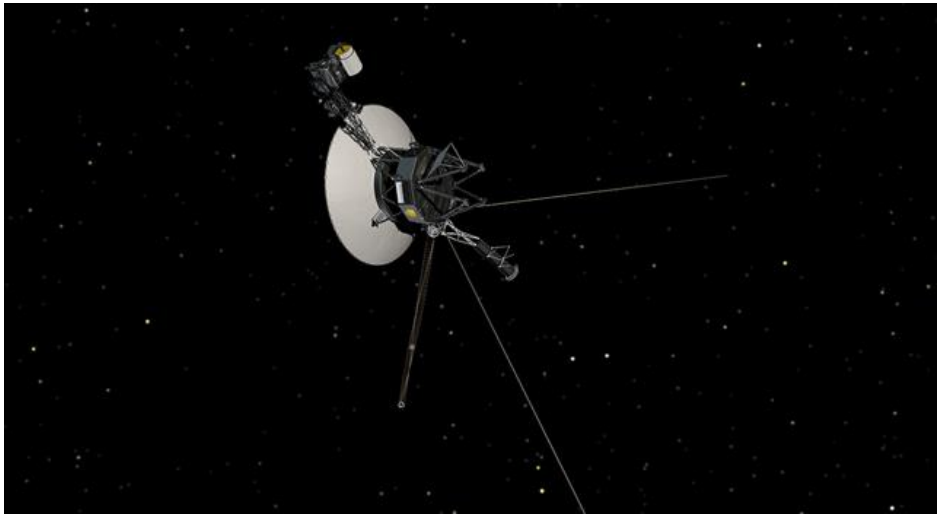
RTA 2.0 to 3.0 Changes

Session Objectives

- **By the end of this session you will**
  - Understand why a new version of RTA was needed
  - Be able to describe the differences between RTA2 and RTA 3
  - Know how Q-Scores are assigned on the NovaSeq
  - Be aware of software that is compatible with RTA3 outputs

Why Create Another RTA?



# of pixels to process are roughly the distance to pluto in feet

+

Processing Speed

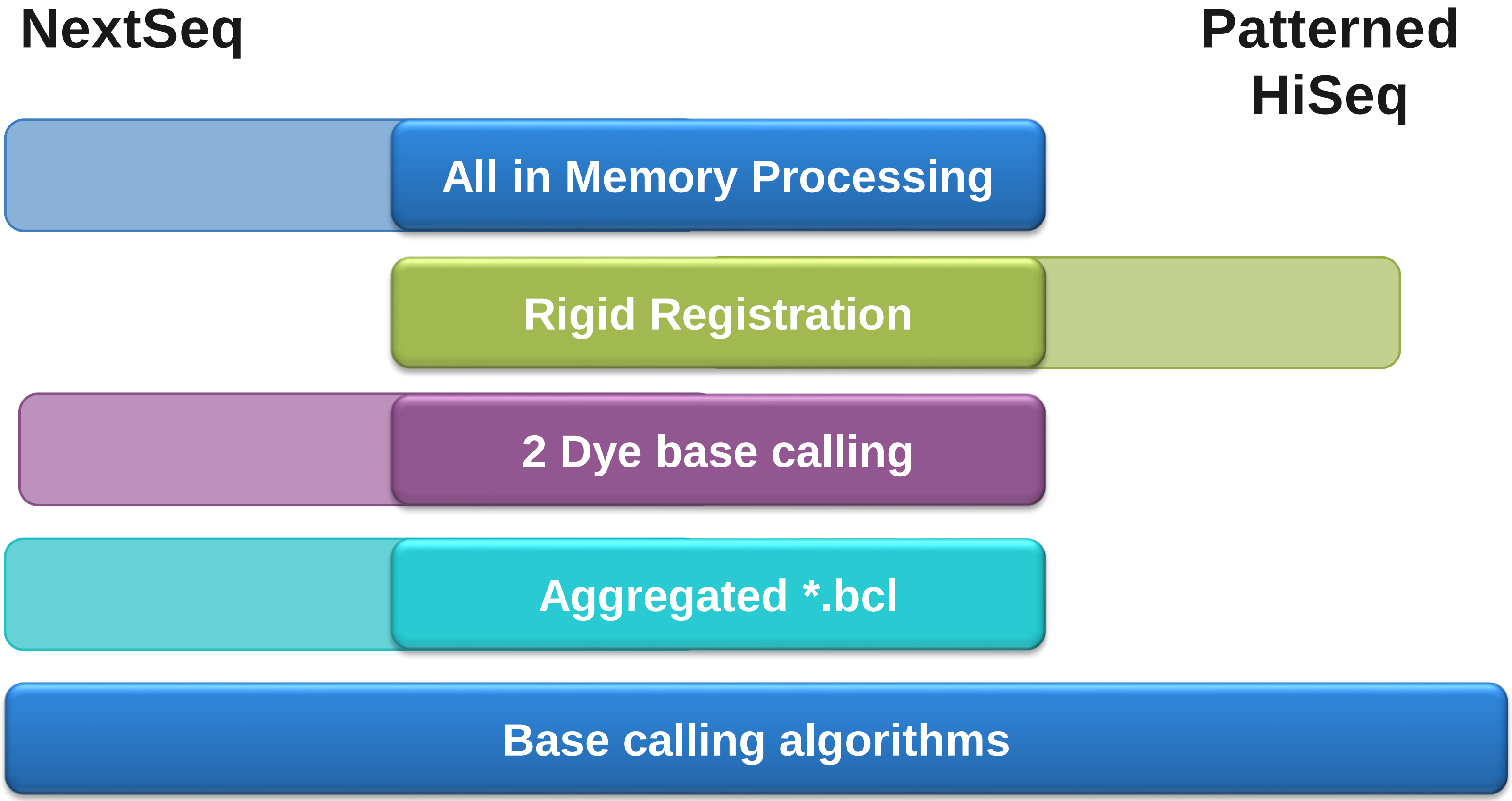
Run time 44 hours

(158,400 pixels per second)

=

2.5x increase in speed required

RTA3 and RTA 2: What Is The Same?



