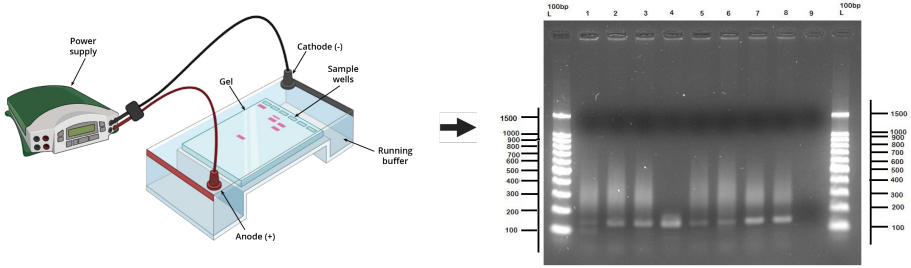
Qubit

$$\text{Concentration in nM} = \frac{\text{concentration in ng/µl}}{\text{660g/mol x average library size in bp}} \times 10^6$$

Capillary or gel electrophoresis

Library prep QC - Length

Gel electrophoresis

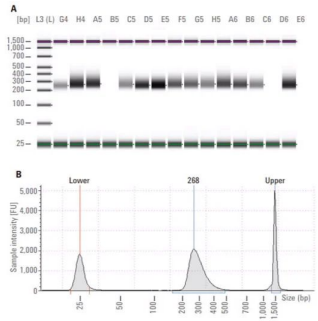


1. PCR amplify an aliquot of the library using universal P5/P7 Illumina primers
2. Gel electrophoresis of ladder and samples

- Takes ~ 4 hrs

Capillary electrophoresis

E.g. Tapestation, Bioanalyzer



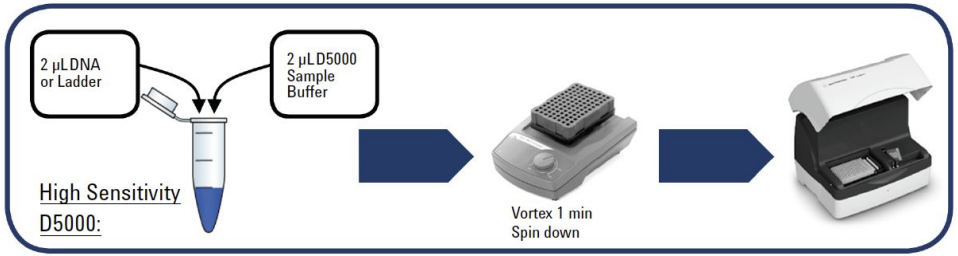
1. Add buffer to aliquot of library
2. Run machine

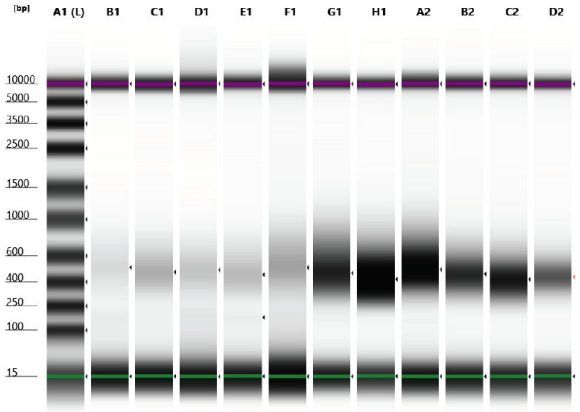
- Tapestation ~ 10 minutes (depending on # samples)
- Bioanalyzer ~ 45 minutes
- No PCR bias
- Can calculate concentration in nM

Library prep QC - Length

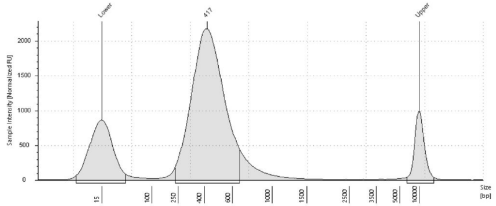
Tapestation

- High Sensitivity D1000 or D5000 kits (1kb or 5kb detection limit)





Peak: H1



Location	H1
Concentration	1470
Description	
Observations	

Size [bp]	Calibrated Conc. [pg/µl]	Assigned Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area	Peak Comment	Observations
15	346	-	35500	-		Lower Marker
417	1470	-	5430	100.00		
10000	180	180	27.7	-		Upper Marker

Library prep QC

Concentration

vs

Length

vs

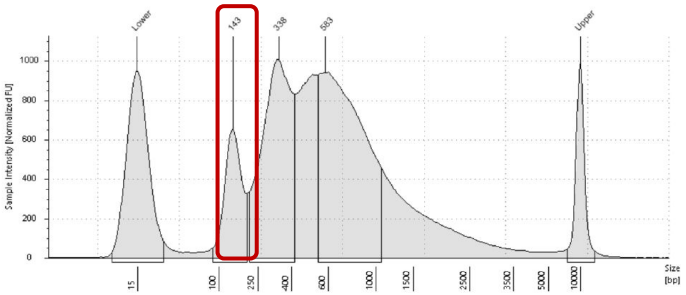
or

Remove adapter dimers

vs

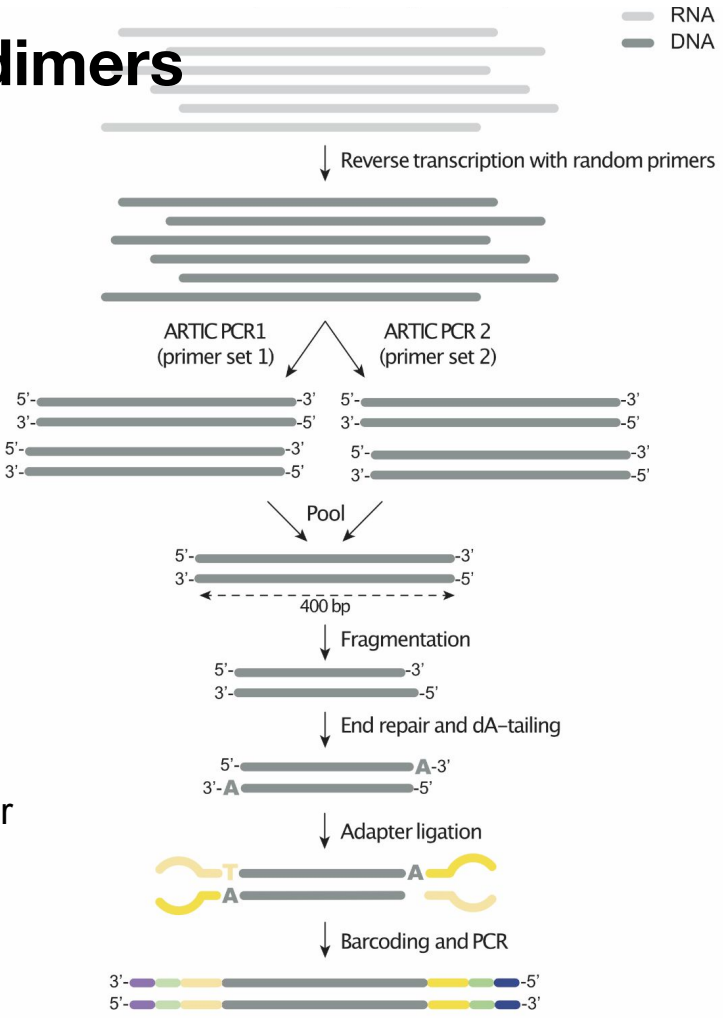
Library prep QC - Clean up adapter dimers

- Adapter dimers are:
  - A byproduct of the adapter ligation step during library prep (TruSeq/NEBNext Ultra II DNA kits)
  - ± 140 bp long at the final QC step (after barcoding PCR)
  - Use capillary electrophoresis (preferred) or PCR + gel electrophoresis to detect adapter dimers

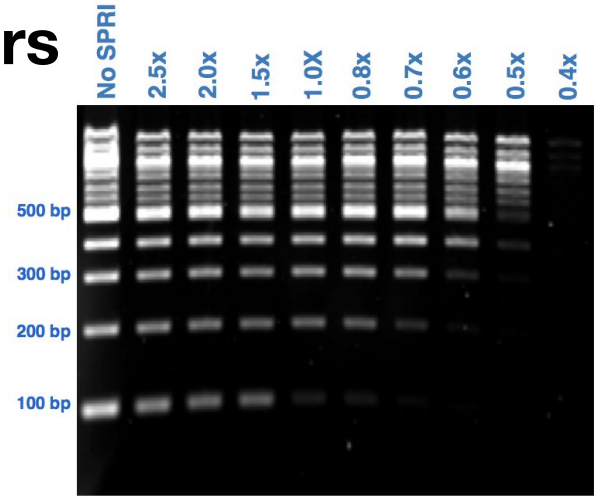


Adapter dimers

Adapter dimer



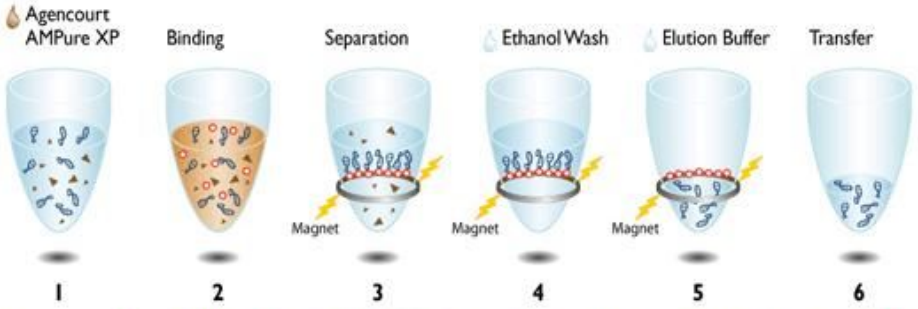
Library prep QC - Clean up adapter dimers



