Post-NGS Data Analysis

Preonath Chondrow Dev

Bioinformatician

Child Health Research Foundation

Date: 15 October 2023

CHRF Child Health Research Foundation

Prevent Infections, Save Lives

Contents

- Sample sheet preparation
- Sequencing quality check (Q30 and PF)
- RNA-virus consensus preparation (CZ ID)
- Virtual tour to basespace
- Troubleshooting

1. BCL to Fastq conversion
2. Requeue

- Quality Checking of fastq files
- Quality Quantrol
- Genome Assembly

Sample sheet preparation

A. Required Fields:

1. Experiment Name

2. Sample_ID

3. Sample_Name

4. Description

5. Index

6. I7_Index_ID

Same but different field name

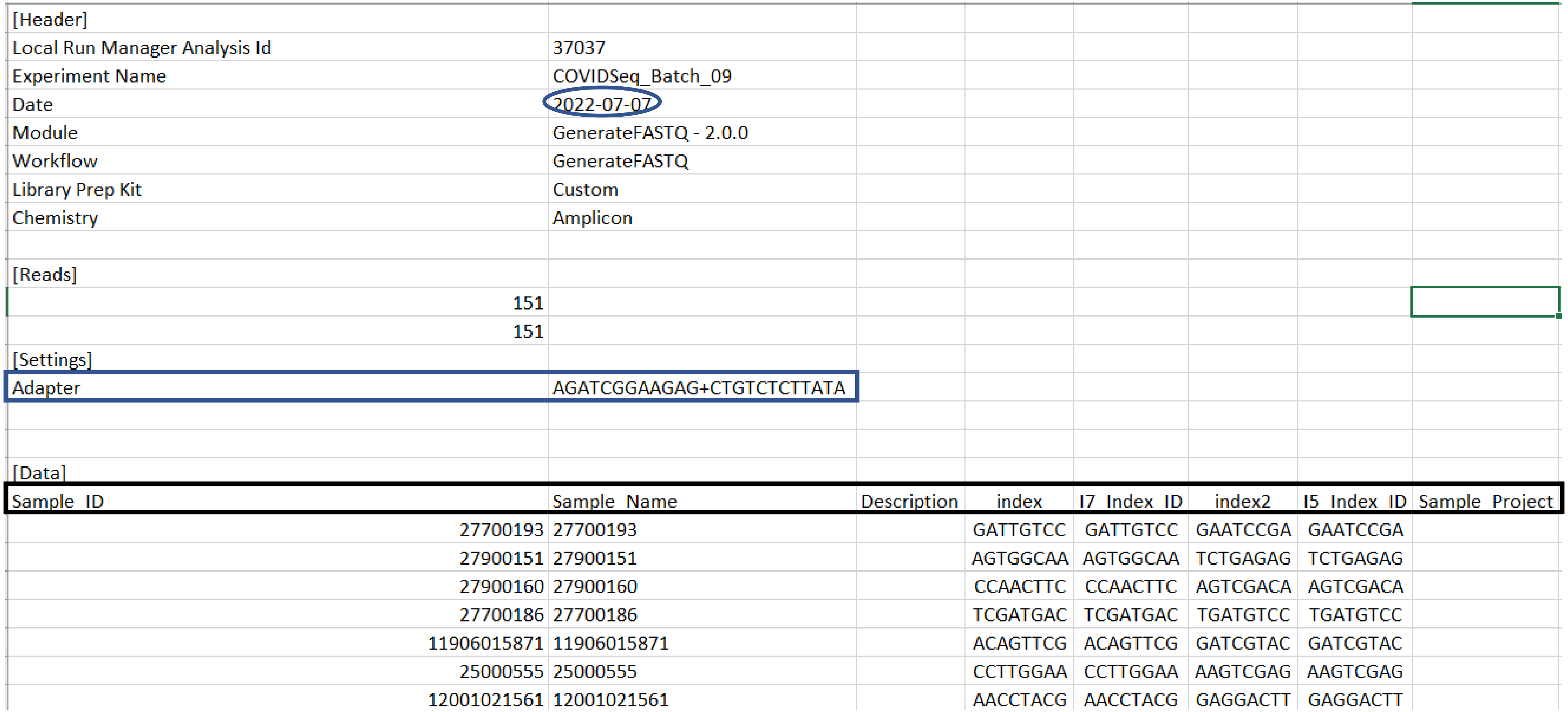
7. index2

8. I5_Index_ID

Same but different field name

10. Sample_Project

Sample sheet Template



[Header]							
Local Run Manager Analysis Id	37037						
Experiment Name	COVIDSeq_Batch_09						
Date	2022-07-07						
Module	GenerateFASTQ - 2.0.0						
Workflow	GenerateFASTQ						
Library Prep Kit	Custom						
Chemistry	Amplicon						
[Reads]							
151							
151							
[Settings]							
Adapter	AGATCGGAAGAG+CTGTCTCTTATA						
[Data]							
Sample_ID	Sample_Name	Description	index	I7_Index_ID	index2	I5_Index_ID	Sample_Project
27700193	27700193		GATTGTCC	GATTGTCC	GAATCCGA	GAATCCGA	
27900151	27900151		AGTGGCAA	AGTGGCAA	TCTGAGAG	TCTGAGAG	
27900160	27900160		CCAACTTC	CCAACTTC	AGTCGACA	AGTCGACA	
27700186	27700186		TCGATGAC	TCGATGAC	TGATGTCC	TGATGTCC	
11906015871	11906015871		ACAGTTCG	ACAGTTCG	GATCGTAC	GATCGTAC	
25000555	25000555		CCTTGGAA	CCTTGGAA	AAGTCGAG	AAGTCGAG	
12001021561	12001021561		AACCTACG	AACCTACG	GAGGACTT	GAGGACTT	

Sample sheet preparation

A. Precautions:

1. Never use space between words (e.g. KPN 132) . Corrected form (KPN_132 , CHRF_COVID_098)

2. Delete last 2 bp from index if Barcode has **10bp** sequences.

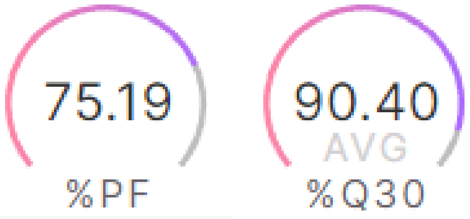
3. Delete first 2 bp from index2 Barcode has **10bp** sequences

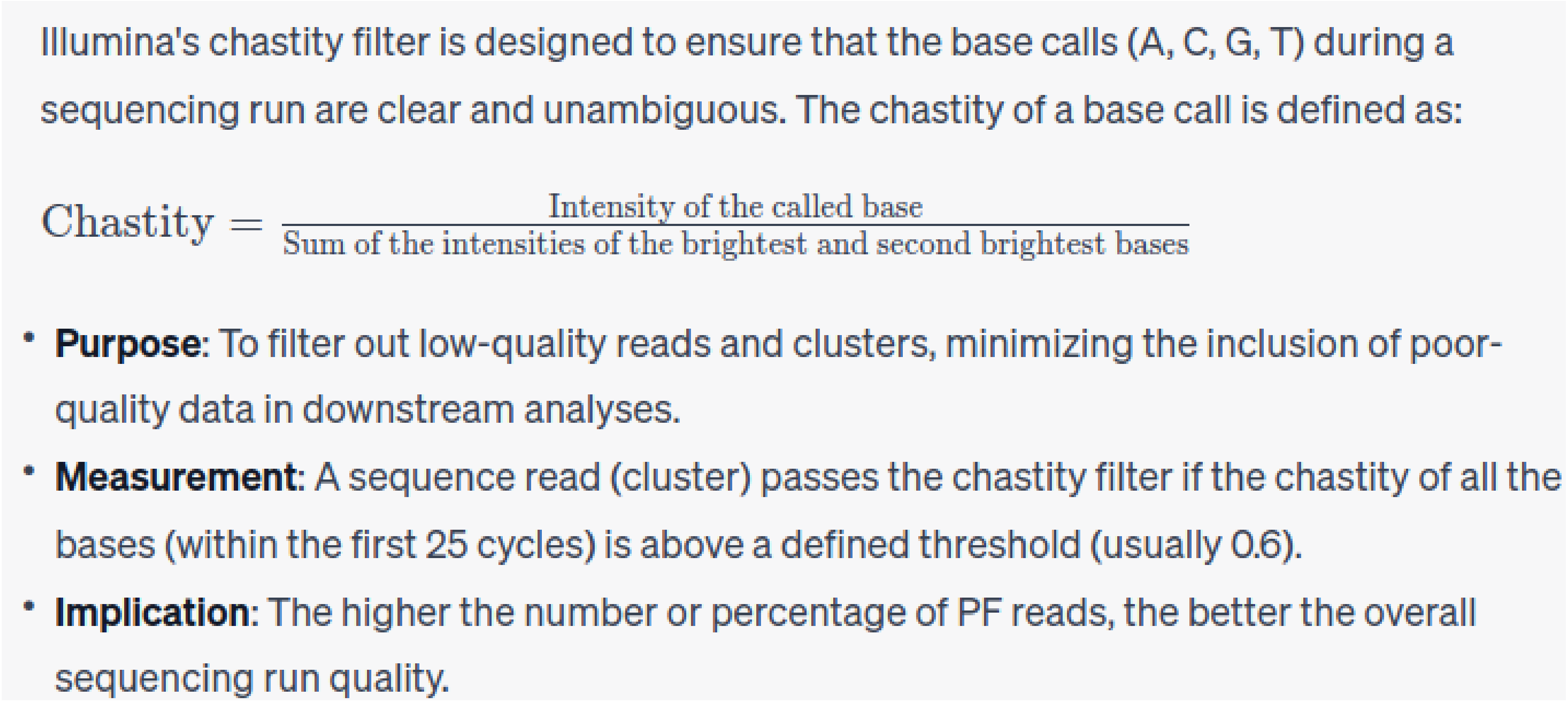
4. Date format should be this way “YYYY-MM-DD”

