Data Analysis Overview

illumina®

Learning Objectives

By the end of this lesson, you will be able to:

- Identify the outcomes of each step of the Illumina data analysis workflow
- Define the components of primary data analysis
  - Instrument Control Software (ICS)
  - Real Time Analysis (RTA)
- Describe the image analysis steps
- Describe the base calling and filtering processes

Data Analysis Workflow

Illumina Data Analysis Workflow

1 **Primary Analysis**

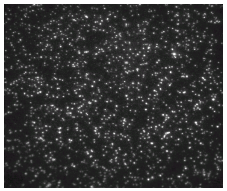
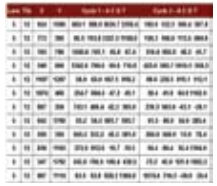
2 **Secondary Analysis**

MiSeq Reporter

CASAVA

3 **Data Visualization**

Primary and Secondary Analysis Overview

Analysis Type	Software	Outputs
Sequencing	 ICS/RTA	 Images/TIFF files
Primary Analysis	 ICS/RTA	 Intensities  Base Calling
Secondary Analysis	BaseSpace CASAVA MiSeq Reporter	 Alignments and Variant Detection

Illumina Data Analysis Workflow

1 **Primary Analysis**

2 **Secondary Analysis**

MiSeq Reporter

CASAVA

3 **Data Visualization**

Primary Data Analysis Workflow

RunFolders

RunFolders

Initiated on the Instrument Control Software computer

Contain all the data for a particular run

Each of the processes, image analysis and base calling write out their data to a run folder

Run folder data is transferred to Secondary Analysis Server where additional data is computed

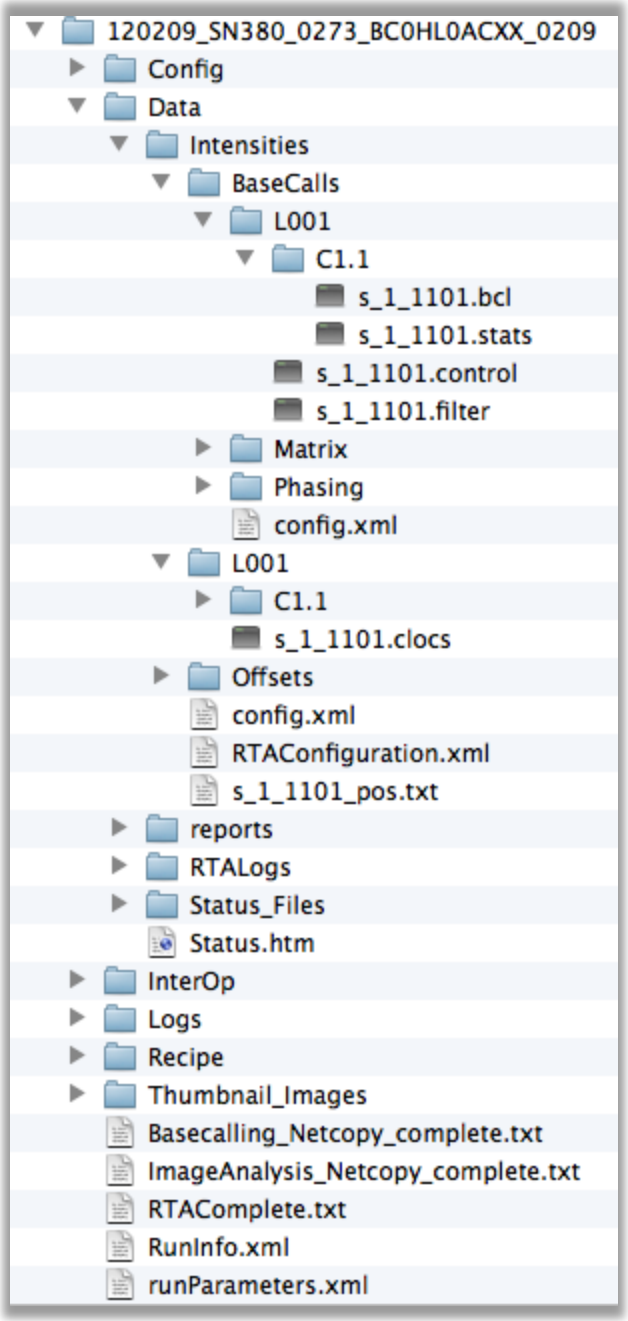


Image Analysis Workflow

Image Analysis

Performed by ICS and RTA

- Default workflow
- Analysis of the images in real time
- Automatically invokes when instrument software is started