AMPure XP SPRI select beads (paramagnetic)

or homemade SPRI beads (>10x cheaper)

- The longer a fragment, the better it will bind to the beads
- The fewer beads you add, the more short fragments you'll lose
- **Perform size selection using these beads to remove adapter dimers**



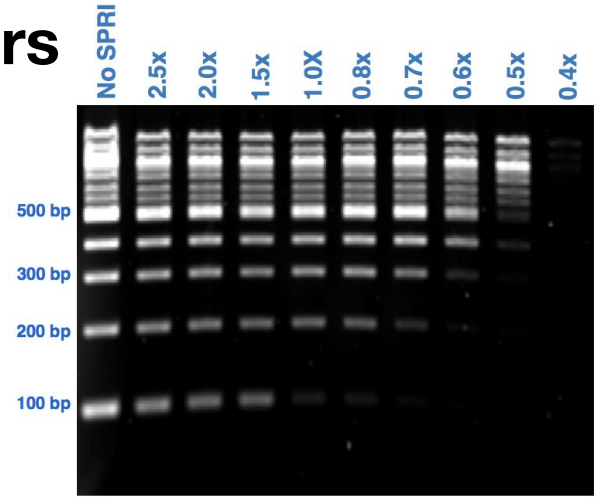
1. PCR reaction 2. Binding of PCR amplicons to magnetic beads 3. Separation of PCR amplicons bound to magnetic beads from contaminants 4. Washing of PCR amplicons with Ethanol 5. Elution of PCR amplicons from the magnetic particles 6. Transfer away from the beads into a new plate

$$\text{SPRI ratio} = \frac{\text{Volume beads}}{\text{Volume reaction}}$$

Example: for a 1.8x SPRI clean-up:

- Add 18 µl of SPRI beads to 10 µl of sample

Library prep QC - Clean up adapter dimers

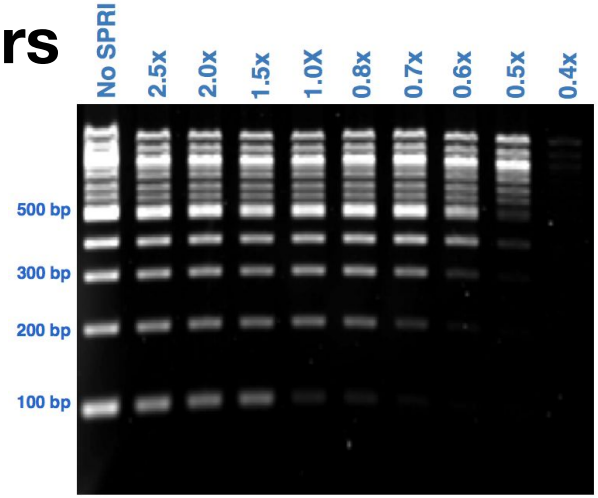


AMPure XP SPRI select beads (paramagnetic)

- Size selection using beads to remove adapter dimers
- Rule of thumb: if you see a clear trace of adapter dimers, performs another round of cleanup

Step	To remove	SPRI ratio
After adapter ligation	Adapter dimers (64 bp) & reagents	0.9x
After P7/P5 PCR	P5-adapter dimer-P7 (143 bp) & reagents	0.8x
QC of library	P5-Adapter dimer-P7 (143 bp)	0.8x

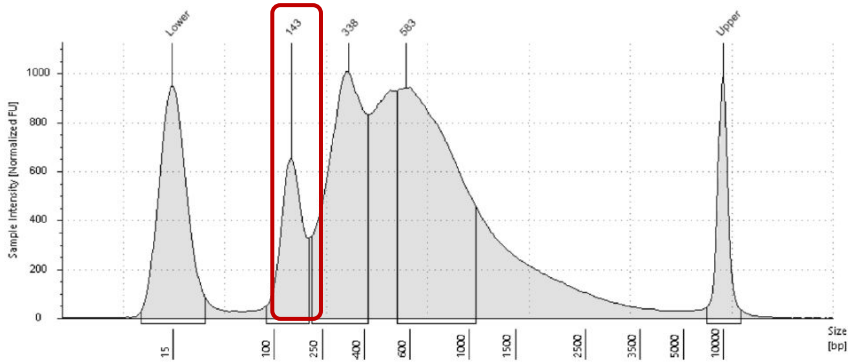
Library prep QC - Clean up adapter dimers



AMPure XP SPRI select beads (paramagnetic)

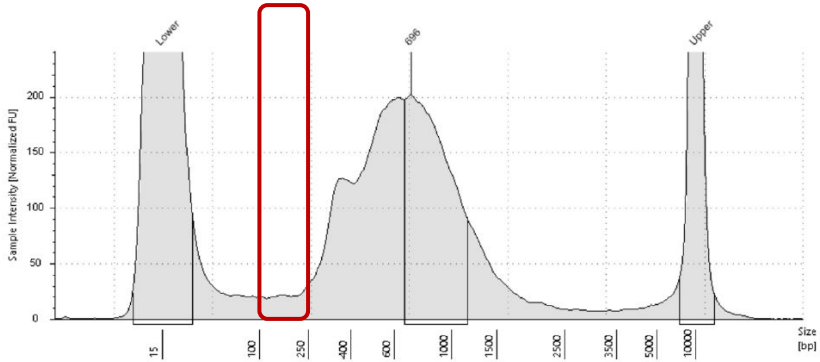
- Size selection using SPRI beads to remove adapter dimers
- Rule of thumb: if you see a clear trace of adapter dimers, performs another round of cleanup

After PCR amplification



3 rounds SPRI (0.85x)

After SPRI bead clean-up



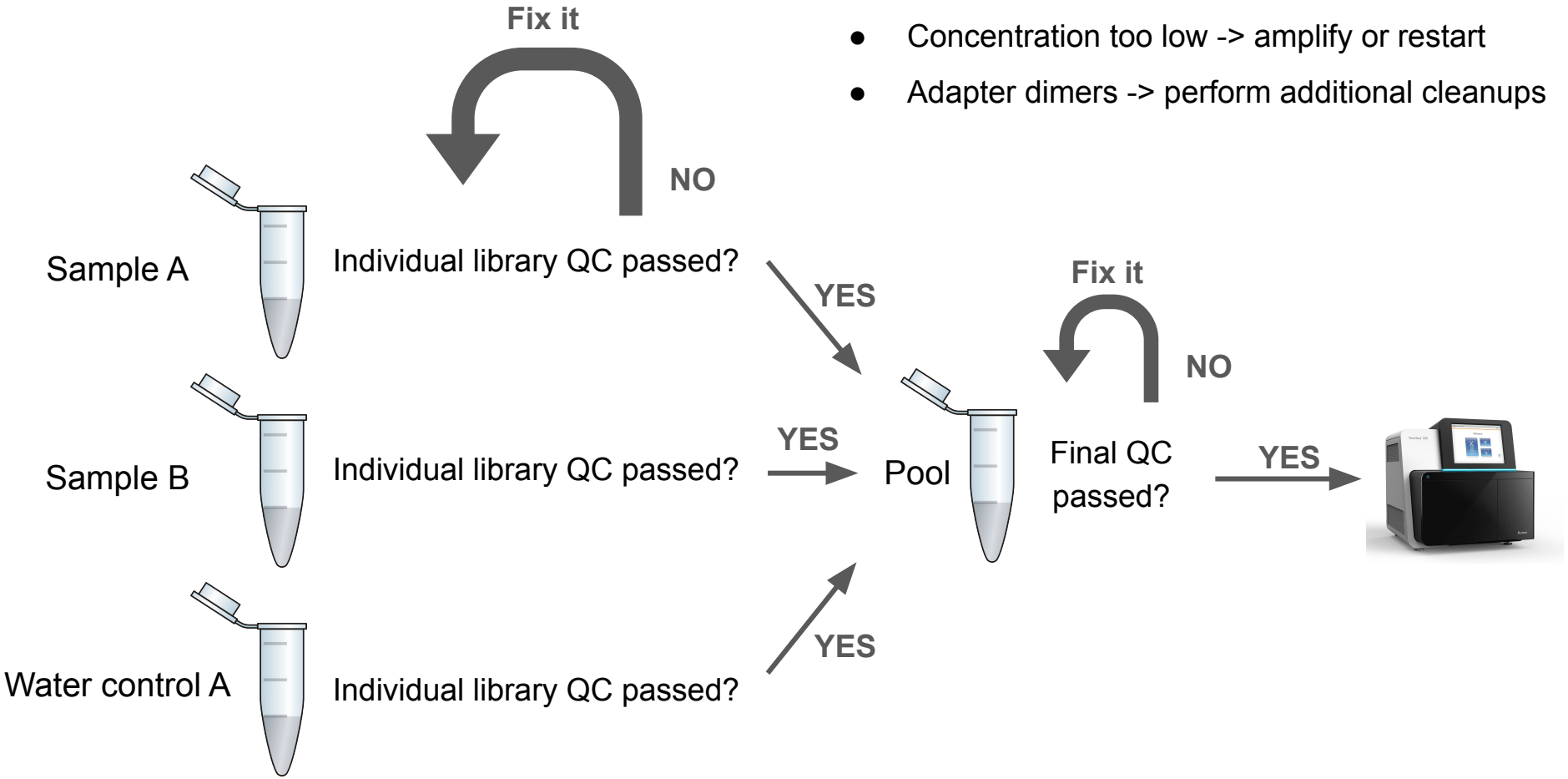
Adapter dimers

ALL DIMERS SHOULD BE REMOVED!