5. It should be a CSV file.

Sequencing quality check (PF)

PF (Pass filter)





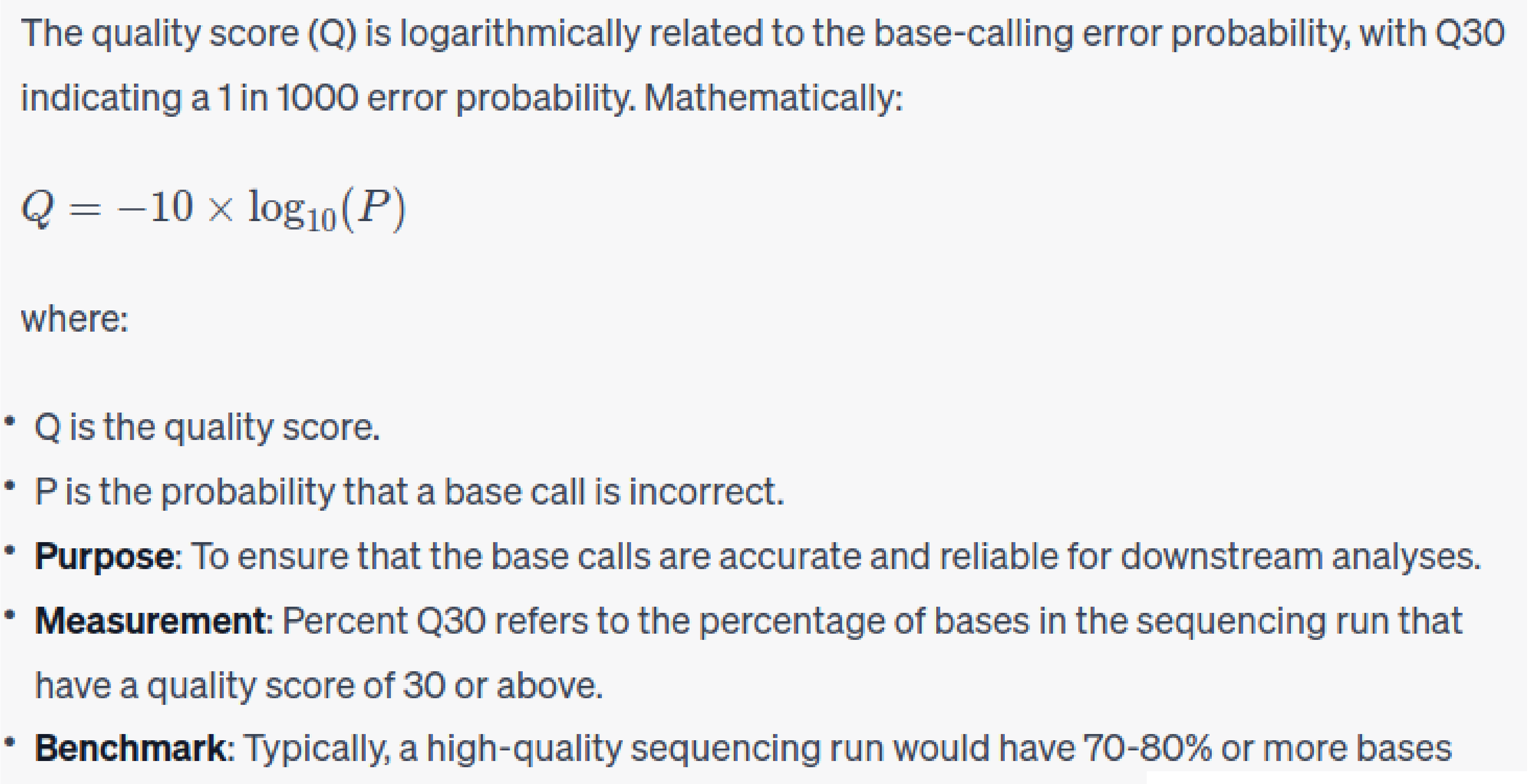
Illumina's chastity filter is designed to ensure that the base calls (A, C, G, T) during a sequencing run are clear and unambiguous. The chastity of a base call is defined as:

$$\text{Chastity} = \frac{\text{Intensity of the called base}}{\text{Sum of the intensities of the brightest and second brightest bases}}$$

- **Purpose**: To filter out low-quality reads and clusters, minimizing the inclusion of poor-quality data in downstream analyses.
- **Measurement**: A sequence read (cluster) passes the chastity filter if the chastity of all the bases (within the first 25 cycles) is above a defined threshold (usually 0.6).
- **Implication**: The higher the number or percentage of PF reads, the better the overall sequencing run quality.

Sequencing quality check (Q30)

Q30

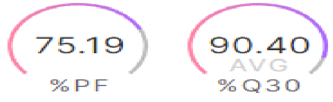


The quality score (Q) is logarithmically related to the base-calling error probability, with Q30 indicating a 1 in 1000 error probability. Mathematically:

$$Q = -10 \times \log_{10}(P)$$

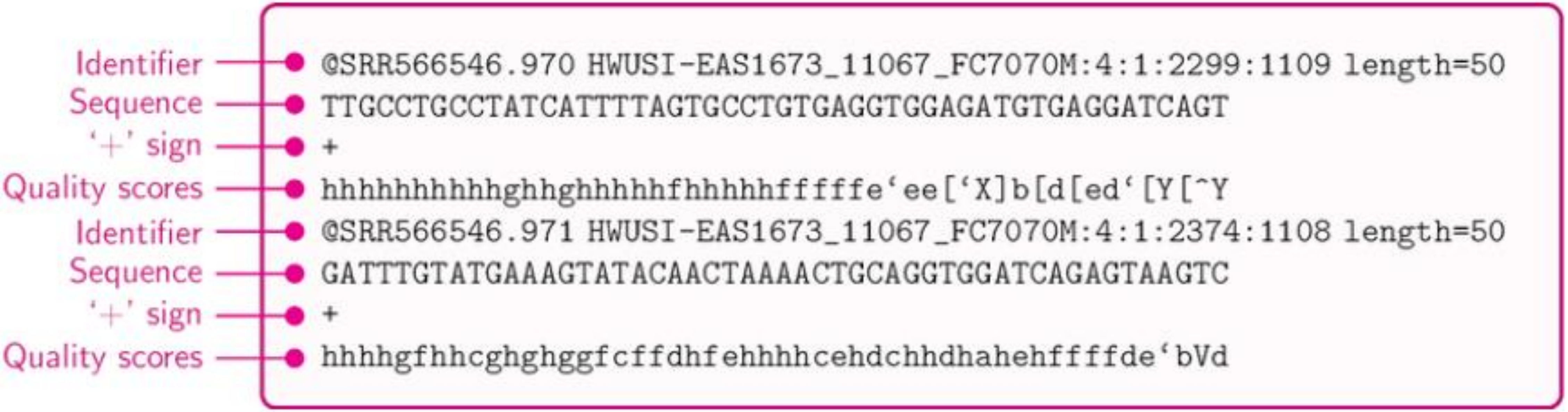
where:

- Q is the quality score.
- P is the probability that a base call is incorrect.
- **Purpose**: To ensure that the base calls are accurate and reliable for downstream analyses.
- **Measurement**: Percent Q30 refers to the percentage of bases in the sequencing run that have a quality score of 30 or above.
- **Benchmark**: Typically, a high-quality sequencing run would have 70-80% or more bases with a Q30 score.