NextSeq

Patterned HiSeq

All in Memory Processing

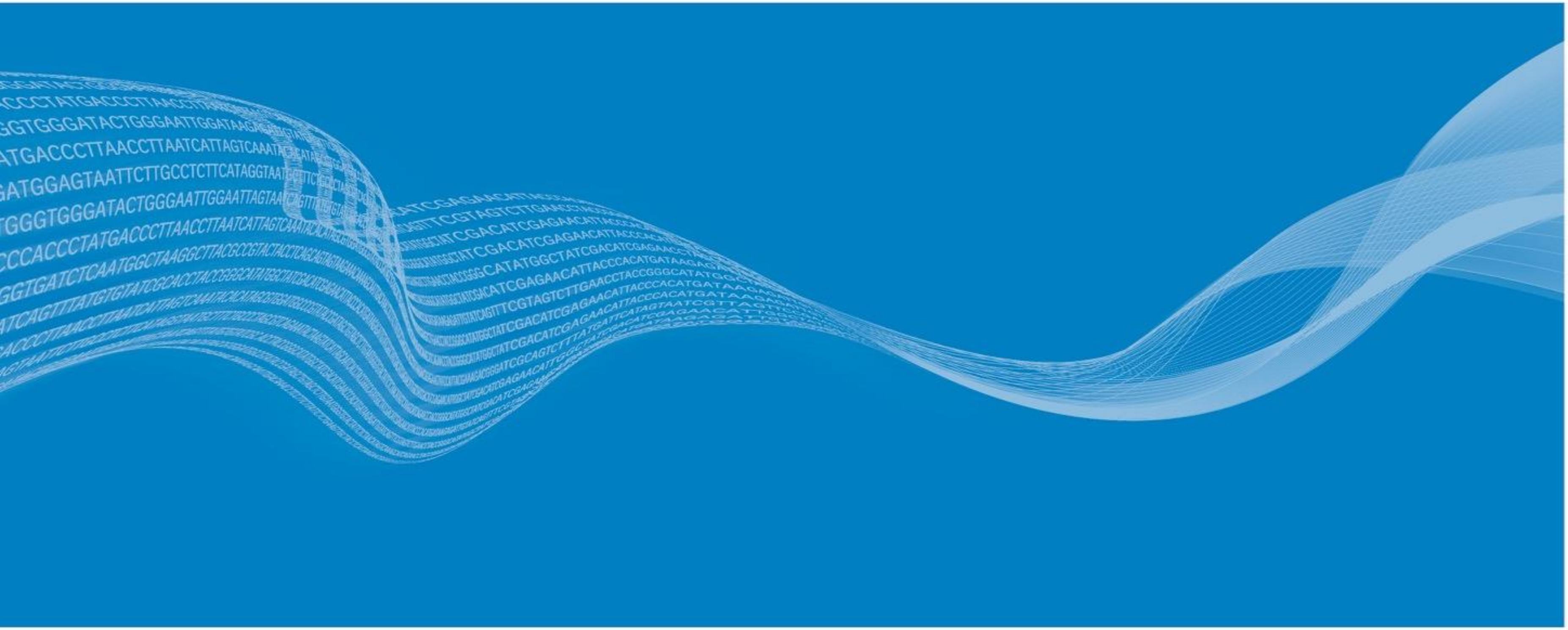
Rigid Registration

2 Dye base calling

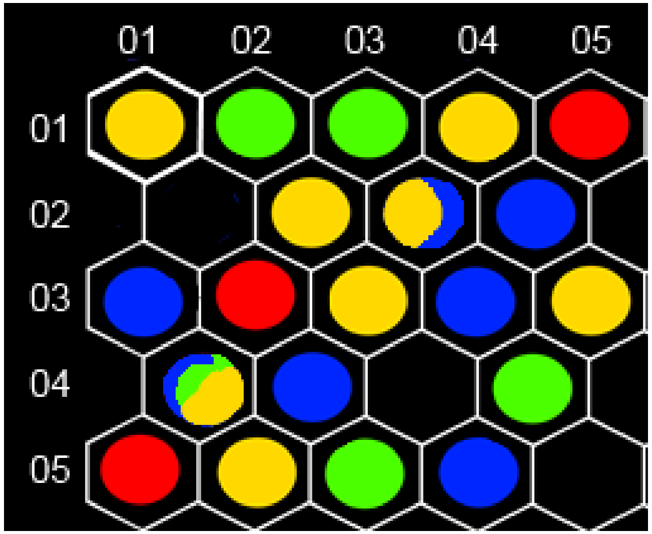
Aggregated *.bcl

Base calling algorithms

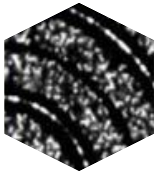
Review of Algorithms Shared Between RTA2 and RTA3



Rigid Registration For Patterned Flow Cells



Preset hexagonal lattice of cluster locations is aligned to the images

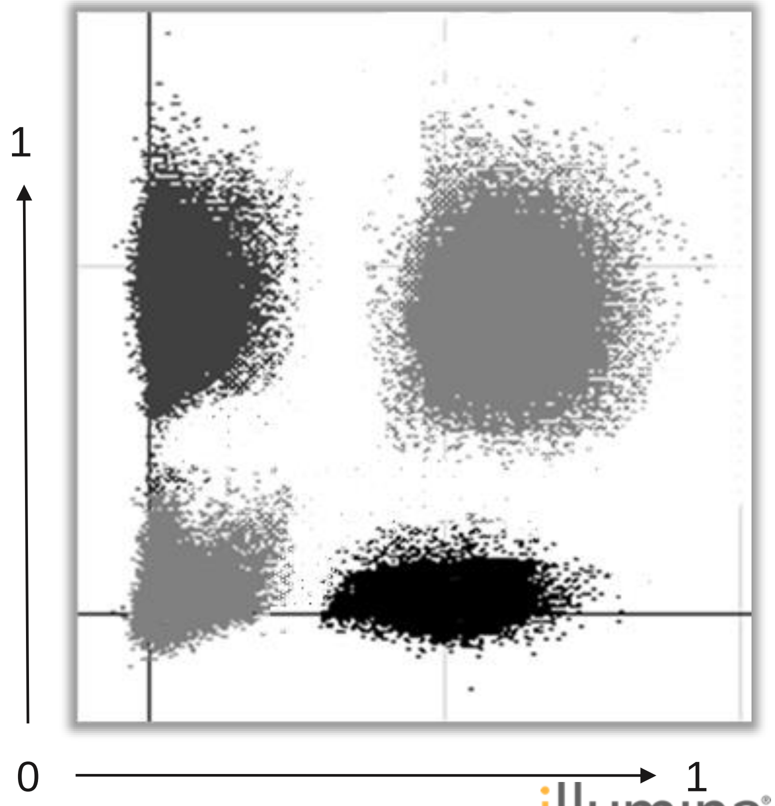


Circular fiducials are used in aligning lattice to flow cell

2 Color Base Calling Normalization

Scale all intensities so their P05 and P95 intensities represent 0 and 1

	Background subtracted, Spatial Normalized Intensities	**Base call Normalized Intensities**
	99	1
P95 =1	98	1
	97	0.99
	85	0.87
	84	0.86
	79	0.89
	76	0.76
	71	0.72
	63	0.69
	62	0.63
	61	0.62
	50	0.51
	48	0.49
	25	0.25
	22	0.22
	20	0.20
	15	0.15
	13	0.13
P05 =0	10	0
	3	0