Library prep QC

How to fix samples with QC issues?

- Concentration too low -> amplify or restart
- Adapter dimers -> perform additional cleanups



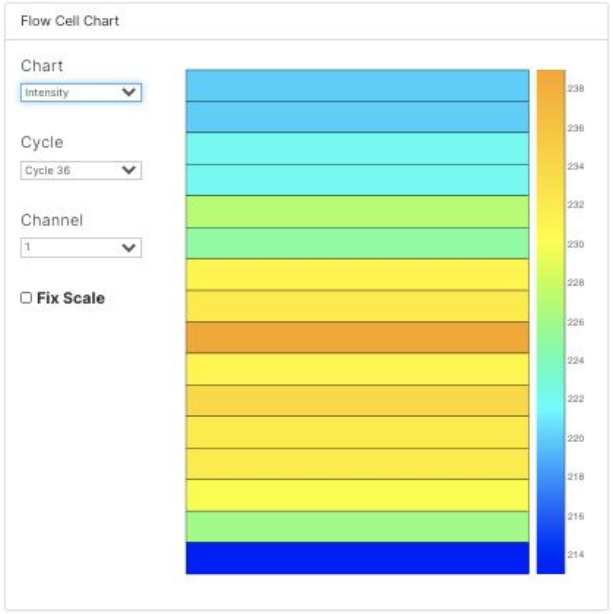
PhiX control

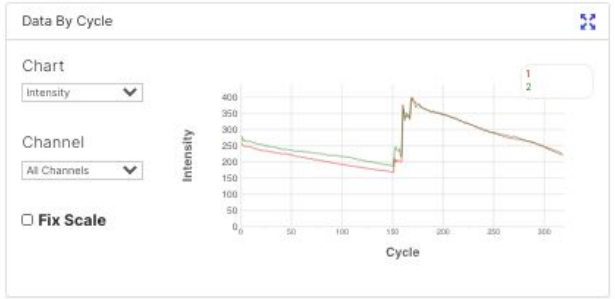
- PhiX is a control library that can be spiked into your library right before loading the sequencer
  - PhiX is a bacteriophage, single stranded DNA genome of 5386 -> known sequence
  - PhiX spike-in comes already fragmented and adapter-ligated -> ready to sequence
  - Spike-in concentration usually 5%

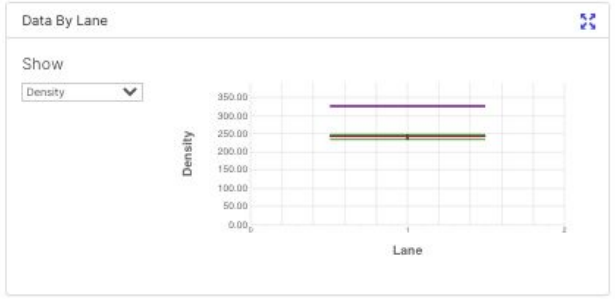
- Advantages of using PhiX:
  - Enables calculation of error rate
  - If a sequencing run fails, Illumina will evaluate PhiX data to assess whether or not the kit was the issue (and send a replacement if so)

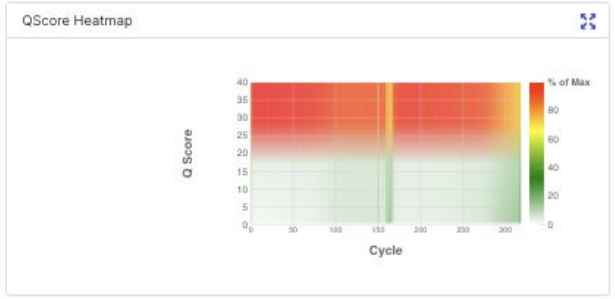
Sequencing Run QC

Flow Cell: BPL20321-2824 Extracted: 318 Called: 318 Scored: 318







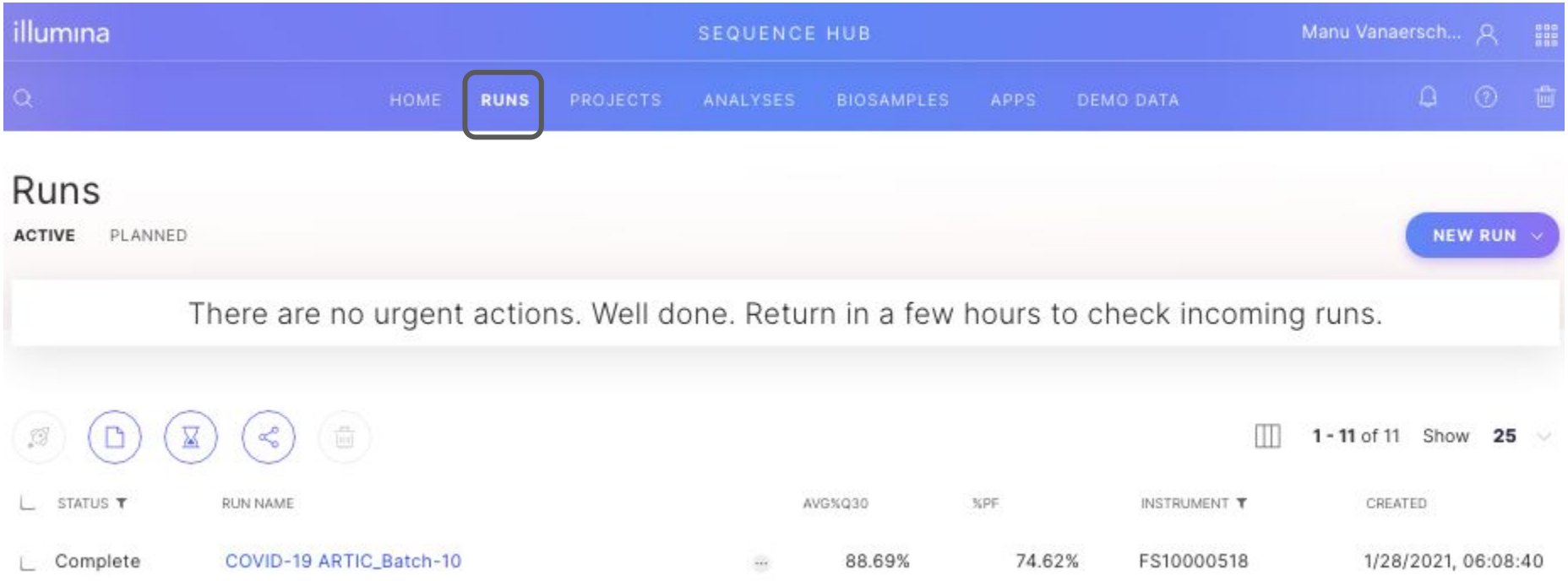


Did you get the expected amount of sequencing data?

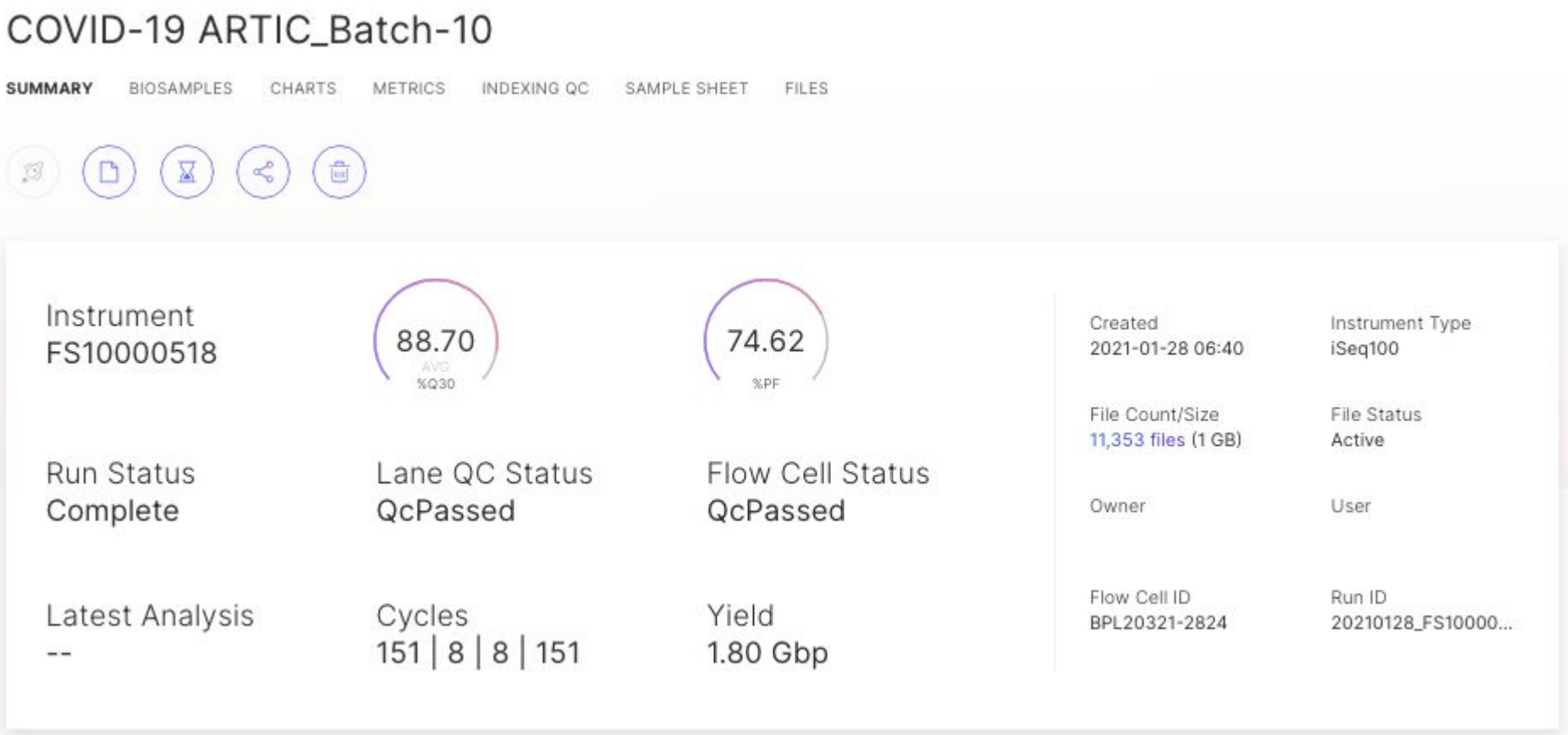
If not, why?