Sequencing quality check (Q30 and PF)

Fastq File:



```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
+
hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde‘bVd
```

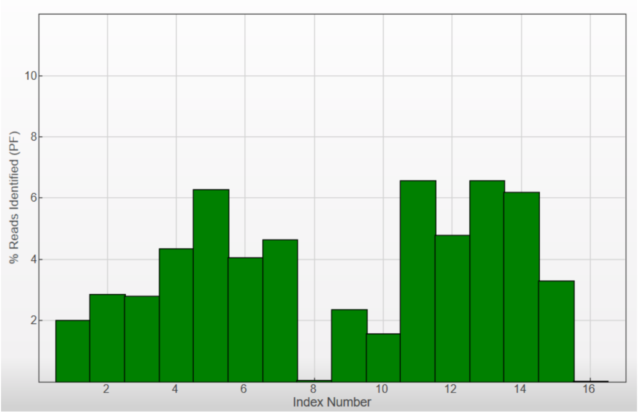
Child Health Research Foundation

CHRF

Prevent Infections, Save Lives

Illumina Sequence Analysis Viewer

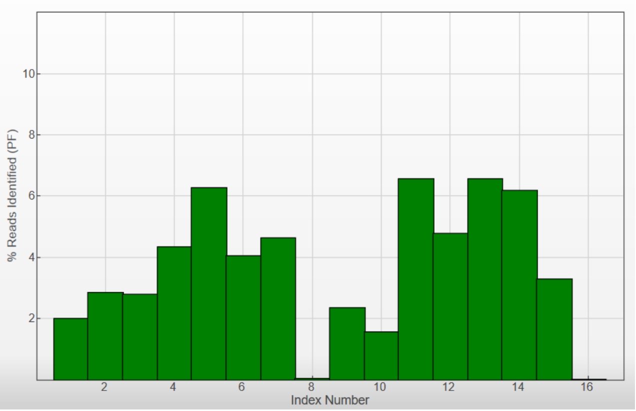
1. **Real-time Monitoring**
    - SAV allows users to monitor the ongoing sequencing run,
    - Providing insights into various parameters, such as
        - cluster density,
        - cluster passing filter (%PF)
        - Phred quality score (e.g., %Q30).



Illumina Sequence Analysis Viewer

2. Indexing Ratio

- SAV enables users to view the indexing ratio
  - Refers to the distribution of indexed reads among all the samples in a multiplexed run.
  - An optimal index ratio ensures equal representation of all samples
  - An optimal index ratio is crucial for de-multiplexing and downstream analyses.

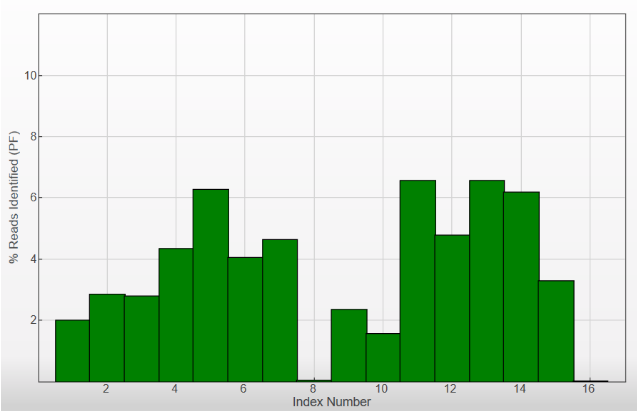


Illumina Sequence Analysis Viewer

3. **Loading Concentration Quantification**

- The software helps users quantify the loading concentration by providing metrics like cluster density,

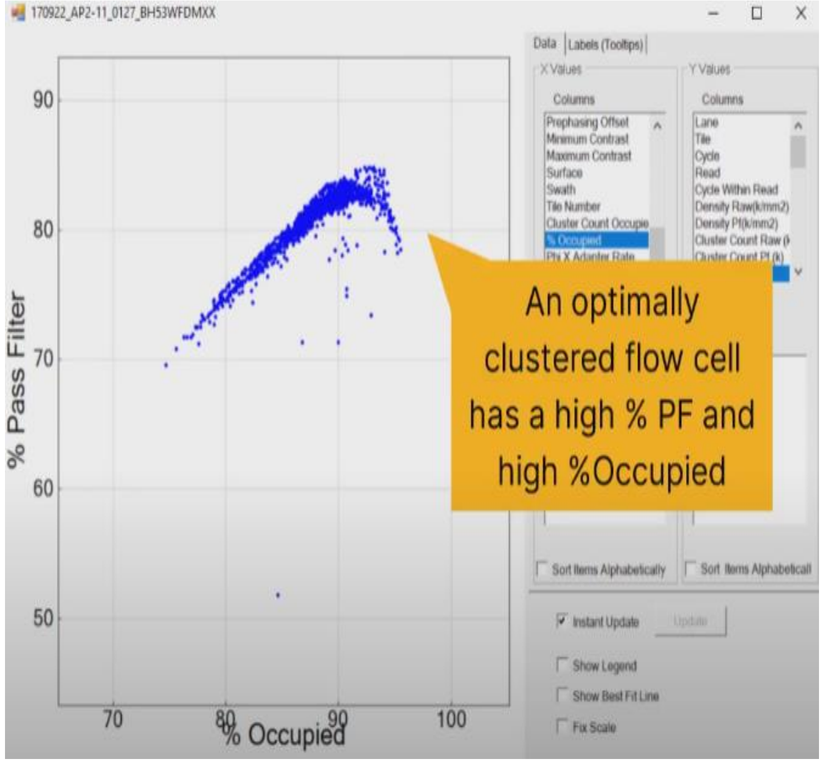
- It is crucial to ensure optimal data output and to prevent issues like over-clustering or under-clustering.



Illumina Sequence Analysis Viewer

Example 1: **Optimal Loading Concentration**

- **Scenario**: A run with optimal loading concentration.
- **SAV Visualization**:
  - **X-Axis (% Occupied)**: Moderately high – indicating a good number of clusters.
  - **Y-Axis (%PF)**: High – indicating that a large portion of clusters is of good quality.
- **Interpretation**: This suggests that the library concentration was optimal, yielding a high-quality sequencing run.



An optimally clustered flow cell has a high % PF and high %Occupied

- **% Occupied**:
  - Refers to the percentage of total tiles that contain clusters.
  - High % Occupied could mean high cluster density,
  - Low % Occupied indicates fewer clusters.