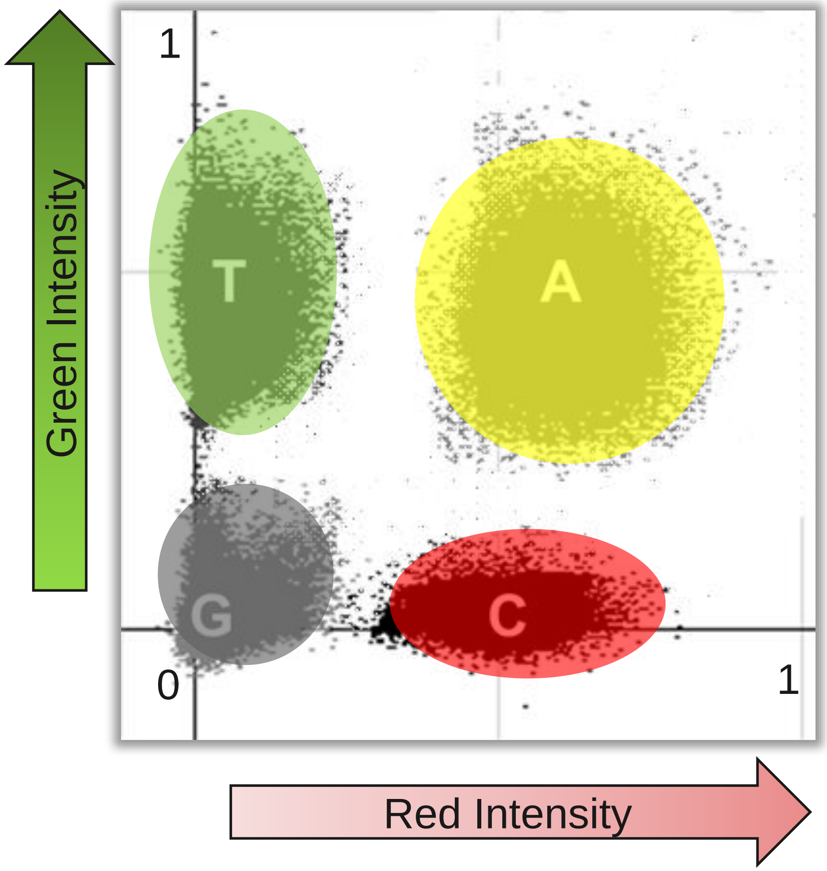


1

2 Color Population-based Base Calling

- **Scatterplot of 4 distinct populations (nucleotides) is created from extracting intensities from one image versus the other image**

- **Base calls are made according to which channel is on (1) or off (0) for each cluster according to (x, y):**
  - (1, 0) → C
  - (0, 1) → T
  - (1, 1) → A
  - (0, 0) → G

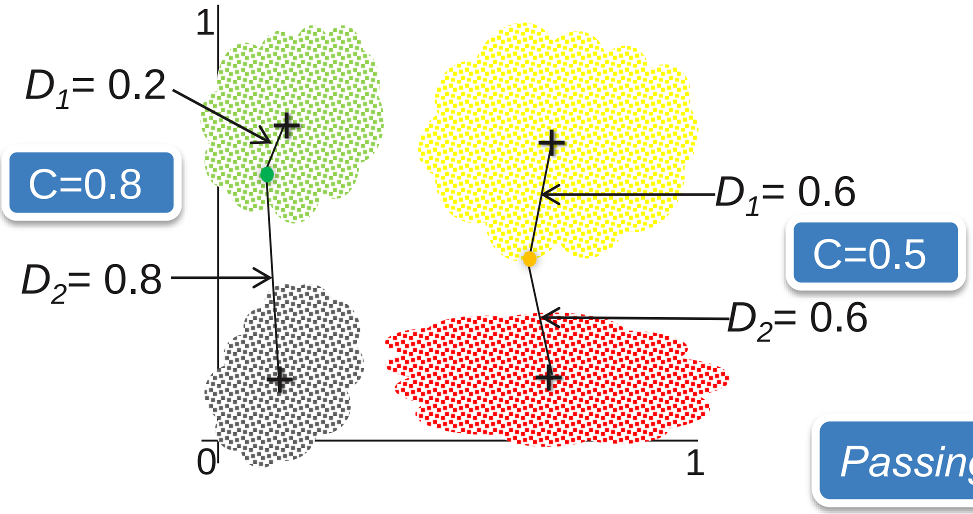


2-Color Calculating Clusters Passing Filter

Pass filter is:

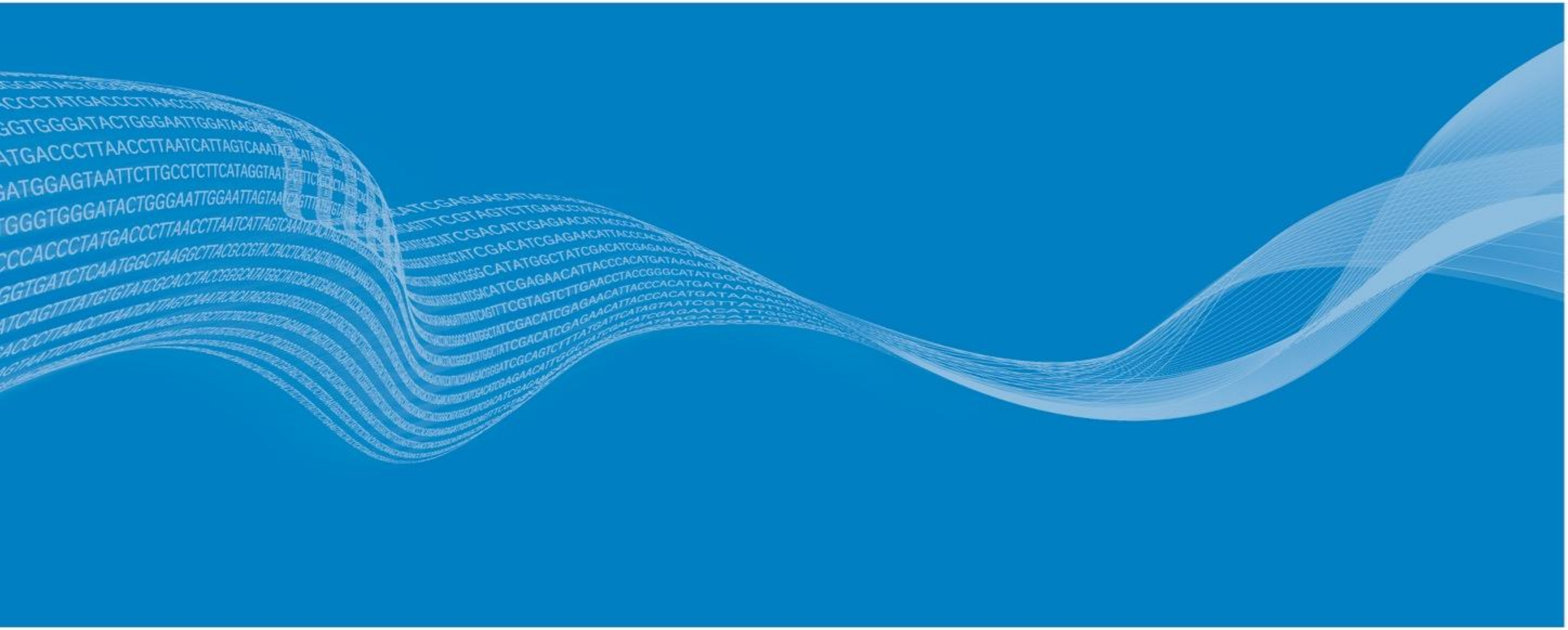
$$C = 1 - \frac{D_1}{D_1 + D_2}$$

- The ratio of the sum of the most prominent and second most prominent population intensities
- Calculated for each cluster over the first 25 bases of the sequence
- Filters cluster by signal purity
  - Removes overlapping and low-intensity clusters