Sequencing Run QC using BaseSpace

After logging into BaseSpace, Click on ‘Runs’ to view a list of all your runs.
Click on a Run Name to see it’s metrics.



Sequencing Run QC using BaseSpace

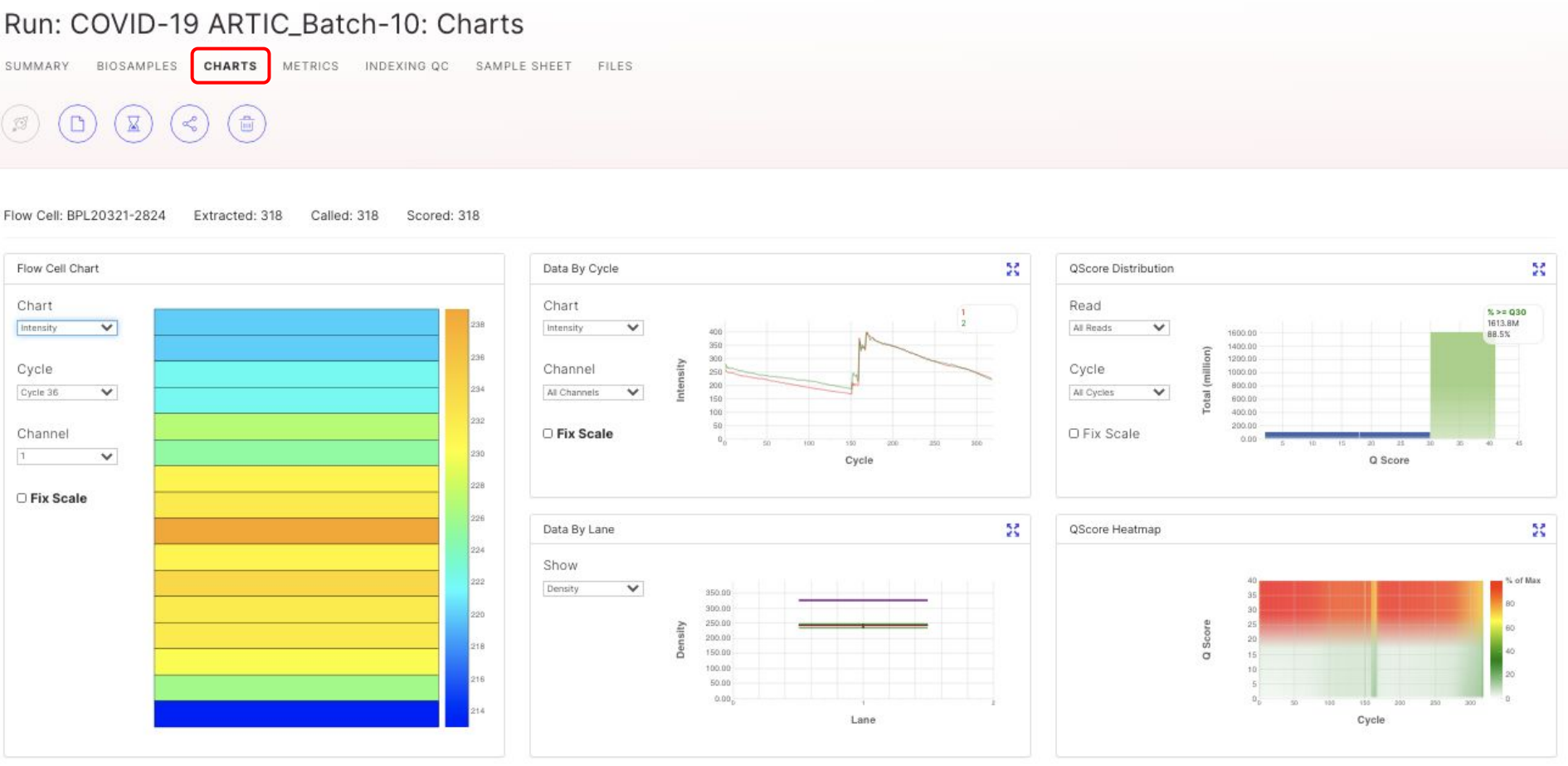


AVG%Q30 and %PF

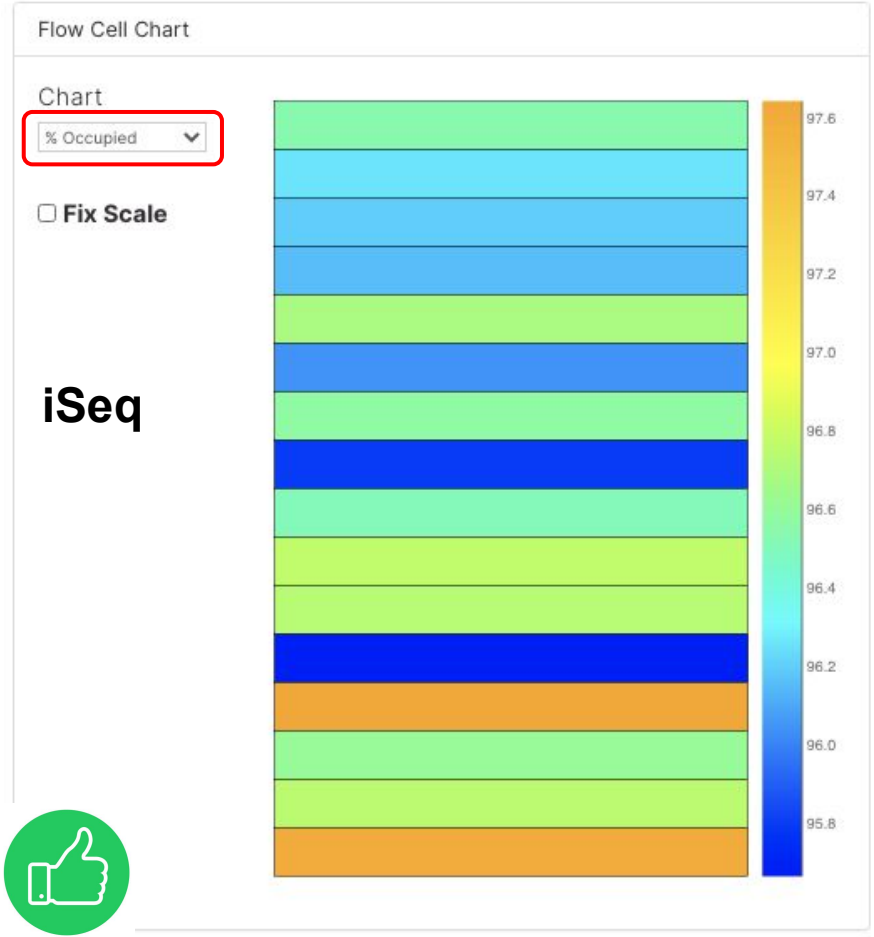
- **AVG%Q30** = percentage of reads that have bases with an average quality score >Q30
  - Ideally AVG%Q30 >90%, but >70% is good too
  - If <70%: evaluate why the quality is lower than expected. Possible reasons:
    - Overclustering: too much material was loaded on the flow cell
    - Adapter dimers: these will yield reads with only ~70bp of good signal, and the rest low quality ‘nothingness’.