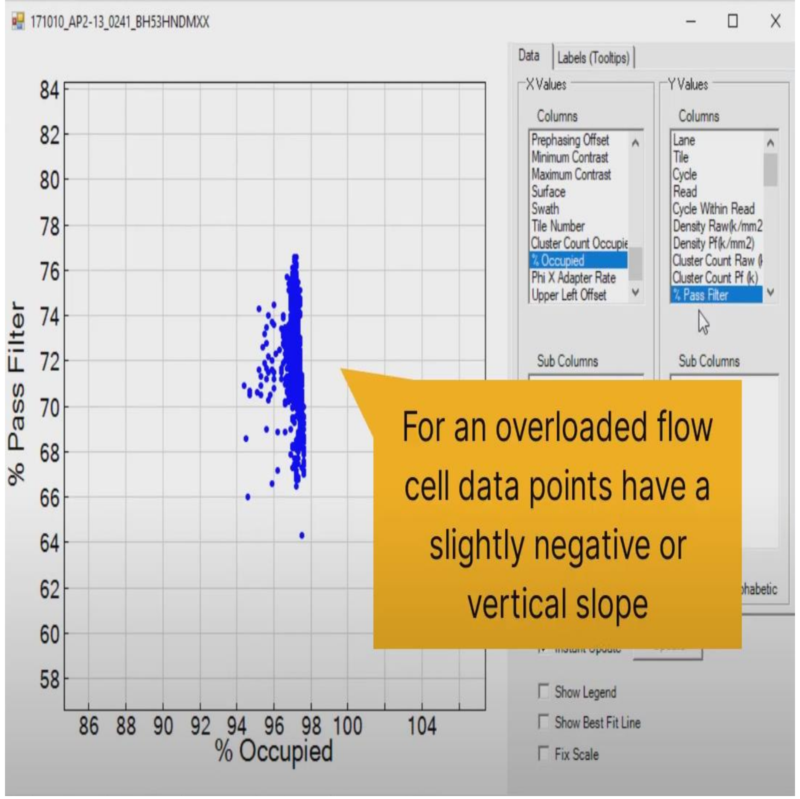
Illumina Sequence Analysis Viewer

Example 2: **Overloaded**

- **Scenario**: A run with too high loading concentration.

- **SAV Visualization**:
  - **X-Axis (% Occupied)**: Very high – indicating an excess of clusters.
  - **Y-Axis (%PF)**: Possibly lower – due to increased signal overlap and noise, reducing the number of clusters that pass the chastity filter.

- **Interpretation**: Too many DNA fragments were loaded, resulting in over-clustering, which can affect the quality of the sequencing data due to overlapping signals and increased noise.



171010_AP2-13_0241_BH53HNDMXX

X Values – Columns: % Occupied

Y Values – Columns: % Pass Filter

□ Show Legend

□ Show Best Fit Line

□ Fix Scale

For an overloaded flow cell data points have a slightly negative or vertical slope

Illumina Sequence Analysis Viewer

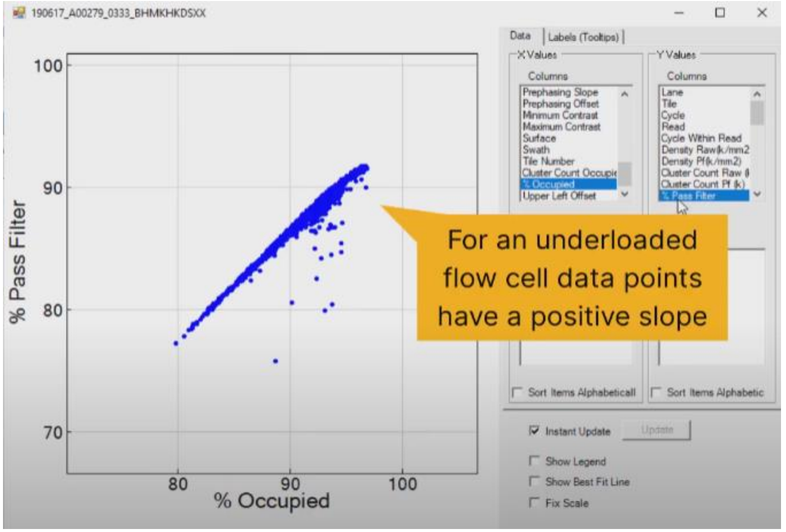
Example 3: Underloaded

- **Scenario**: A sequencing run where the data quality is suboptimal.

- **SAV Visualization**:
  - **X-Axis (% Occupied)**: Low – indicating sparse clusters across the flow cell.
  - **Y-Axis (%PF)**: Can be variable – may still be high if the clusters present are of good quality.

- **Interpretation**:
  - Underloading means fewer clusters are generated.
    - Which might be due to low library concentration.
    - While the data might still be of high quality (%PF)

  - the overall yield (data output) of the sequencing run would be reduced,



190617_A00279_0333_BHMKHKDSXX

X Values: % Occupied

Y Values: % Pass Filter

☑ Instant Update

□ Show Legend

□ Show Best Fit Line

□ Fix Scale

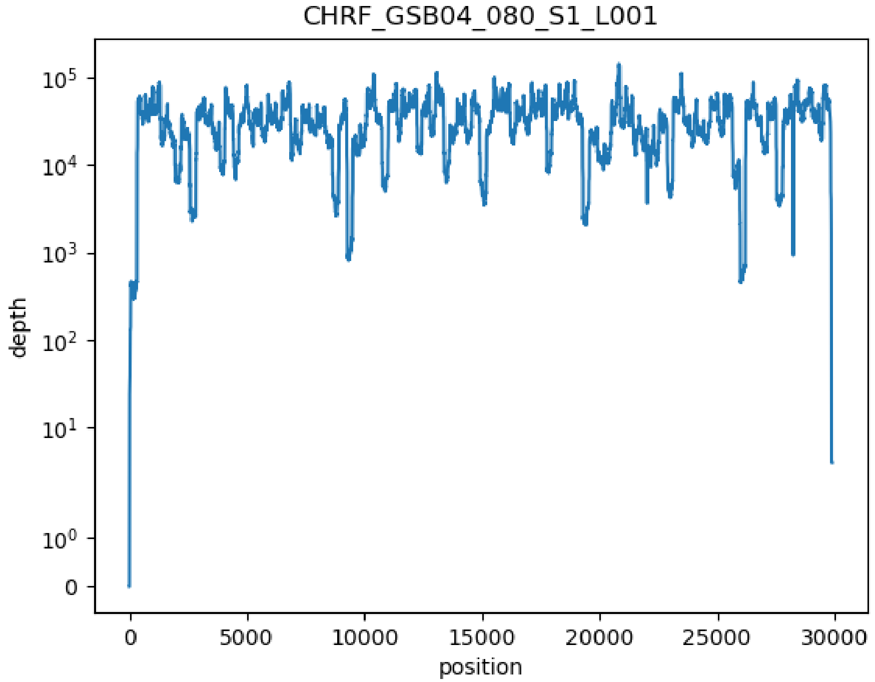
For an underloaded flow cell data points have a positive slope

RNA-virus Consensus Genome Preparation (CZ ID)

- sc2-illumina-pipeline output:

1. Fasta file (consensus)
2. Multiqc result
3. Coverage plot

Coverage depth plot →



sample_name	depth_avg	mapped_reads	total_reads	n_actg	n_missing	n_gap	n_ambiguous
CHRF_GSB04_080_S1_L001	33891.354	9472195	14415600	29838	49	0	7
CHRF_GSB04_081_S2_L001	37037.561	10301531	14074768	29846	12	0	12

RNA-virus Consensus Genome Preparation (CZ ID)

- **Sample_name**: The unique identifier for each sequenced sample.
- **Depth_avg**: The average depth of sequencing, which represents the average number of times a base is sequenced.
  - Higher depth increases confidence in the identified bases.
- **Mapped_reads**: The number of reads that were successfully mapped (aligned) to a reference genome.
  - High mapped read counts suggest good specificity in the sequencing.
- **Total_reads**: The total number of reads obtained after sequencing.

RNA-virus Consensus Genome Preparation (CZ ID)

- **n_actg**: Likely the number of bases (A, C, T, G) in the consensus genome.
  - This might give an idea about the length and quality of the consensus sequence.