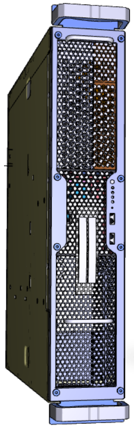


Passing Chastity value: ≥ 0.63

Introducing RTA3



What’s New – RTA 3

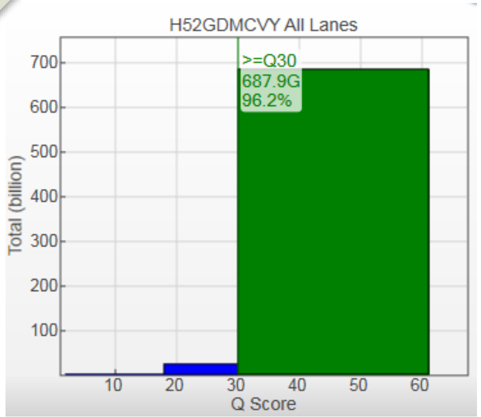


Distributed Compute Architecture

Folder and File structure differences

Compute Framework

Quality Score Reporting



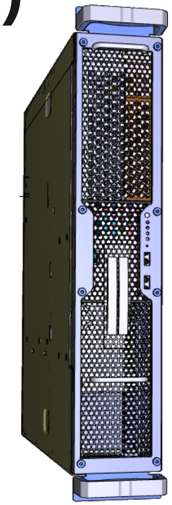
Distributed Compute Architecture

Single Board Computer (SBC)

- Windows 10
- Responsible for:
  - User Interface (hardware and software)
  - UCS (New Run Copy Service)
  - NovaSeq Control Software
  - Storage of logs

Compute Engine (CE)

- Powerful Linux Box
- Responsible for
  - RTA 3
  - Temp run folder



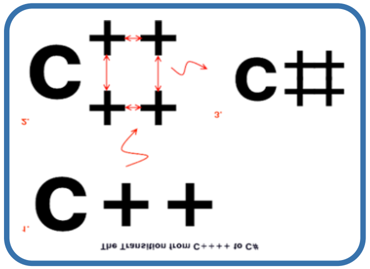
Folder and File Structure Differences

Folder and File structure differences

- **Run Folder Structure**
  - More efficient base calling format
  - 2 base calls per byte before zipping
- ***.CBCL format (Concatenated)**
  - Nonpassing filter clusters removed after cycle 25
  - Dramatically smaller through compression
  - Aggregated by surface and lane
- **InterOp Folder Format**
  - Per Cycle InterOp Files
  - Open-source library to parse new InterOp:
    - github.com/Illumina/interop

Compute Framework Changes

Compute Framework



Written fully in C++

- Mix of C++ and C# converted to C++
- One language results in better CPU utilization

Vectorization

- Execute the same task on multiple values simultaneously
- “Eat 1 candy vs. Eat all the candies”

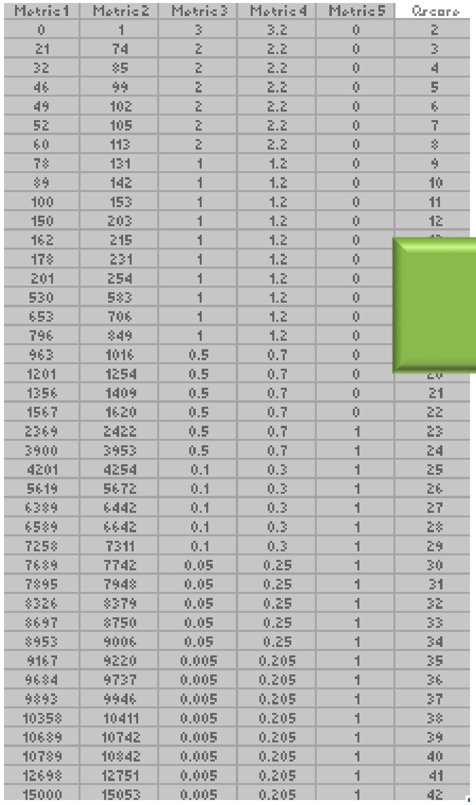
Data “Traffic Flow” Optimizations

- Any tile can be worked on by any thread
- Every tile owns its own cache

Historic Q-Score Generation And Binning

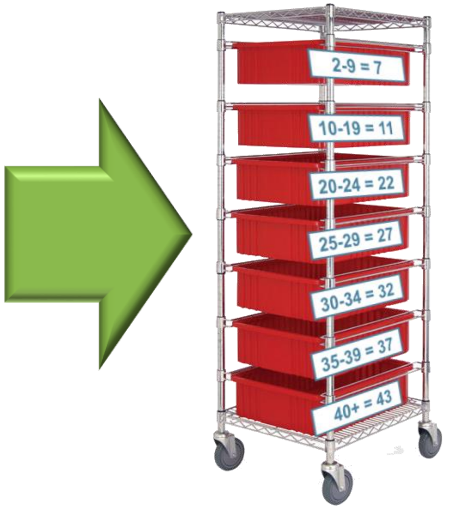
Quality Score Reporting

LookUp Table



Metric1	Metric2	Metric3	Metric4	Metric5	Qscore
0	1	3	3.2	0	2
21	74	2	2.2	0	3
32	85	2	2.2	0	4
46	99	2	2.2	0	5
49	102	2	2.2	0	6
52	105	2	2.2	0	7
60	113	2	2.2	0	8
78	131	1	1.2	0	9
89	142	1	1.2	0	10
100	153	1	1.2	0	11
150	203	1	1.2	0	12
162	215	1	1.2	0	13
178	231	1	1.2	0	[illegible]
201	254	1	1.2	0	[illegible]
530	583	1	1.2	0	[illegible]
653	706	1	1.2	0	[illegible]
796	849	1	1.2	0	[illegible]
963	1016	0.5	0.7	0	[illegible]
1201	1254	0.5	0.7	0	20
1356	1409	0.5	0.7	0	21
1567	1620	0.5	0.7	0	22
2369	2422	0.5	0.7	1	23
3900	3953	0.5	0.7	1	24
4201	4254	0.1	0.3	1	25
5619	5672	0.1	0.3	1	26
6389	6442	0.1	0.3	1	27
6589	6642	0.1	0.3	1	28
7258	7311	0.1	0.3	1	29
7689	7742	0.05	0.25	1	30
7895	7948	0.05	0.25	1	31
8326	8379	0.05	0.25	1	32
8697	8750	0.05	0.25	1	33
8953	9006	0.05	0.25	1	34
9167	9220	0.005	0.205	1	35
9684	9737	0.005	0.205	1	36
9893	9946	0.005	0.205	1	37
10358	10411	0.005	0.205	1	38
10689	10742	0.005	0.205	1	39
10789	10842	0.005	0.205	1	40
12698	12751	0.005	0.205	1	41
15000	15053	0.005	0.205	1	42

Binned Q-scores



Binning reduces data footprint, however the large lookup table is a processing bottleneck

RTA3 Outputs Four Quality Scores

Quality Score Reporting

Simplified Q-Score Assignment

Q Score	Probability of Incorrect Base	Base Call Accuracy
2	Qscore not assigned	
12	6.3 in 100	~94%
23	5 in 1,000	~99.5%
37	2 in 10,000	~99.98%

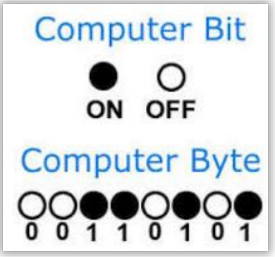
Actual Q-scores subject to change

We will discuss how Q-Scores are assigned in more details in subsequent slides

Fewer reported quality scores reduce data footprint

Why Only Four Quality Scores?