- **%PF** = percentage of passed filter reads = reads based on a cluster of a single molecule
  - Ideally >70%, but >50% is good too
  - If <50%, check for indications of overclustering

QC charts


Run: COVID-19 ARTIC_Batch-10: Charts
SUMMARY
BIOSAMPLES
CHARTS
METRICS
INDEXING QC
SAMPLE SHEET
FILES
Flow Cell: BPL20321-2824
Extracted: 318
Called: 318
Scored: 318
Flow Cell Chart
Chart
Intensity
Cycle
Cycle 36
Channel
1
Fix Scale
238
236
234
232
230
228
226
224
222
220
218
216
214
Data By Cycle
Chart
Intensity
Channel
All Channels
Fix Scale
Intensity
Cycle
QScore Distribution
Read
All Reads
Cycle
All Cycles
Fix Scale
% >= Q30
1613.8M
88.5%
Total (million)
Q Score
Data By Lane
Show
Density
Density
Lane
QScore Heatmap
Q Score
Cycle
% of Max

Evaluate clustering for iSeq

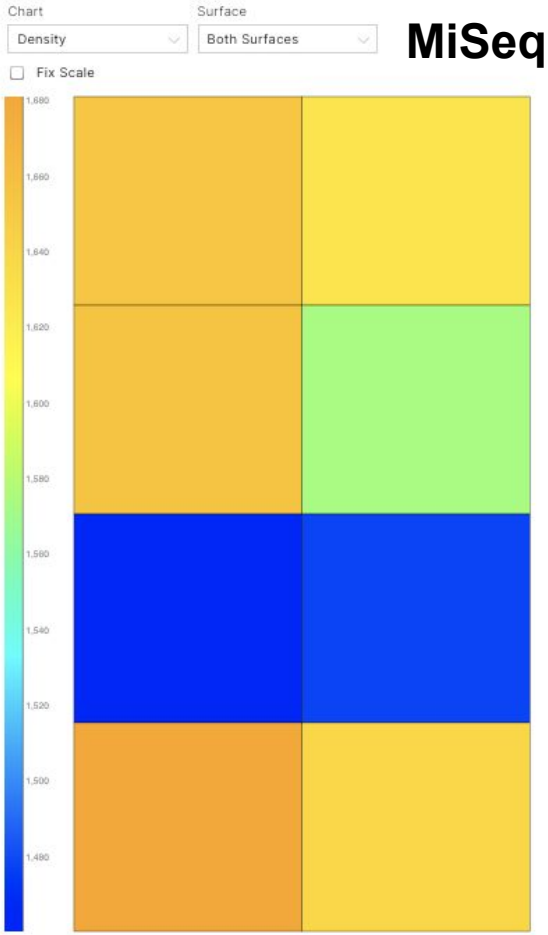


Flow Cell Chart: select % occupied under ‘Chart’

-> Signs of overclusterig or underclustering?

- Best range for values is 90-98%
- >98%: possible overclustering, which will reduce %PF and AVG%Q30 scores. Review library quantification data to check for errors or consider reducing the loading quantity for the next sequencing runs.
- <90%: Review library quantification data to check for errors or consider increasing the loading concentration for the next sequencing runs to get more reads from a run.

Evaluate clustering for MiSeq



MiSeq

Flow Cell Chart: select Density under ‘Chart’

Optimal Raw cluster density MiSeq:

- V2 kits: 1000-1200 K/mm^2
- V3 kits: 1200-1400 K/mm^2

This V3 example shows a raw cluster density ~1600 K/mm^2

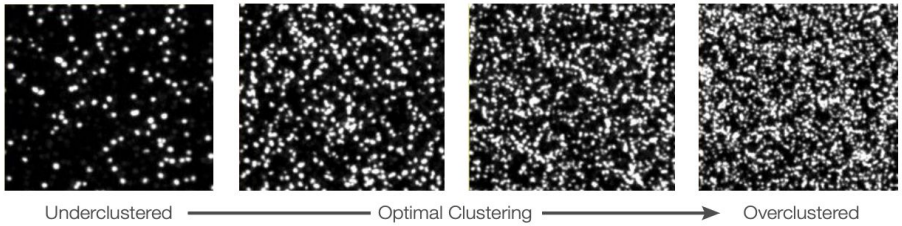
-> **The run was overclustered!**

-> lower %PF reads

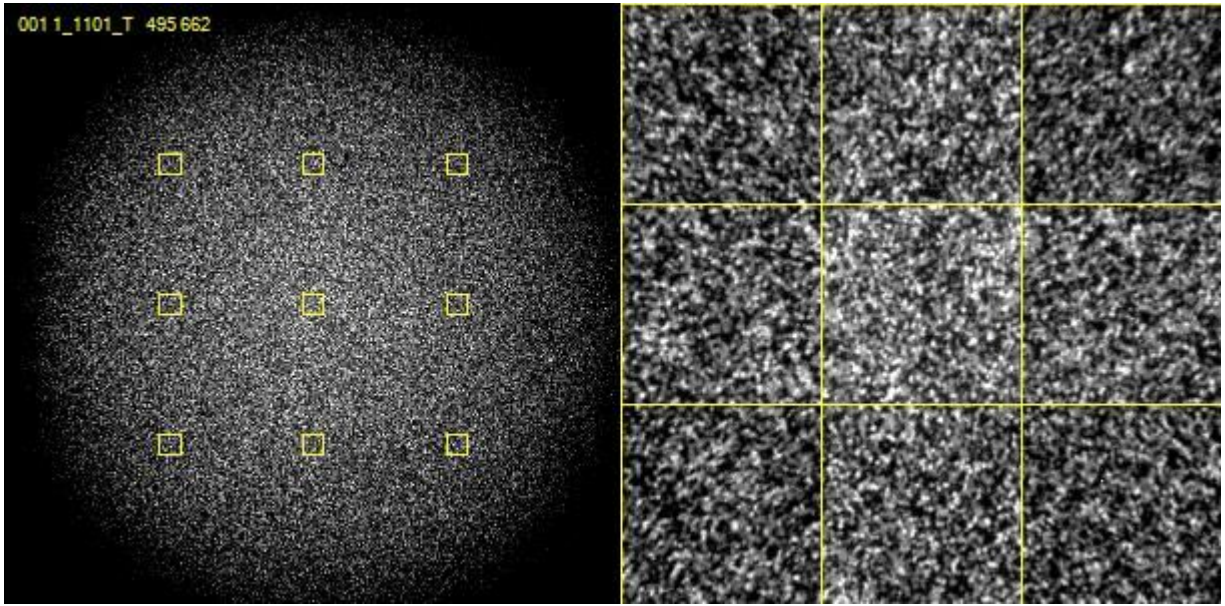
-> lower yield of PF reads total

Check QC data to see why concentration in nM of final library was underestimated (length overestimated?)