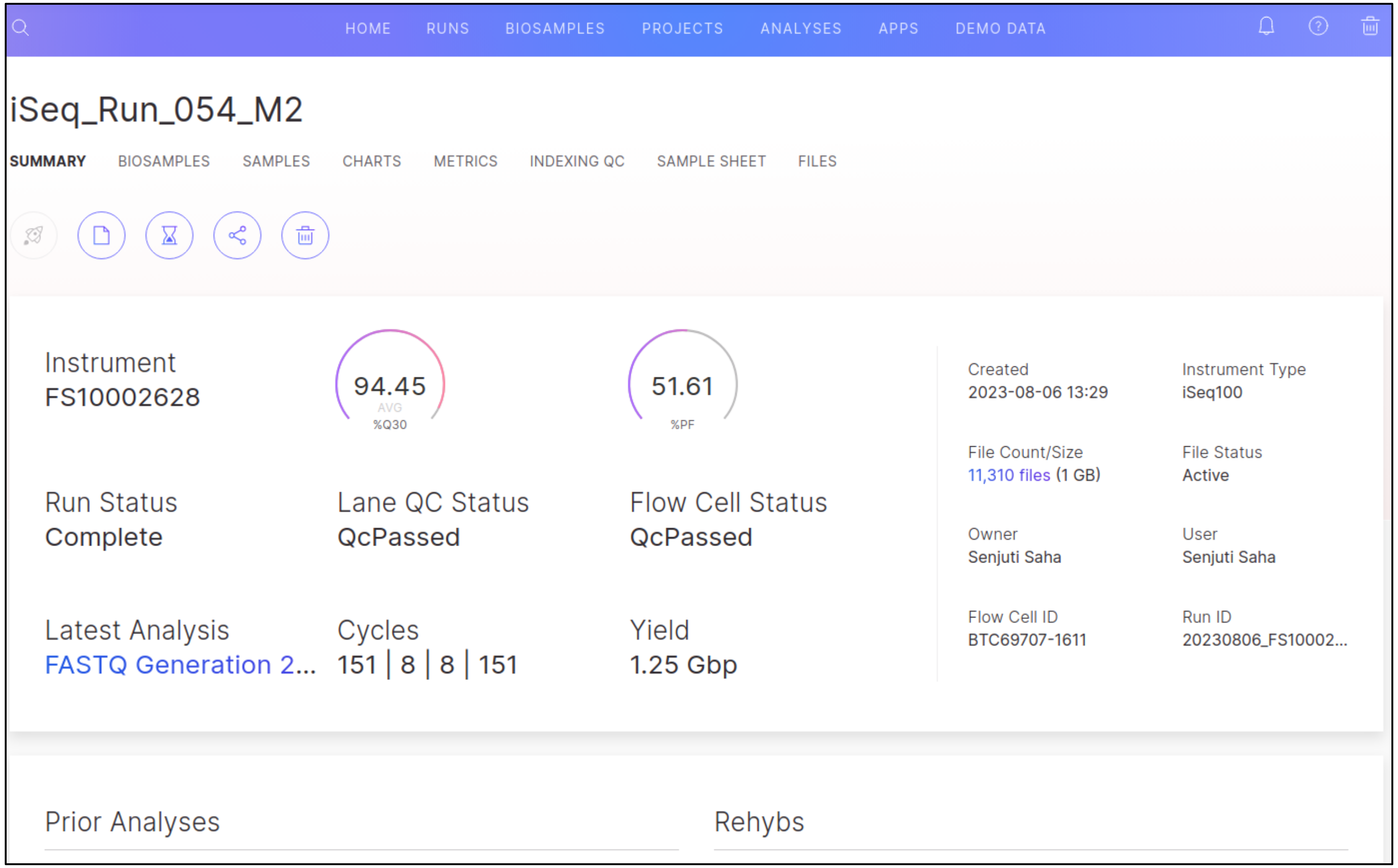
- **n_missing**: The number of positions in the consensus genome with no base called due to insufficient data or uncertainty in base calling.

- **n_gap**: The number of gap characters ("-") in the consensus genome,
  - indicating areas where the sequence is broken or unavailable.

- **n_ambiguous**: The number of ambiguous base calls (typically represented by "N") in the consensus genome,
  - indicating positions where the base could not be confidently

Basespace Tour

Basespace



iSeq_Run_054_M2

Instrument
FS10002628

94.45 AVG %Q30

51.61 %PF

Run Status
Complete

Lane QC Status
QcPassed

Flow Cell Status
QcPassed

Latest Analysis
FASTQ Generation 2...

Cycles
151 | 8 | 8 | 151

Yield
1.25 Gbp

Created
2023-08-06 13:29

Instrument Type
iSeq100

File Count/Size
11,310 files (1 GB)

File Status
Active

Owner
Senjuti Saha

User
Senjuti Saha

Flow Cell ID
BTC69707-1611

Run ID
20230806_FS10002...

Prior Analyses

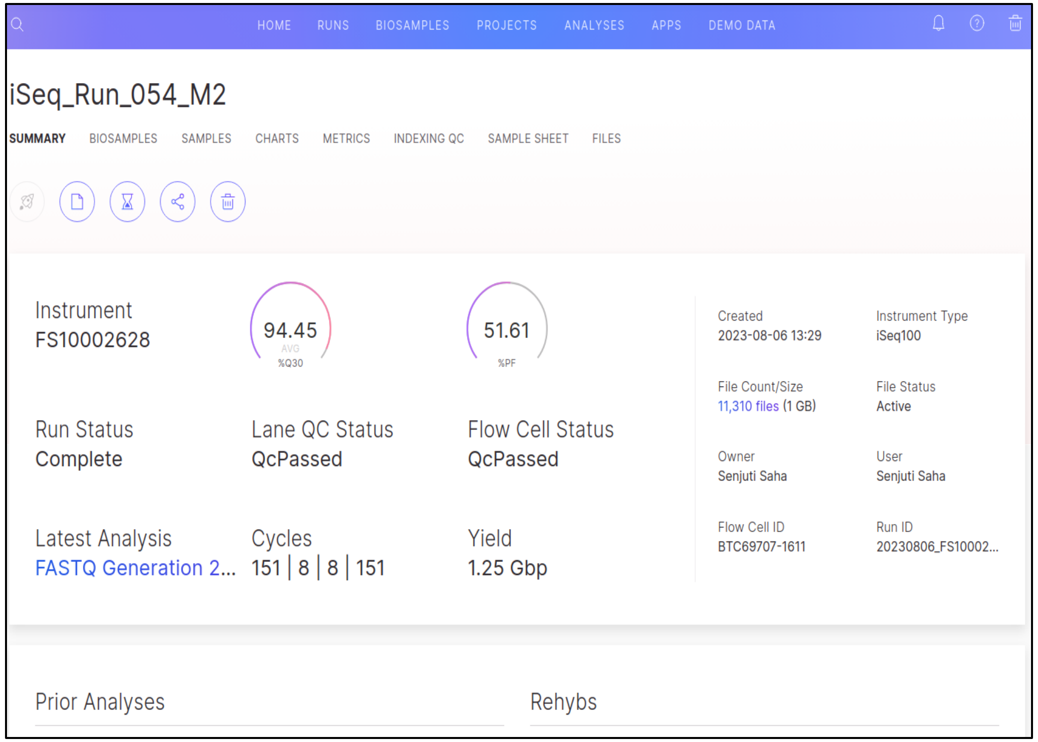
Rehybs

Basespace Summary

Instrument FS10002628: This is probably the **identifier** or **name** for the sequencing instrument used.

AVG 94.45: This could refer to an average quality score (AVG).

- A quality score of 94.45 is exceptionally high if we assume it's on the Phred scale which is commonly used in sequencing data quality assessments.



iSeq_Run_054_M2

Instrument
FS10002628

94.45
AVG
%Q30

51.61
%PF

Run Status
Complete

Lane QC Status
QcPassed

Flow Cell Status
QcPassed

Latest Analysis
FASTQ Generation 2...

Cycles
151 | 8 | 8 | 151

Yield
1.25 Gbp

Created
2023-08-06 13:29

Instrument Type
iSeq100

File Count/Size
11,310 files (1 GB)

File Status
Active

Owner
Senjuti Saha

User
Senjuti Saha

Flow Cell ID
BTC69707-1611

Run ID
20230806_FS10002...