A Little Bit Of Computer Science



- **Smaller decimals require fewer bits to store in binary**
  - Bit is short for "binary digit"
  - 8 bits per byte

- **RTA3 *.CBCL files Math:**

  2 bits to store each base

  <u>+ 2 bits to store its Q-score</u>

  4 bits for each base in a *.CBCL

  (two bases per byte)

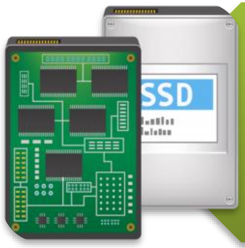
	Decimal	Translated to Binary	
RTA3	0	0	2 bits per Q-score
	1	1	
	2	10	
	3	11	
Q-score Binning	4	100	3 bits per Q-score
	5	101	
	6	110	
	7	111	
Non-binned Q-scores	8	1000	7 bits per Q-score
	16	10000	
	32	100000	
	64	1000000	

Quality Score Reporting Advantages

Quality Score Reporting

Time

- Smaller lookup table = faster lookup



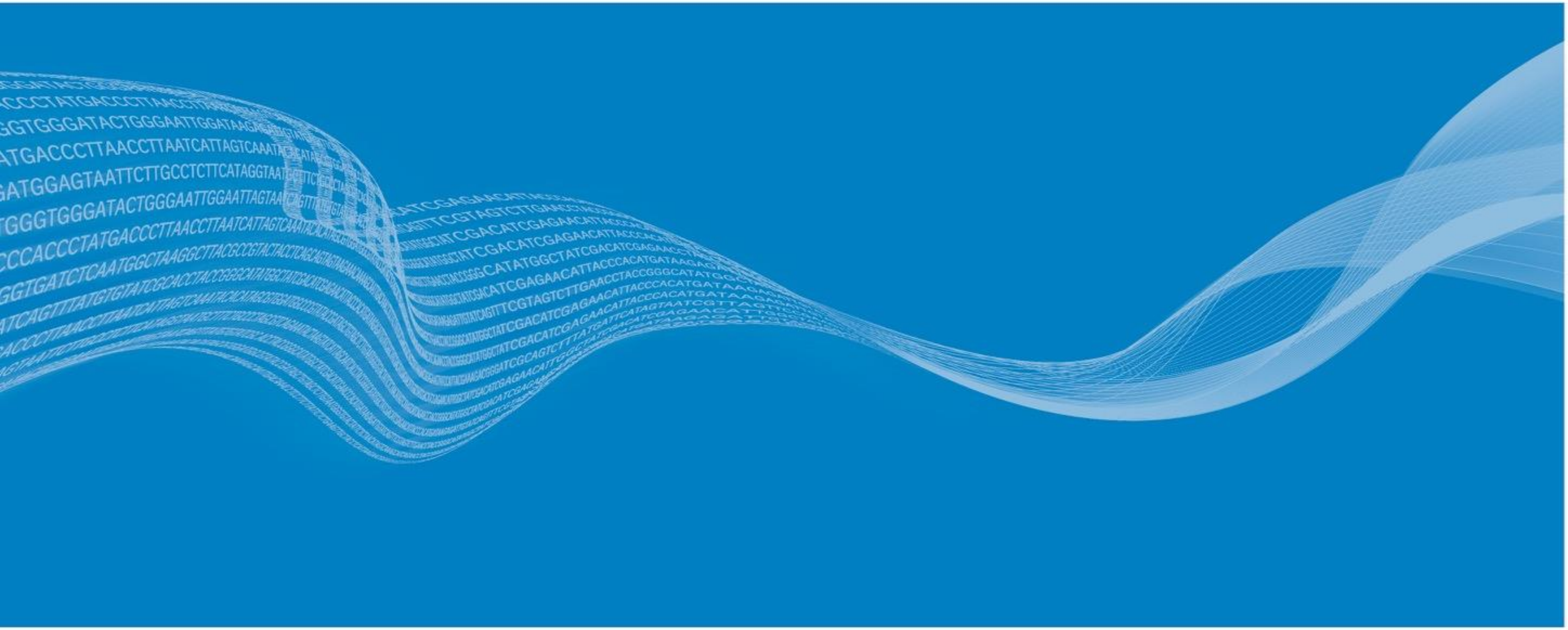
Disk Space

- 4 scores = reduced data footprint

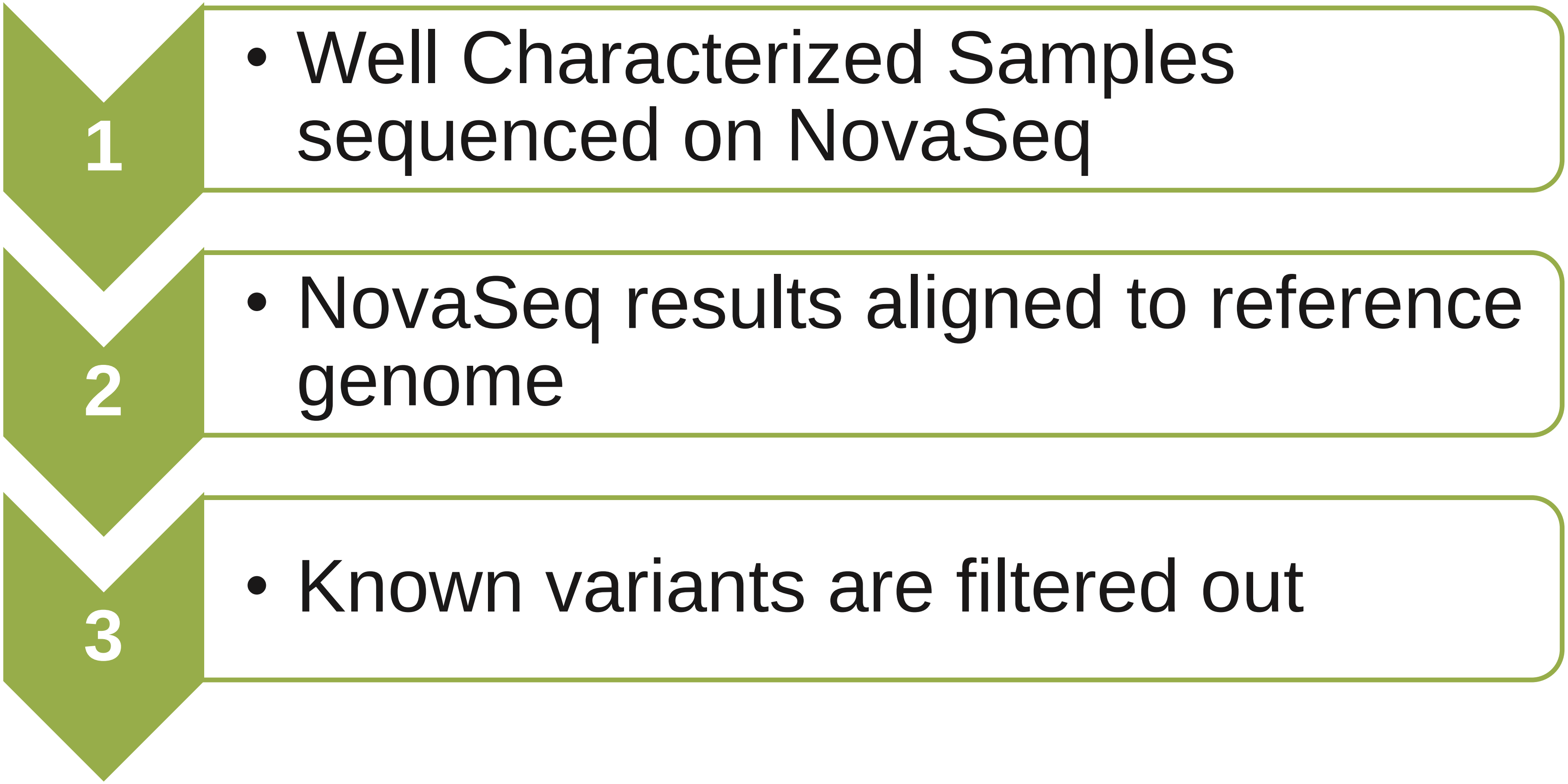
Data Transfer

- Reduced data footprint = reduces bandwidth required compared to what it would have been