Evaluate clustering by looking at Files/Thumbnail_Images



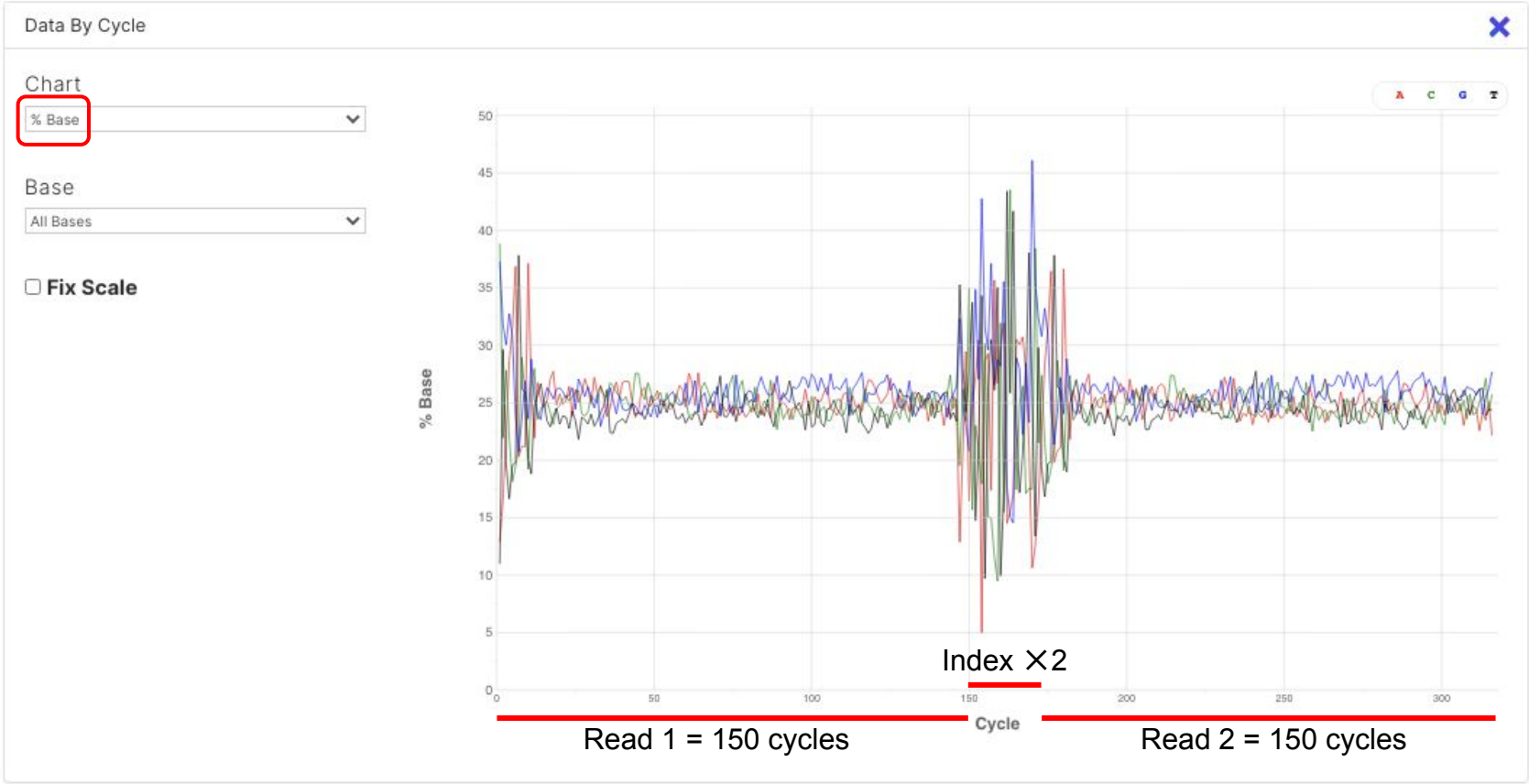
Illumina examples from here



Tiles also show overclustering of this MiSeq run

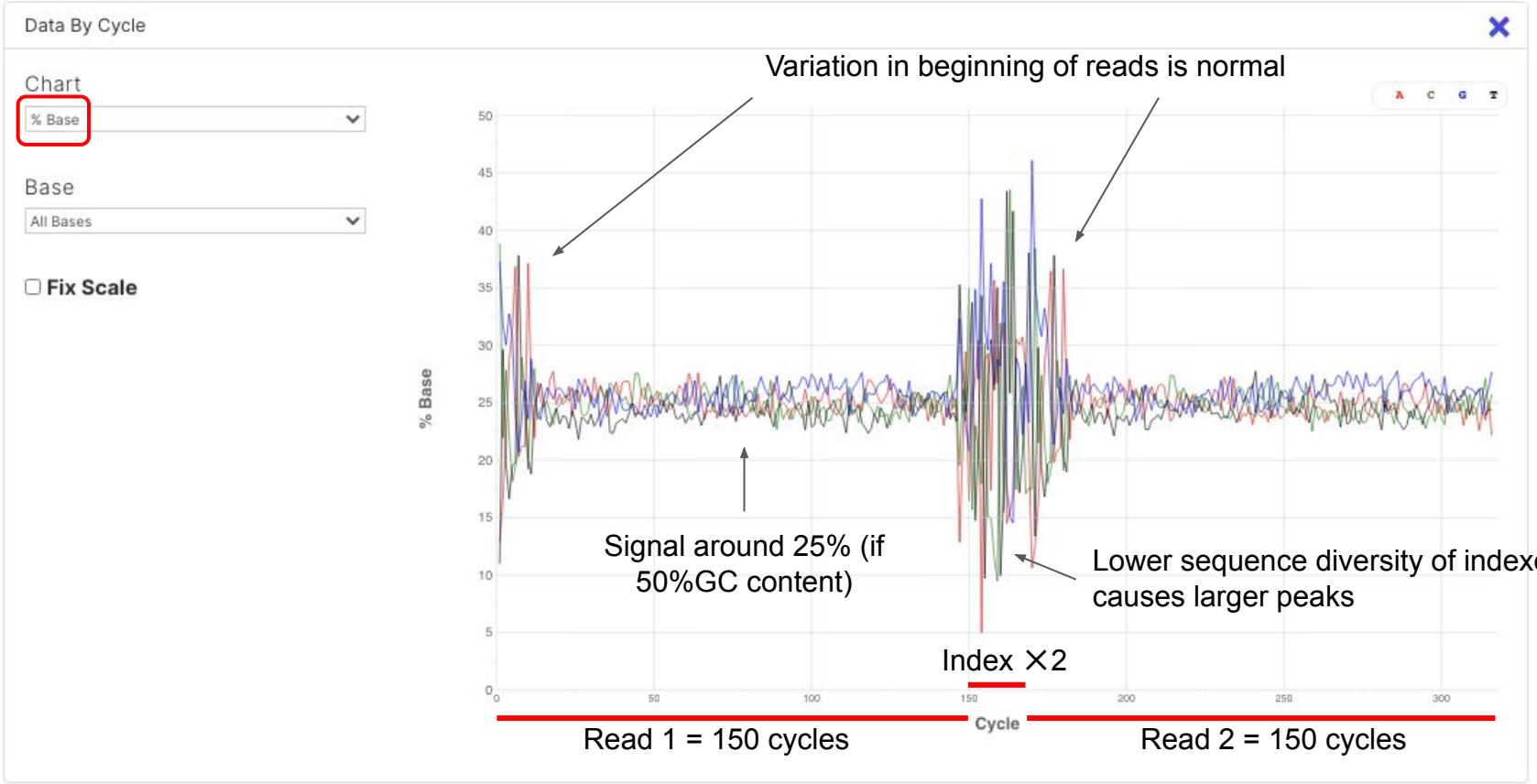
Evaluate level of adapter dimers and fragment length

Data by cycle chart: select % Base under ‘Chart’ -> Adapter dimers? Short fragments?

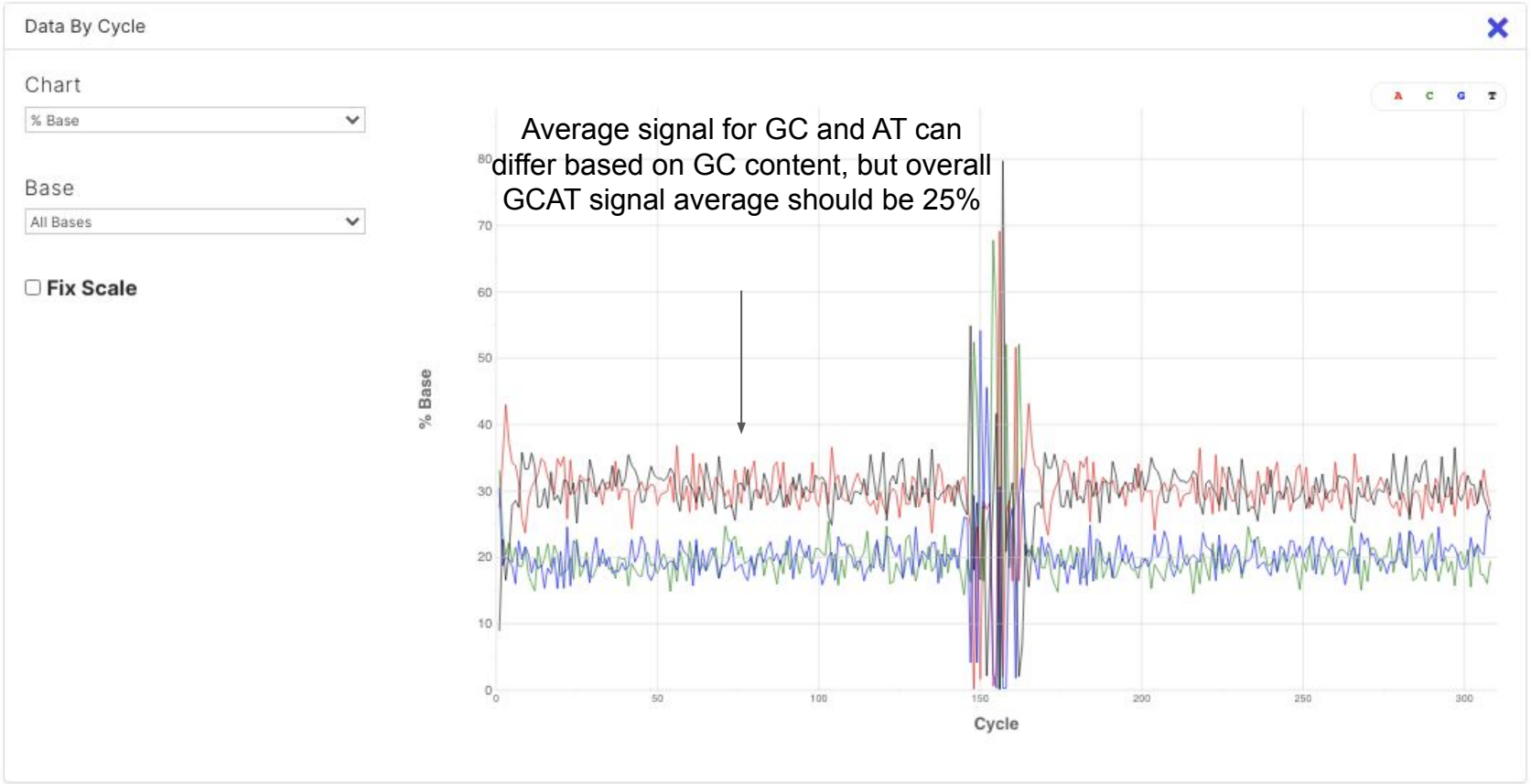


Evaluate level of adapter dimers and fragment length

Data by cycle chart: select % Base under ‘Chart’ -> Adapter dimers? Short fragments?



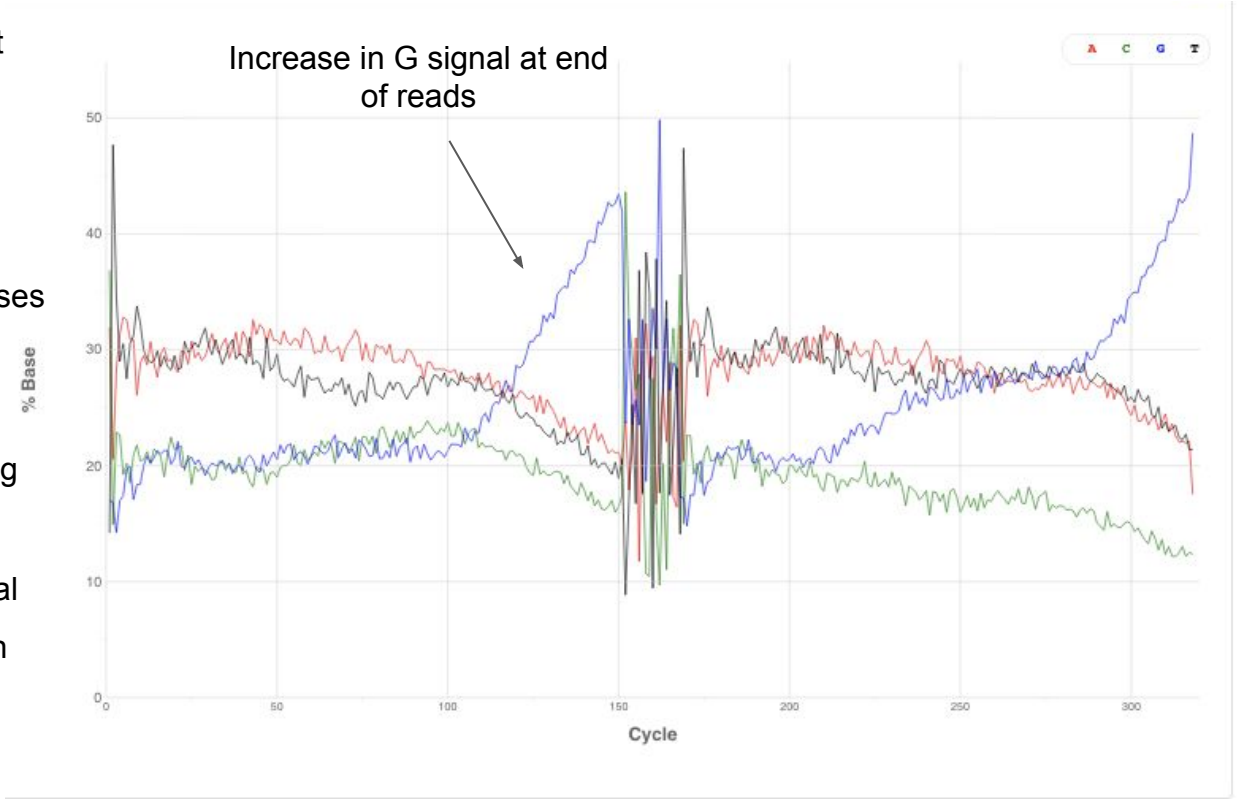
Evaluate level of adapter dimers and fragment length