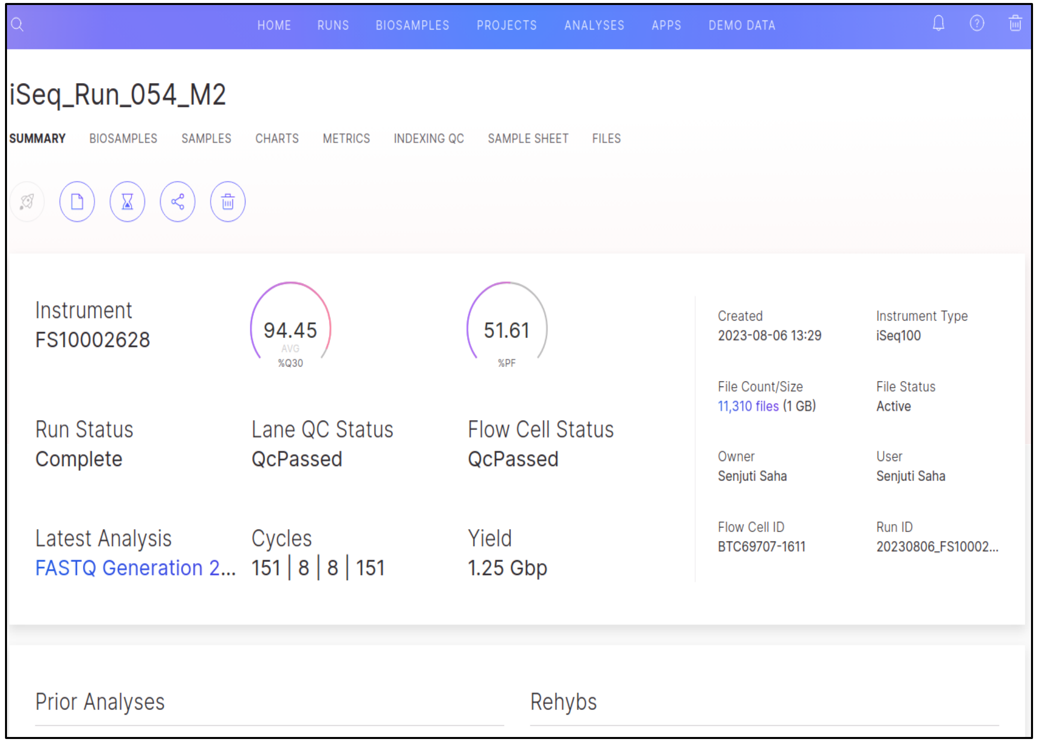
Prior Analyses

Rehybs

Basespace Summary

%Q30 51.61: This refers to the percentage of bases in the sequencing run that have a quality score of 30 or above (%Q30).

%PF: This might stand for "Percent Pass Filter," indicating the percentage of clusters passing the quality filter.



iSeq_Run_054_M2

Instrument
FS10002628

94.45
AVG
%Q30

51.61
%PF

Run Status
Complete

Lane QC Status
QcPassed

Flow Cell Status
QcPassed

Latest Analysis
FASTQ Generation 2...

Cycles
151 | 8 | 8 | 151

Yield
1.25 Gbp

Created
2023-08-06 13:29

Instrument Type
iSeq100

File Count/Size
11,310 files (1 GB)

File Status
Active

Owner
Senjuti Saha

User
Senjuti Saha

Flow Cell ID
BTC69707-1611

Run ID
20230806_FS10002...

Prior Analyses

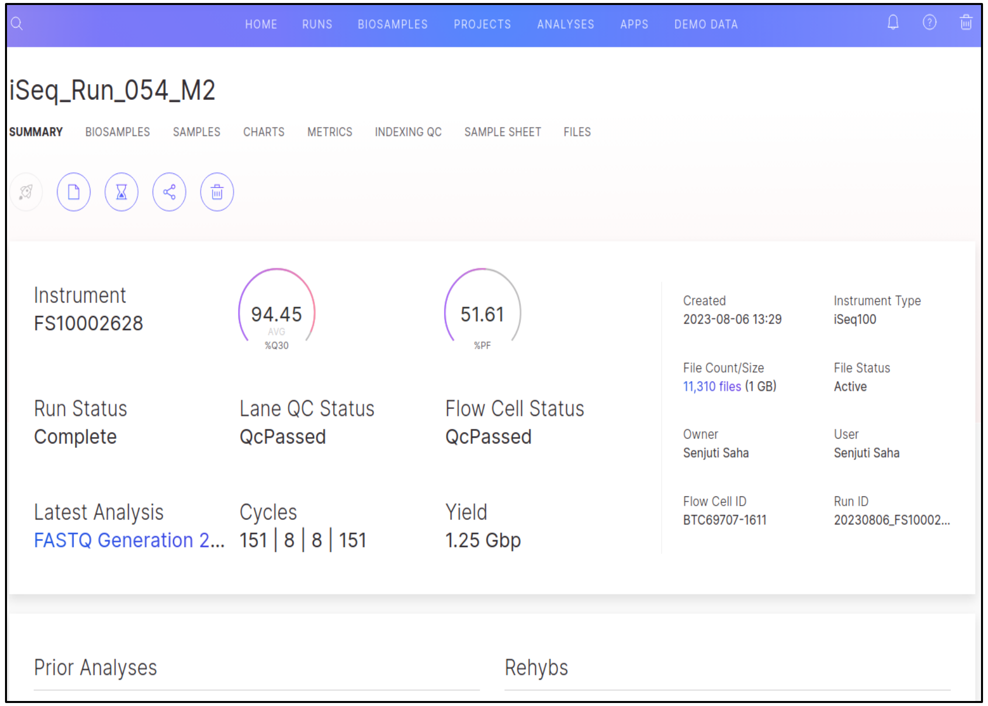
Rehybs

Basespace Summary

Lane QC Status & Flow Cell Status QcPassed: (QC) checks have been passed for both the lane and the entire flow cell.

Cycles 151 | 8 | 8 | 151: Describes the read length configuration for the sequencing run. It's often formatted as [Read 1] | [Index 1] | [Index 2] | [Read 2] for paired-end runs.

Flow Cell ID BTC69707-1611 & **Run ID 20230806_FS10002628_1_BTC69707-1611**: Unique identifiers for the sequencing run and flow cell, which can be vital for tracking and data management.



iSeq_Run_054_M2

Instrument
FS10002628

Run Status
Complete

Lane QC Status
QcPassed

Flow Cell Status
QcPassed

Latest Analysis
FASTQ Generation 2...

Cycles
151 | 8 | 8 | 151

Yield
1.25 Gbp

Created
2023-08-06 13:29

Instrument Type
iSeq100

File Count/Size
11,310 files (1 GB)