Q-Score Assignment on NovaSeq



How Illumina Generated Data to Train NovaSeq Q-Tables



1

- Well Characterized Samples sequenced on NovaSeq

2

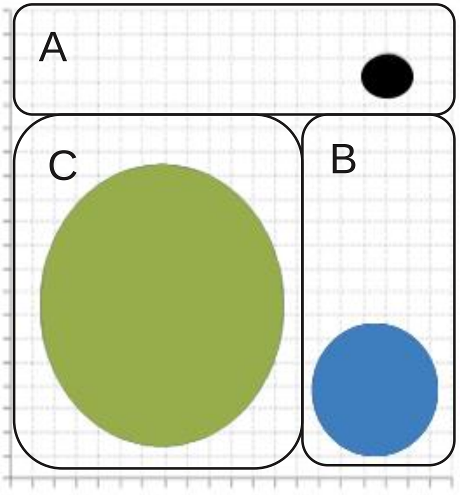
- NovaSeq results aligned to reference genome

3

- Known variants are filtered out

How Illumina Trained NovaSeq Q-Tables

Multiple features predictive of quality plotted



Phred-Scale Quality Scores are Logarithmic

Group	Error Rate	Q-Score
A	6.3 in 100	12
B	5 in 1,000	23
C	2 in 10,000	37

Actual Q-scores reported are subject to change

4

Basecalls are divided into 3 groups based on predictive features

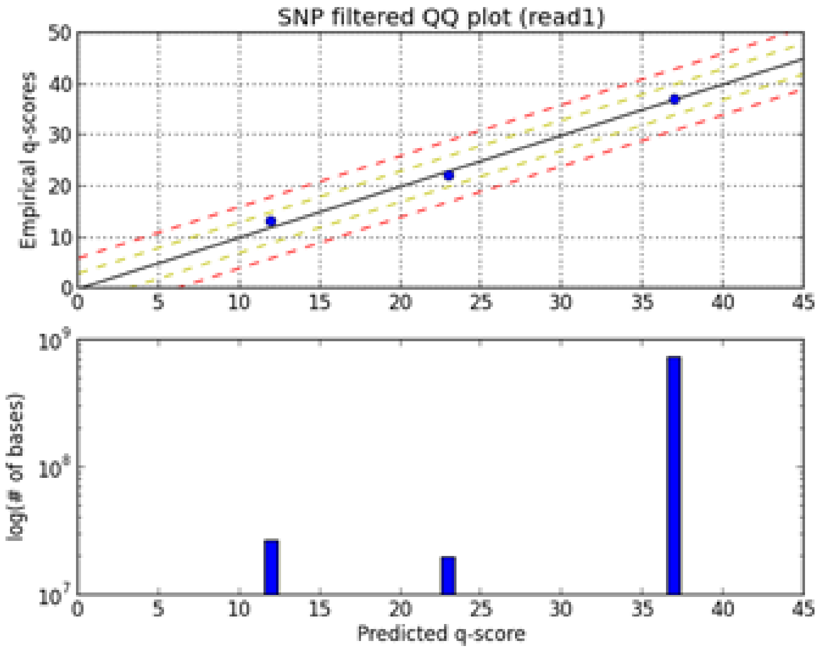
5

Quality score assigned based on group’s empirical error rate

How Q-Tables Provides Quality Prediction

Quality Scores are assigned according to which group the data behaves like most

Feature Behavior Similar to Group:	Q Score Assigned
A	37
B	23
C	12
No Call Assigned	2

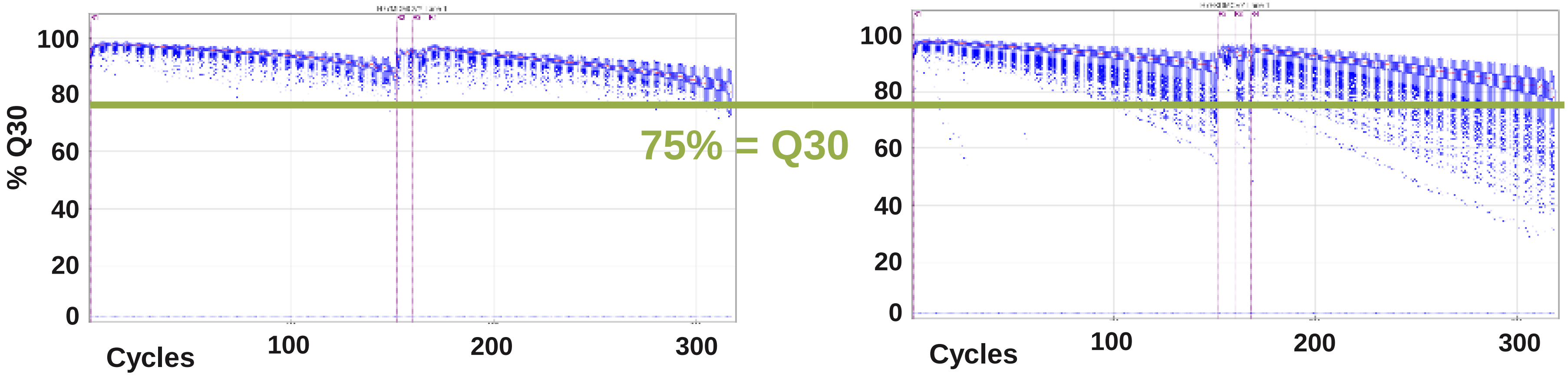


Comparing the empirical Q-Score to the predicted Q-Score in new samples show the tables are well trained

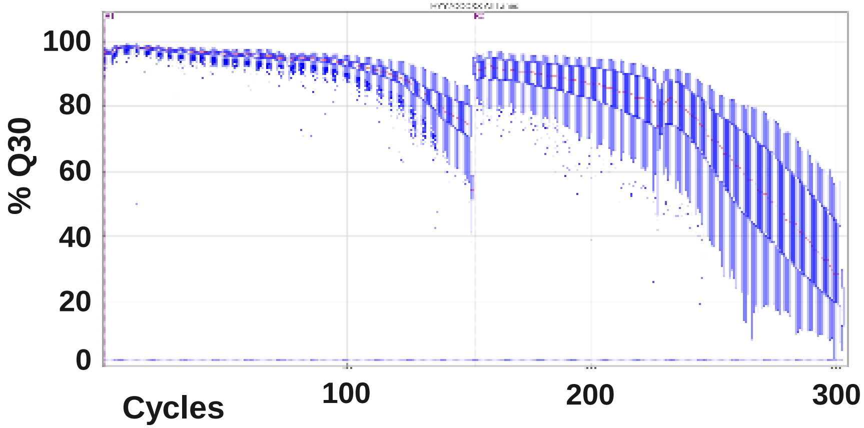
Platform Comparison %Q30 by Cycle

NovaSeq Example 1 (2*151+23 cycles)