Evaluate level of adapter dimers and fragment length

Issue 1 - short reads: Increase in G signal at end of reads

- G is the “dark base” is iSeq , MiniSeq, NextSeq and NovaSeq chemistry.
- Continuous increase in G indicates no bases detected for subsequent cycles
  - Inserts shorter than 150 bp?
  - Issue with elongation of strands during clustering?
- If insert size was indeed shorter than usual (check lib prep QC), reduce fragmentation time for future!

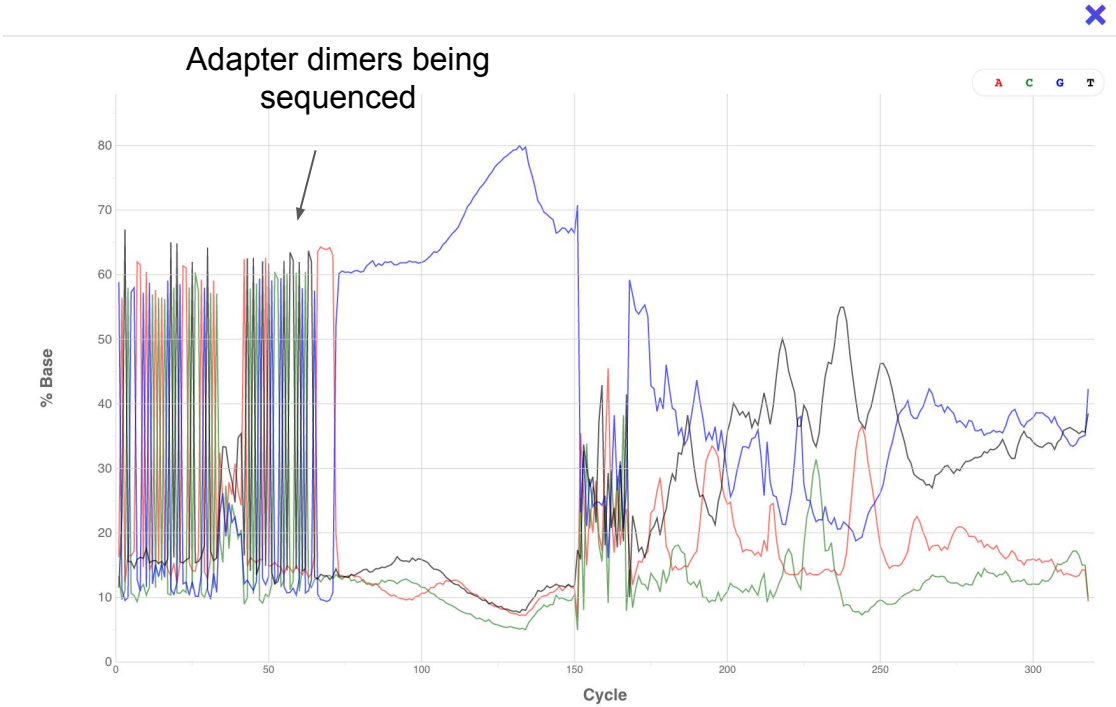


Note: MiSeq and HiSeq would call an A in case nothing is detected).

Evaluate level of adapter dimers and fragment length

Issue 2 - adapter dimers: Large peaks for first 70 cycles, and then an increase in G signal (iSeq, MiniSeq, NextSeq, NovaSeq)

- Adapter dimers will yield reads with low sequence variation (large peaks) that are about 70 bp long
- Adapter dimers are short fragments and thus get preferentially sequenced, taking up valuable sequencing space
- Crucial to remove ALL adapter dimers during library prep QC !!

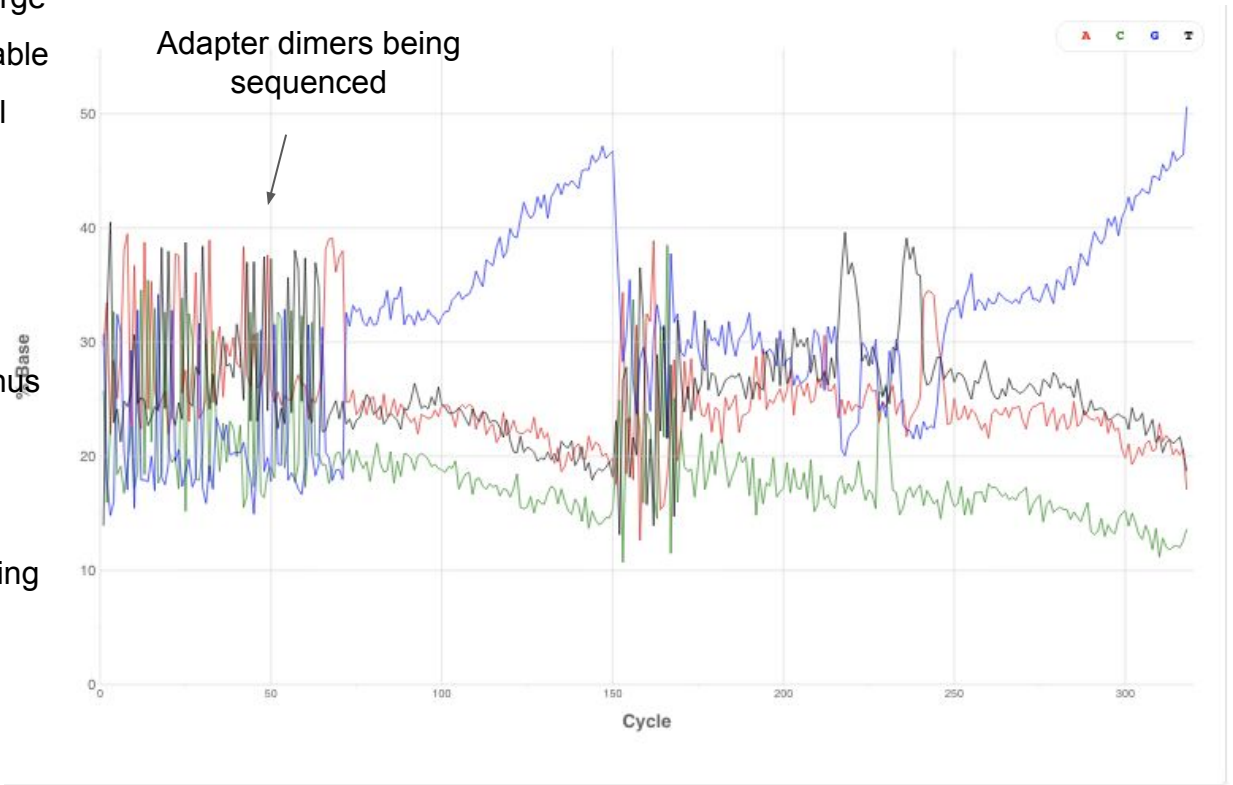


The first ~30 bases are the first portion of the illumina adapter, followed by heterogeneity from the barcode sequences (8-12 bases), followed by the last ~30 bases of the adapter. Then a large spike in the ‘dark’/no signal base ‘G’.

Evaluate level of adapter dimers and fragment length

Issue 3 - adapter dimers and short reads: Large peaks for first 70 cycles followed by more stable signal and then a steady increase in G signal

- Adapter dimers will yield reads with low sequence variation (large peaks) that are about 70 bp long
- Adapter dimers are short fragments and thus get preferentially sequenced, taking up valuable sequencing space
- Crucial to remove ALL adapter dimers during library prep QC !!