File Status
Active

Owner
Senjuti Saha

User
Senjuti Saha

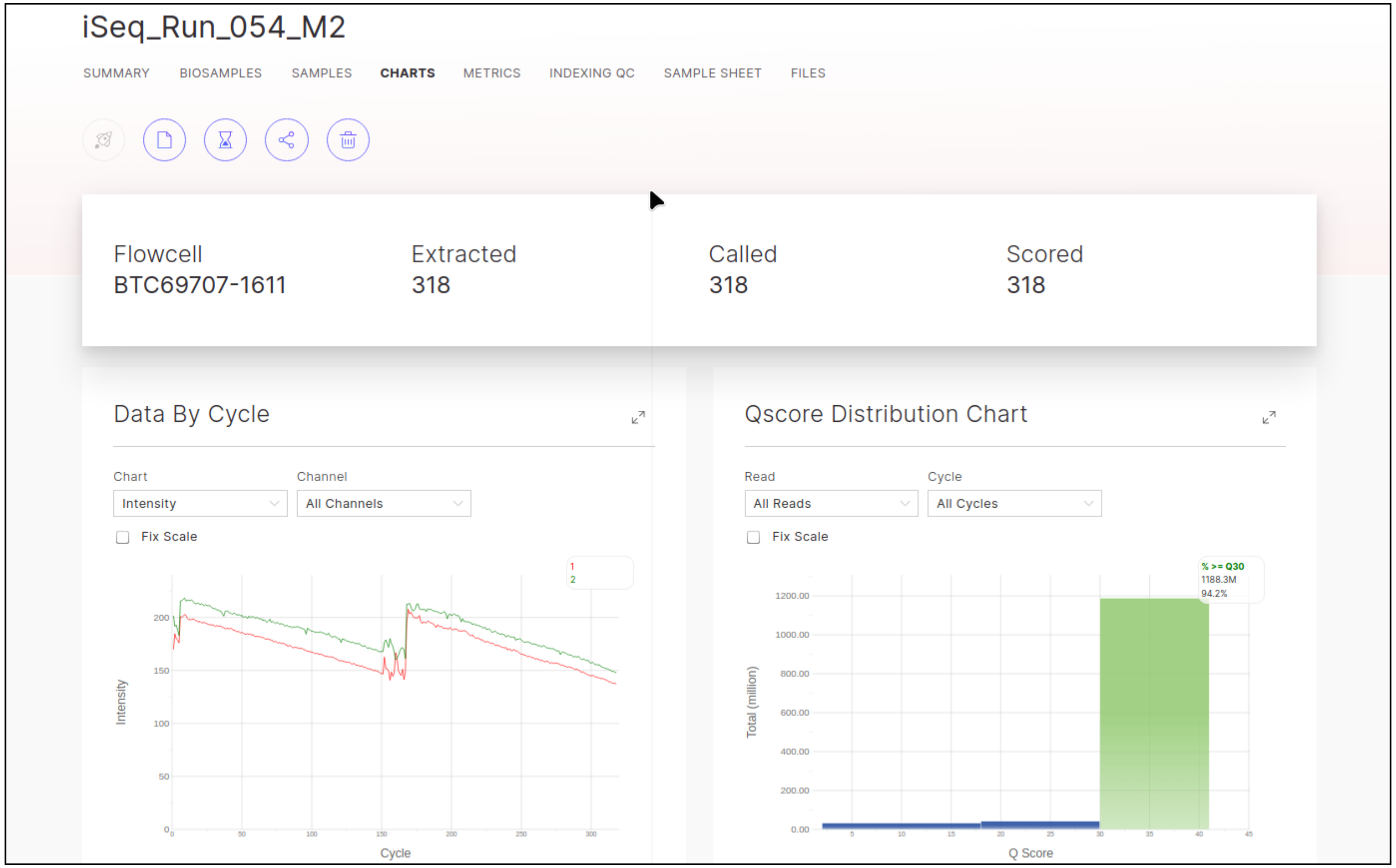
Flow Cell ID
BTC69707-1611

Run ID
20230806_FS10002...

Prior Analyses

Rehybs

Basespace CHARTS Intensity



iSeq_Run_054_M2

Flowcell
BTC69707-1611

Extracted
318

Called
318

Scored
318

Data By Cycle

Chart: Intensity

Channel: All Channels

□ Fix Scale

Qscore Distribution Chart

Read: All Reads

Cycle: All Cycles

□ Fix Scale

Child Health Research Foundation

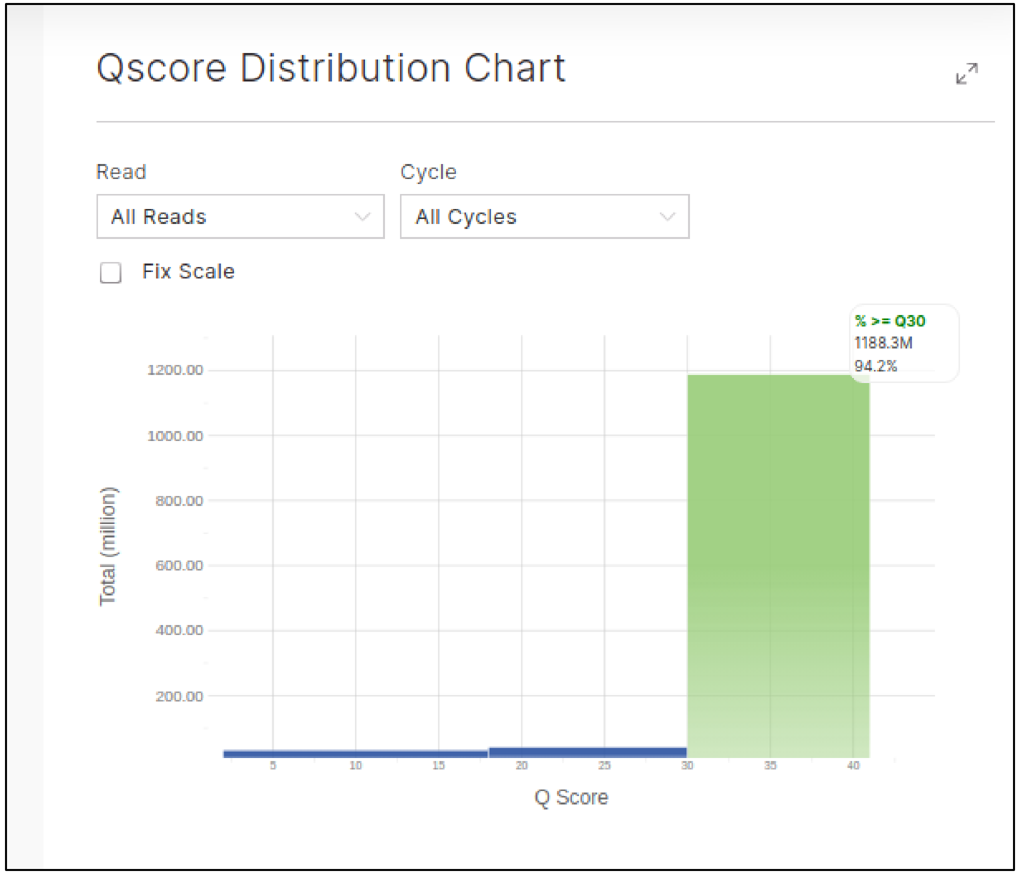
CHRF

Prevent Infections, Save Lives

Basespace CHARTS Intensity

Qscore Distribution Chart

- **Total (million): 5 to 45** and **Q Score: % >= Q30: 1188.3M, 94.2%** — This suggests a visualization of quality scores (Qscore) across reads. 94.2% of the 1188.3 million bases have a Qscore of at least 30, which implies a high level of accuracy.



Qscore Distribution Chart

Read: All Reads

Cycle: All Cycles

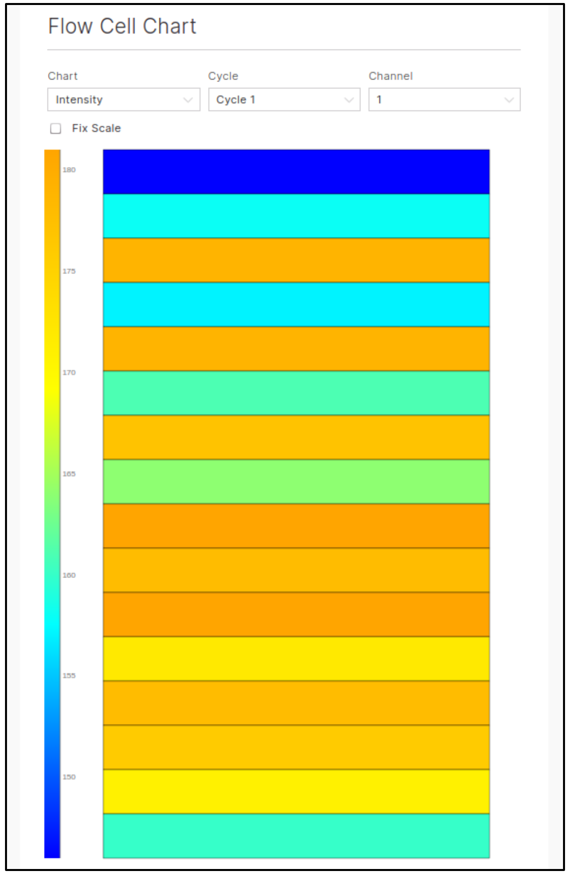
□ Fix Scale

% >= Q30
1188.3M
94.2%

Basespace CHARTS Intensity

Flow Cell Chart

- **Intensity** and **Cycle 1, Channel 1** — This could be representing a visual data showing the fluorescence intensity from channel 1 during the first cycle. The "Fix Scale" might be a user-defined limit to enhance visualization of the data.



Flow Cell Chart

Chart: Intensity

Cycle: Cycle 1

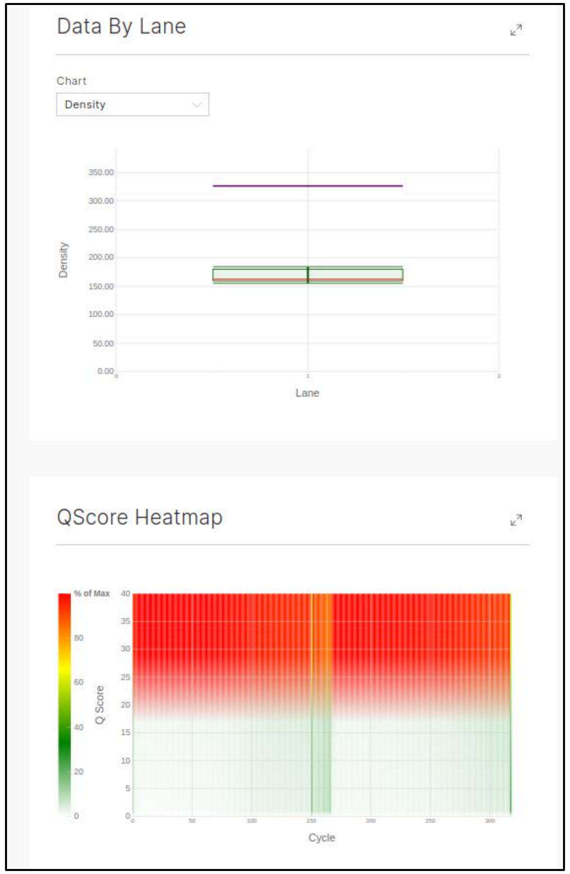
Channel: 1

□ Fix Scale

Basespace CHARTS Intensity

Data By Lane

- **Density: 0.00 to 350.00** — This might refer to a graph showing the cluster density per lane on the flow cell, which could help identify issues like under- or over-clustering.