Images generated on per base/cycle as TIFF files

- TIFF files (*.tif) are deleted after image analysis completes
- *.tif files never accumulate

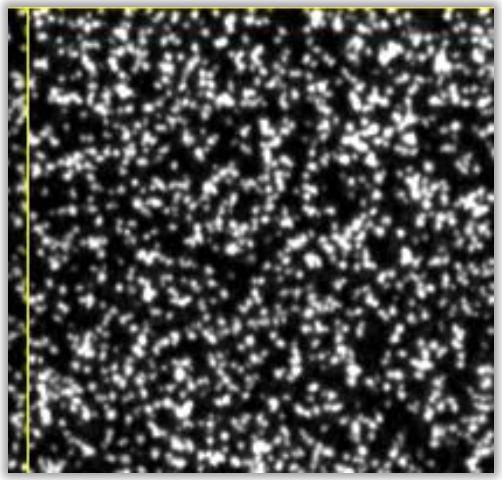


Image Analysis Input and Output

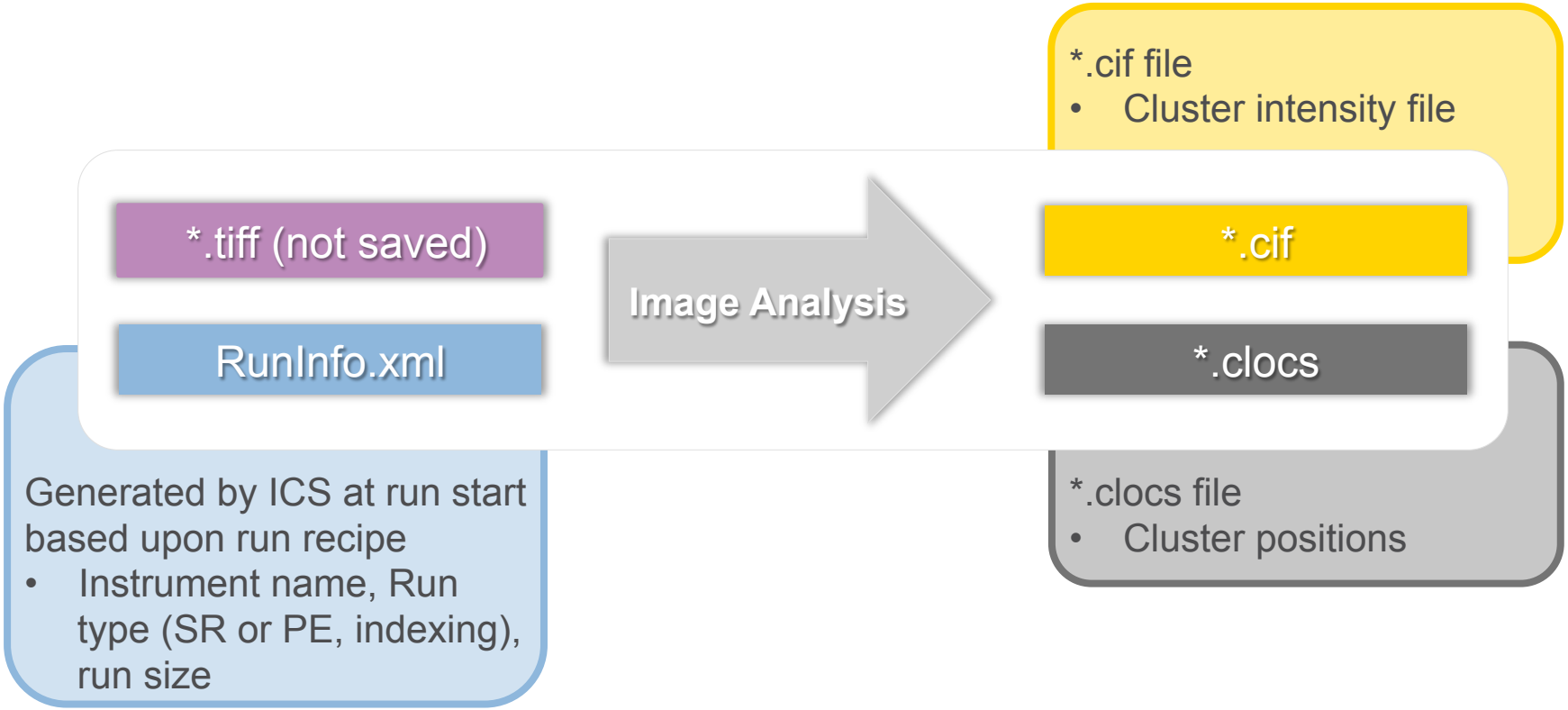
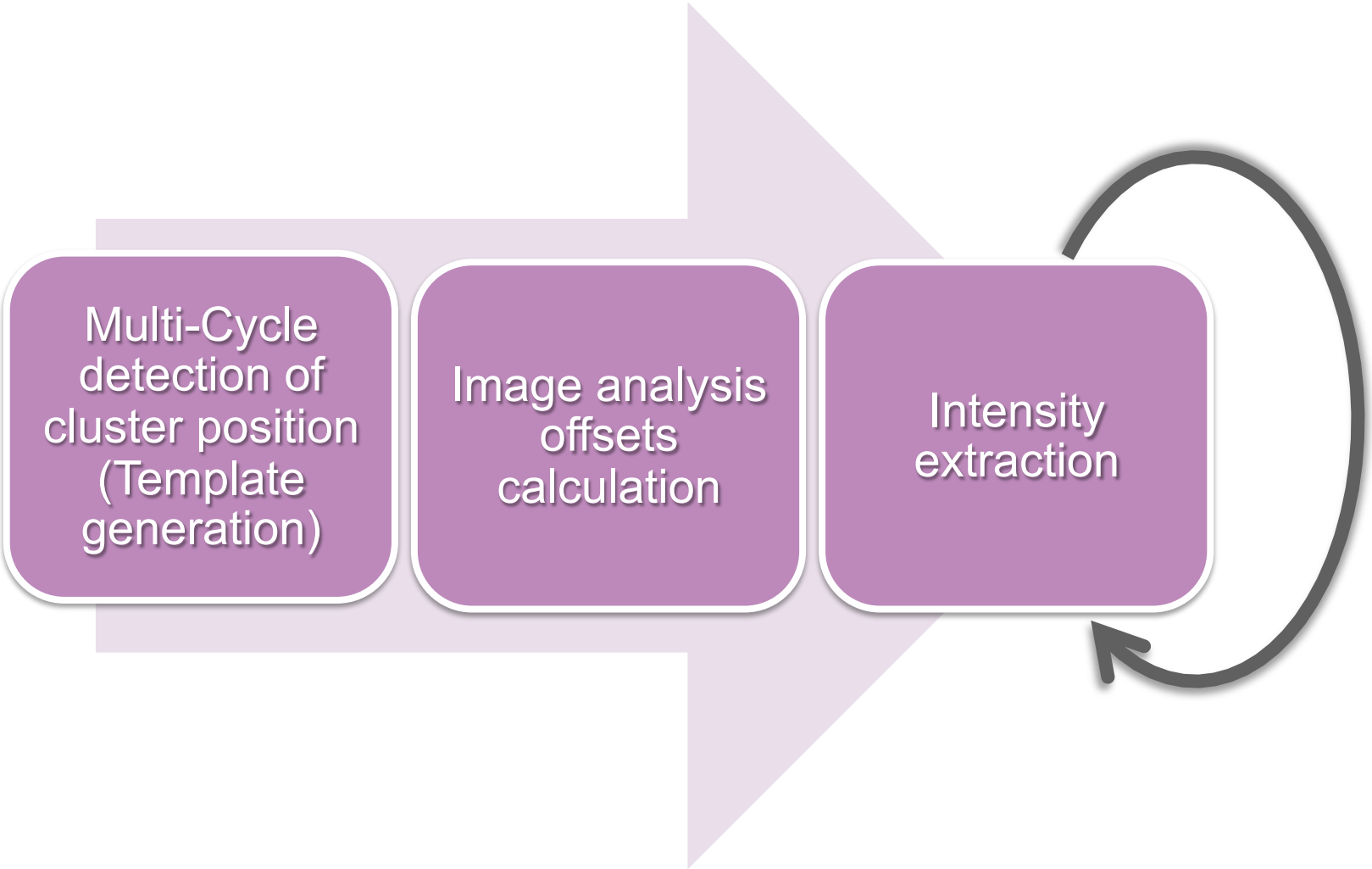