Evaluate error rate based on PhiX

COVID-19 ARTIC_Batch-10

SUMMARY BIOSAMPLES CHARTS **METRICS** INDEXING QC SAMPLE SHEET FILES

Per Read Metrics

READ	CYCLES	YIELD	PROJECTED YIELD	ALIGNED (%)	ERROR RATE (%)	INTENSITY CYCLE 1	%>Q30
Read 1	151	860.56 Mbp	860.56 Mbp	2.72	0.45	259.25	90.79
Read 2 (I)	8	40.16 Mbp	40.16 Mbp	0.00	0.00	210.56	87.55
Read 3 (I)	8	40.16 Mbp	40.16 Mbp	0.00	0.00	376.44	74.09
Read 4	151	860.56 Mbp	860.56 Mbp	2.43	1.09	390.75	87.34
Non-index Reads Total	**302**	**1.72 Gbp**	**1.72 Gbp**	**2.58**	**0.77**	**325.00**	**89.06**
Total	**318**	**1.80 Gbp**	**1.80 Gbp**	**2.58**	**0.77**	**309.25**	**88.69**

Evaluate sequencing data yield

Scroll down to 'Per Lane Metrics'

Number of paired end reads that were generated

Per Lane Metrics

LANE	STATUS	READ	CLUSTER PF(%)	%≥ Q30	YIELD	ERROR RATE(%)	READS PF	DENSITY	TILES	LEGACY PHAS / PREPHAS(%)	COMMENTS	INTENSITY
1	QC Passed	Read 1	74.62±1.18	90.79	0.86 Gbp	0.45 ±0.04	5,737,044	326 ±0	16	0.365 / 0.227		259±8
		Read 2 (I)		87.55	0.04 Gbp	0.00 ±0.00				0.000 / 0.000		211±6
		Read 3 (I)		74.09	0.04 Gbp	0.00 ±0.00				0.000 / 0.000		376±18
		Read 4		87.34	0.86 Gbp	1.09 ±0.06				0.269 / 0.127		391±19

Evaluate level of adapter dimers and fragment length

On the ‘Metrics’ page (if using PhiX)

- Per Read Metrics:
  a. **% Aligned** (PhiX): should be close to % of PhiX that you spiked in.
     i. If it’s higher, PhiX outcompeted your library, meaning you put less library in than you thought (double check your pooling and quantification calculations).
     ii. If it’s lower than expected, small fragments like adapter dimers might have outcompeted PhiX AND/OR you underestimated the quantity of material in your sample (added way more than you thought, double check your quantification data).
  b. **Error rate %** (PhiX): should be less than 1%.
- Per Lane Metrics:
  c. **Reads PF**: number of passed filter reads that were generated in this run. This should be close to (or over) the minimum number of guaranteed reads that your sequencer is advertised to generate. If lower, you probably under- or overclustered. Check your quantification calculations and adjust the loading concentration if needed.

Evaluate demultiplexing

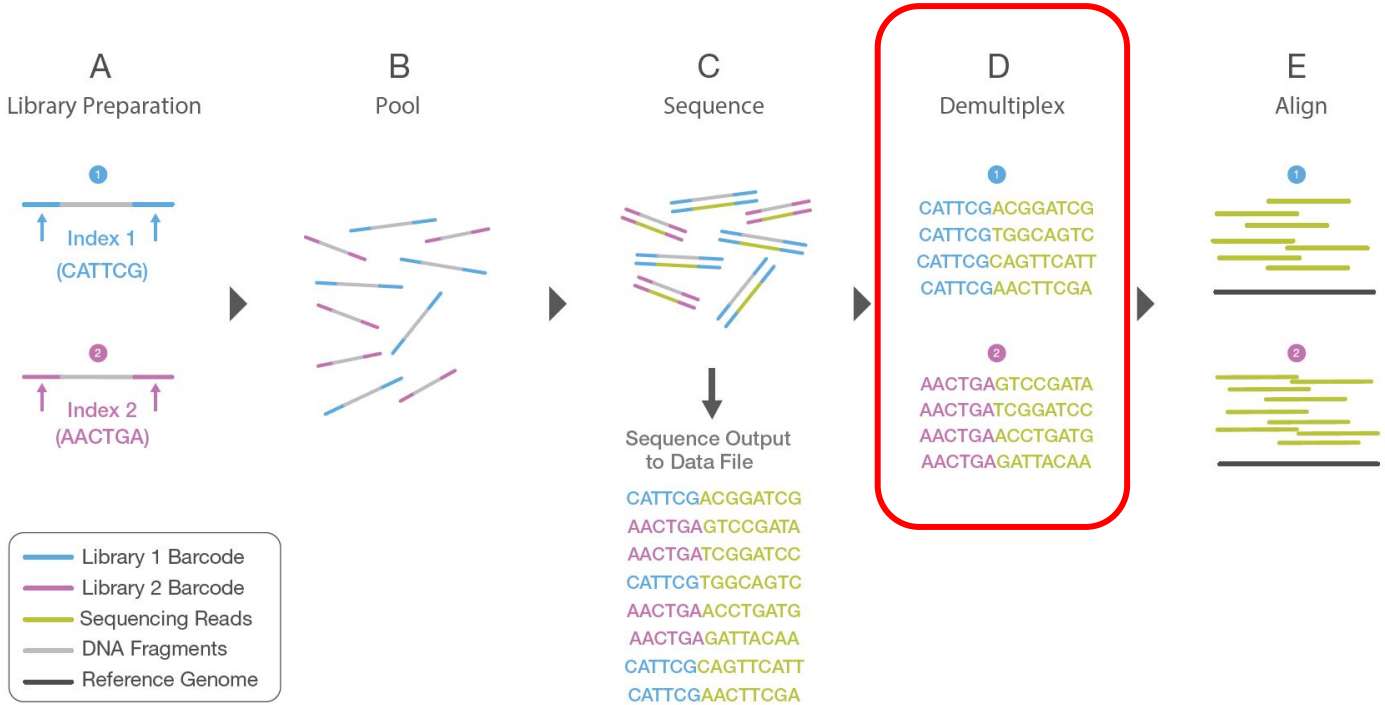


Figure 5: Library Multiplexing Overview.

a. Two distinct libraries are attached to unique index sequences. Index sequences are attached during library preparation.

b. Libraries are pooled together and loaded into the same flow cell lane.

c. Libraries are sequenced together during a single instrument run. All sequences are exported to a single output file.

d. A demultiplexing algorithm sorts the reads into different files according to their indexes.

e. Each set of reads is aligned to the appropriate reference sequence.

Adapted from Illumina.com

Evaluate demultiplexing

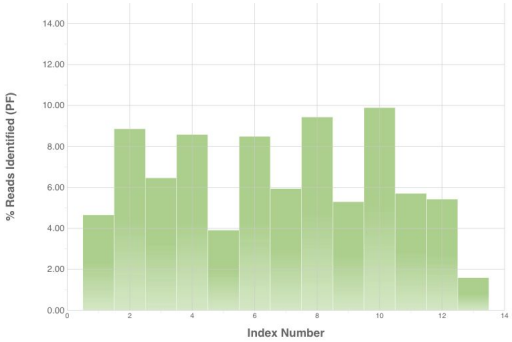
SUMMARY | BIOSAMPLES | CHARTS | METRICS | **INDEXING QC** | SAMPLE SHEET | FILES

TOTAL READS	PF READS	% READS IDENTIFIED (PF)	% READS UNDETERMINED	CV	MIN	MAX
15,376,000	11,474,088	84.2893	15.7107	0.3772	1.5917	9.8988

INDEX	SAMPLE ID	LIBRARY NAME	INDEX 1 (I7)	INDEX 2 (I5)	% READS IDENTIFIED (PF)
1	COVID_0125	n/a	CGATCGAT	TCGAGTGA	4.6542
2	COVID_0126	n/a	GCCTTAAC	CGCTTAAC	8.8601
3	COVID_0127	n/a	AGTCAGGT	GATAGGCT	6.4675
4	COVID_0128	n/a	CTAGGTTG	ATGGAAGG	8.5830
5	COVID_0129	n/a	CCTCATCT	TAATGCCG	3.9148

... ...

If correctly pooled, all samples should have more or less the same # of reads assigned



Evaluate demultiplexing

On the ‘Indexing QC’ page:

a. **% read identified =** reads with identified barcodes: should be >80% (the higher the better). If this % is low, you should check to make sure your sample sheet assigned the correct index/barcode sequences for each sample.

b. **% reads undetermined =** reads without an associated known barcode. If this % is high, you should check to make sure your sample sheet was correct.

c. **PF reads**: this number details the amount of (unpaired) single end reads. This number is always double than your true amount of paired-end reads shown on the ‘Metrics’ page.

Troubleshoot demultiplexing

- A high % of undetermined reads and missing samples (samples with no reads assigned) indicate an issue with demultiplexing!

    - Double check the sample sheet to verify if the correct indices/barcodes were assigned to each sample.

    - Look at the barcodes sequences identified for ‘undetermined reads’ listed in DemuxSummaryF1L1.txt, if there is barcode combination (paired end) that occured in a high number of reads, it might be an unidentified sample.

Info on finding troubleshooting and finding the demux file in BaseSpace here
How to edit sample sheet and requeue in Basespace here
Info on finding troubleshooting and finding the demux file using Local Run Manager here
Info on finding troubleshooting and finding the demux file using MiSeq Reporter here
How to edit sample sheet and requeue using MiSeq Reporter here

Summary library and sequencing run QC

- Concentration
  - Accurate measurements are crucial, use fluorometry-based method (e.g. Qubit),
    - do NOT use spectrophotometry (e.g. no Nanodrop)
  - Make sure values are within dynamic range of assay
  - If too low, amplify library if needed, or adjust # PCR cycles of indexing PCR step for future runs

- Library length
  - Remove all adapter dimers to prevent them taking up sequencing space
  - Capillary electrophoresis is the fastest and most accurate method
  - PCR + gel electrophoresis is an alternative method if capillary electroph. is not available

- Sequencing run QC:
  - Optimal loading concentration will yield maximum amount and quality of data
  - Monitor sequencing runs and adjust loading concentration of future runs as needed