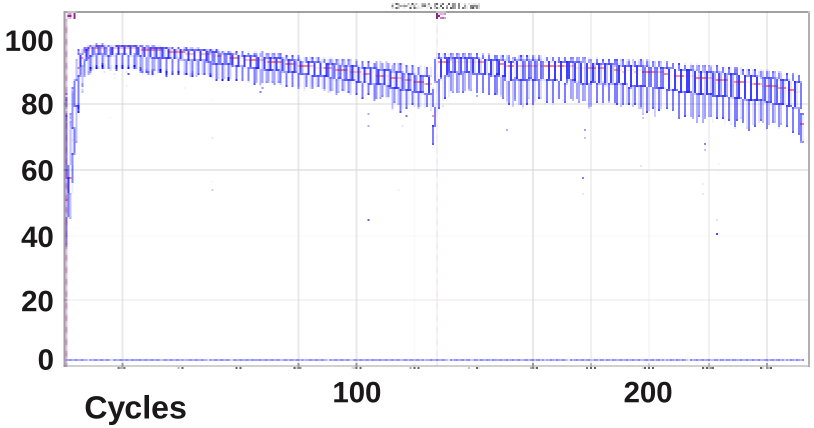
NovaSeq Example 2 (2*151+23 cycles)



HiSeq X (2*151 cycles)



HS2000 v4 (2*126 cycles)



Note: Runs use an updated, although not final, Q-table which may affect the accuracy of the quality scores.

Waterfall in % Q30 Data By Cycle

Jumps between Q Score groups are clearly separated

- Visual artifact thought to be caused by groups of tiles shifting together

More tile based features used in NovaSeq

- Previous Q tables used more cluster-based features which resulted in smoother plots

Comparing HiSeq X and NovaSeq data

- Shows comparable human genome build quality
- Suggests this is a cosmetic issue, not a data quality issue

Advice From Illumina’s Data Analysis Experts

Visual artifact makes the % ≥Q30 per cycle plot less informative

Q20 per cycle plots correlate better with error rate

Overall %≥Q30, Q20 per cycle, and error rate are better measures of data quality

Bioinformatics Details - Quality Scores

Data set comparisons show extremely high correlation between down stream analysis regardless of how this plot looks

Chr20	8 Q-score (HiSeq X)	4 Q-score (NovaSeq)
Total variant positions	100,795	100,875
In Platinum regions	83,659	83,669
In Platinum regions and PASSes FILTER	82,473	82,442
In Platinum regions and PASSes FILTER and not in other vcf	361	371
In Platinum regions and PASSes FILTER only in 8score/4Qscore	184	216

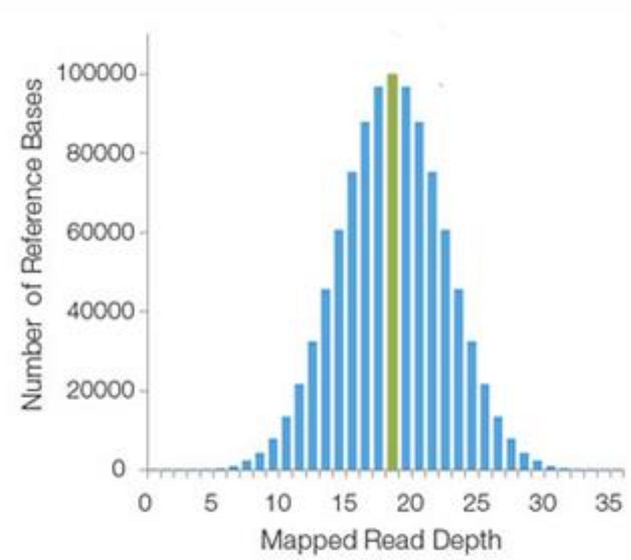
Human Genome Performance on NovaSeq

Genome build quality highly concordant with HiSeq

	NovaSeq (n4)	HiSeq X (n2)	HiSeq v4 (n2)	NextSeq (n2)
Genome Coverage (x)	**30.6**	**30.5**	**29.8**	**30.1**
Autosome Coverage	**95%**	**95%**	**91%**	**94%**
Autosome Callability	**95%**	**95%**	**93%**	**93%**
Autosome Exon Callability	**98%**	**98%**	**91%**	**95%**
SNV Precision	**100%**	**100%**	**100%**	**100%**
SNV Recall	**97%**	**97%**	**96%**	**96%**
Indel Precision	**97%**	**98%**	**97%**	**96%**
Indel Recall	**95%**	**95%**	**88%**	**88%**

NovaSeq Prototype Instruments running S2 flow cell

Coverage And Callability Defined



Coverage

- Better defined as the mean mapped read depth
- Sum of mapped read depths divided by the number of known (sequence-able) bases in the reference

Callability

- Can the genotype be definitively determined at a specified confidence threshold after multiple filters (such as read depth and Q Score) have been applied

Callable States:

-Did the base have enough coverage?

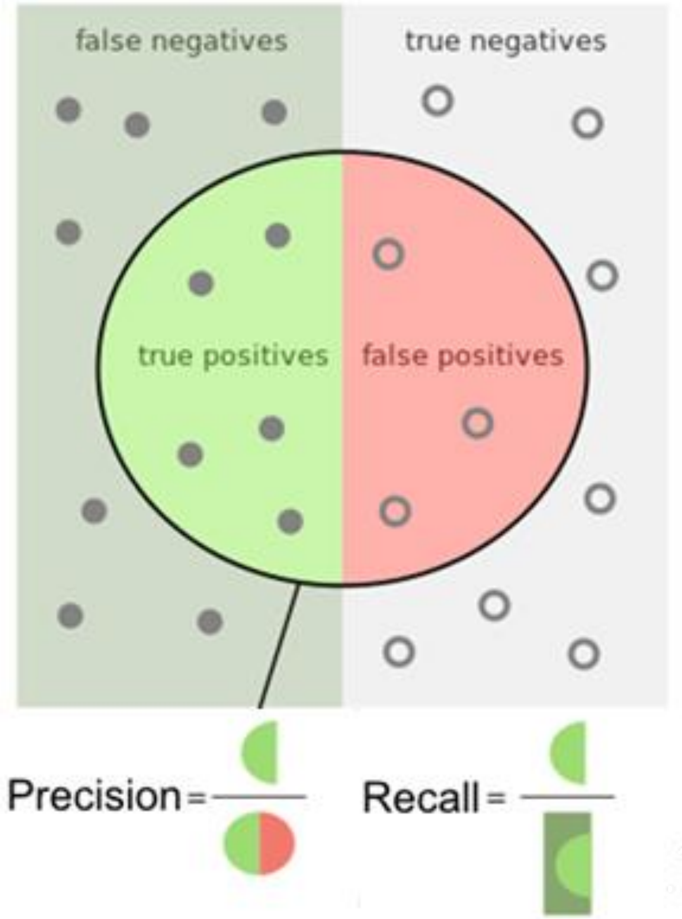
-Was the read able to be mapped?