Image Analysis Workflow



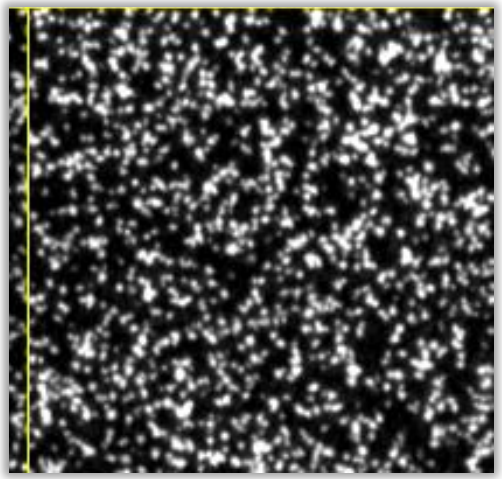
Template Generation

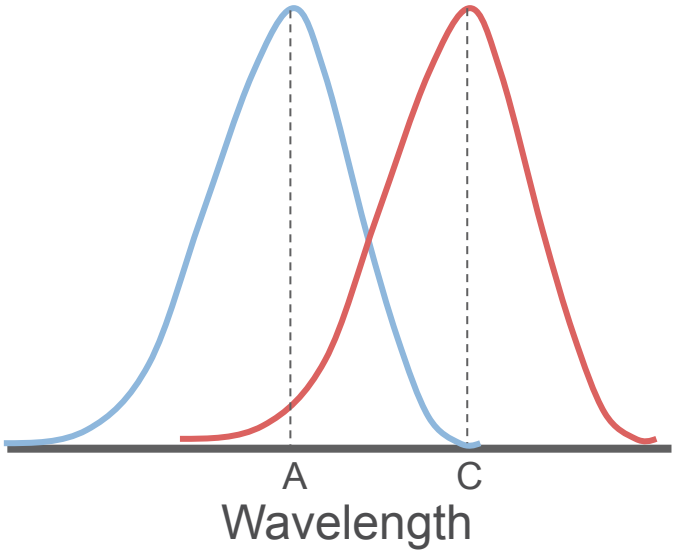
Clusters are bright spots on an image

Each cluster represents approximately 1000 copies of the same DNA strand in a 1 micron spot

Each image represents fluorescence in the G, A, T, or C channel

Spots that produce signal in the C channel (i.e., C nucleotides in a cycle) also have some signal in the A channel due to spectral overlap





Building a template (map of spot locations)

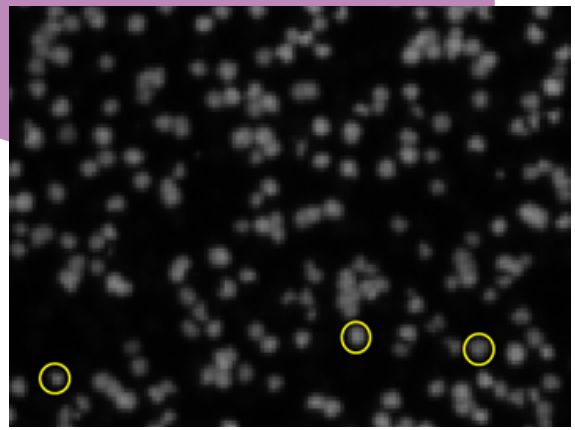
Goal:
Identify the location of every cluster (spot)