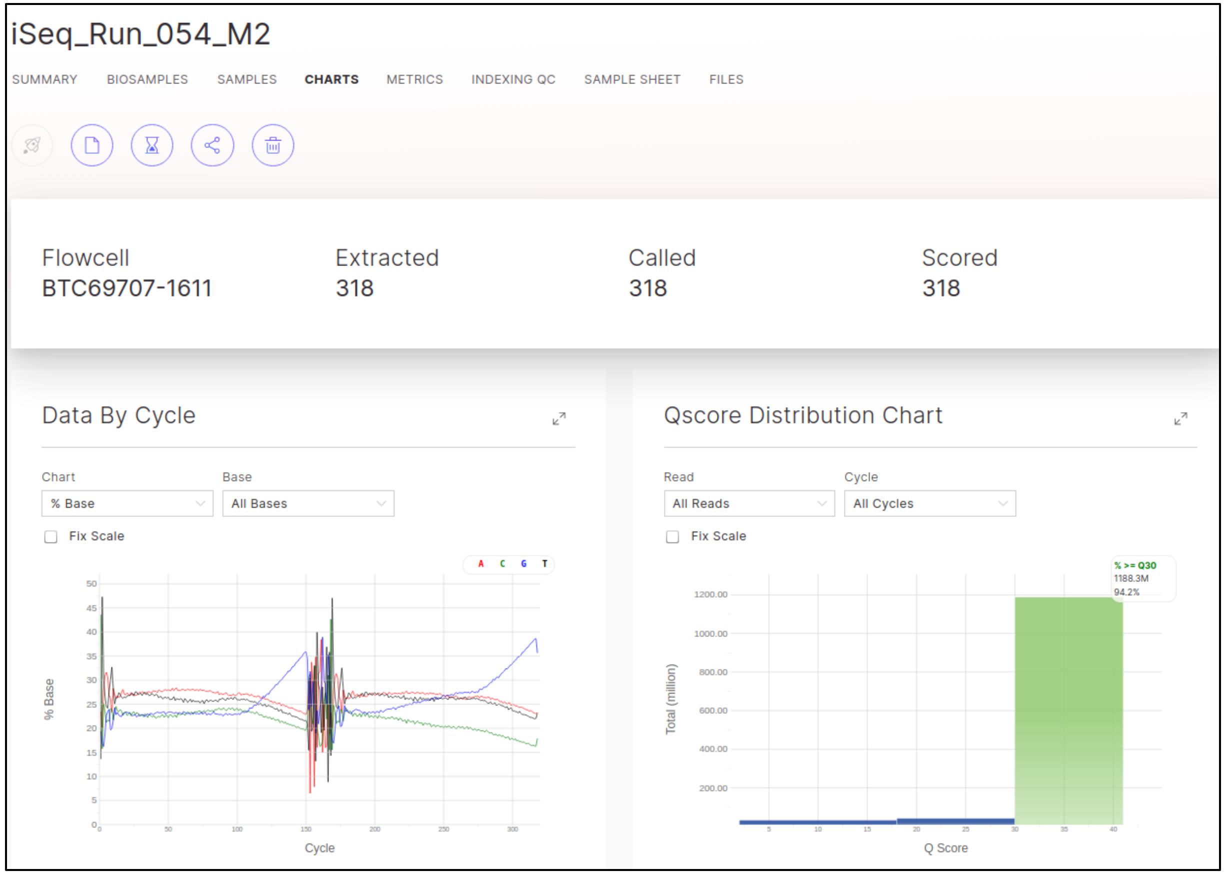


Data By Lane

Chart: Density

QScore Heatmap

Basespace CHARTS Base



iSeq_Run_054_M2

Flowcell BTC69707-1611

Extracted 318

Called 318

Scored 318

Data By Cycle

Chart: % Base

Base: All Bases

□ Fix Scale

Qscore Distribution Chart

Read: All Reads

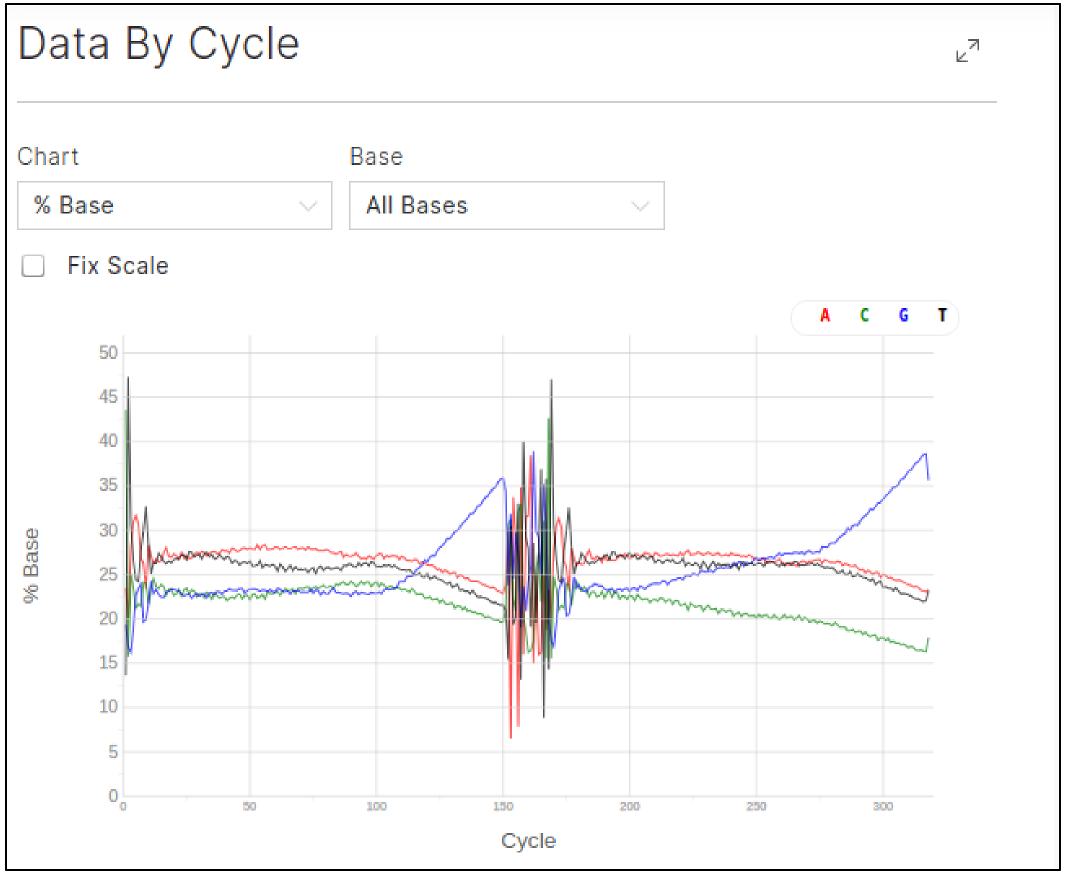
Cycle: All Cycles

□ Fix Scale

Basespace CHARTS Base

- **% Base**: This refers to the percentage of each nucleotide (A, C, G, T) at a specific position in the sequencing reads.

- **Cycle**: This refers to the sequencing cycles, each cycle corresponds to the incorporation of one nucleotide in the sequencing process.



Data By Cycle

Chart: % Base

Base: All Bases

□ Fix Scale

Thank You !