-Was the reference base an N?

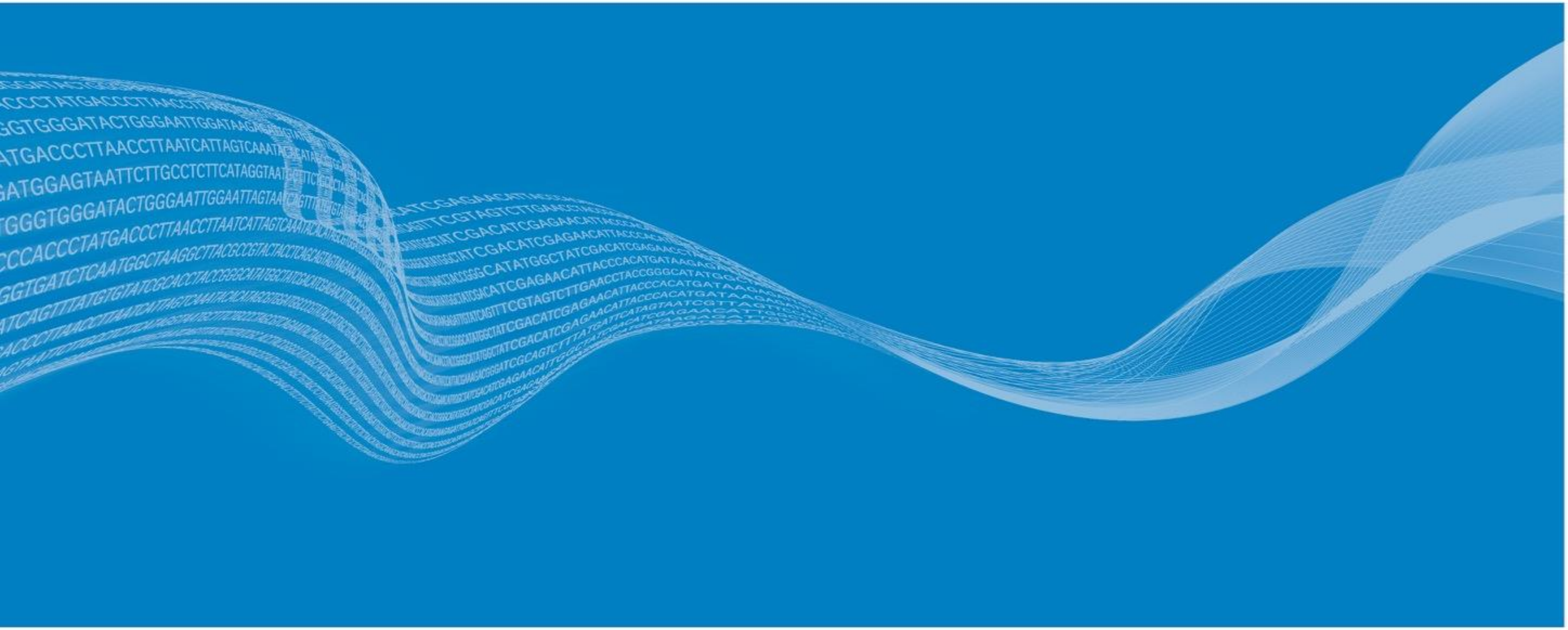
Precision and Recall Defined

Precision and Recall

- Precision: What percent of variant calls made are correct?
- Recall: What percent of known variants were detected?



New Software To Support RTA3



New Software to Support RTA3

UCS

- Combines BaseSpace Broker and Run Copy Service
- More robust, unlimited retries, seamless restart

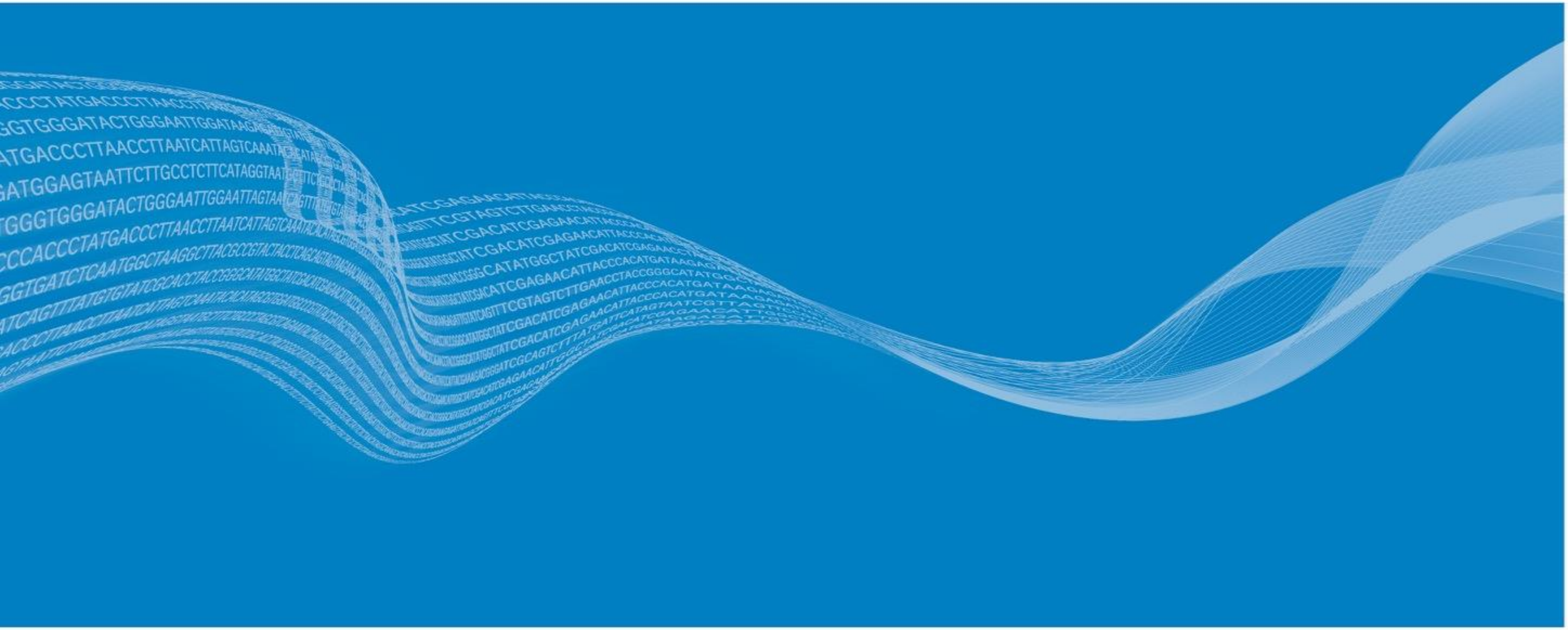
SAV

- New version required to handle new InterOp Folder Structure
- Not pre-installed on instrument
- Does not Autolaunch when starting a run

BCL2FastQ2

- Not required if sending data to BaseSpace Sequence Hub

Questions?



Revision History

Version	Updates
B	• Updated slide 4 to clarify content and remove typos • Changed slide 11 to prevent people from thinking there are 4 bins • Added Slides 19-29 to better explain how RTA3 assigns Q-Scores • Changed “reduced Quality Score Bins” to “Quality Score Reporting” • Updated info on UCS on slide 31