Common Spots Across Cycles

Multiple cycles are used to generate the template

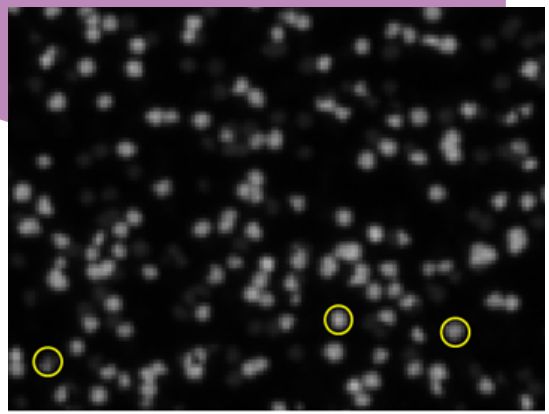
Some clusters have the same base in consecutive cycles, so the same spot produces signal in both cycles

Cycle 1



A Base Image

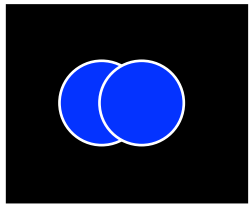
Cycle 2



A Base Image

Multi-Cycle Detection of Cluster Positions

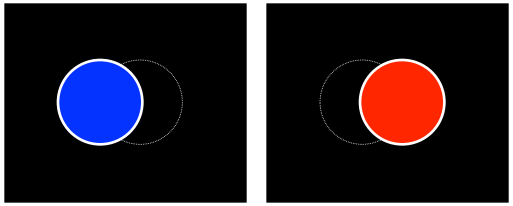
CYCLE 1



G channel

Difficult to resolve overlapping clusters in one cycle only, when there are overlapping clusters of the same base

CYCLE 2



G channel C channel

By detecting cluster positions in multiple cycles there is a better chance of resolving overlapping clusters since they are less likely to be of the same base in multiple cycles

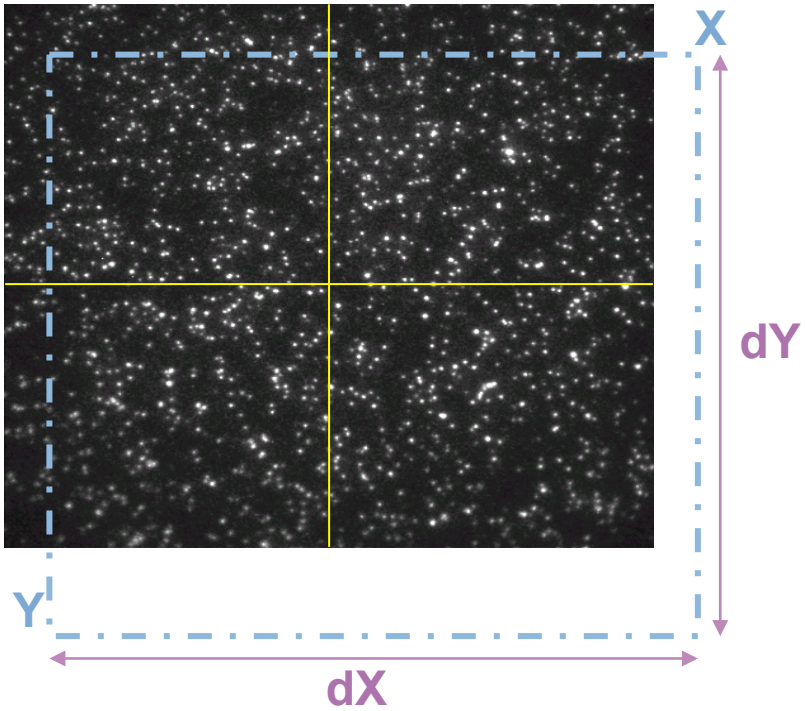
Image analysis offsets calculations

Each channel has slightly difference position and scale

- Images must be aligned

For each channel

- Identify X, Y shift between the 4 corner regions of image and template that maximizes correlation
- Apply the offset to the image



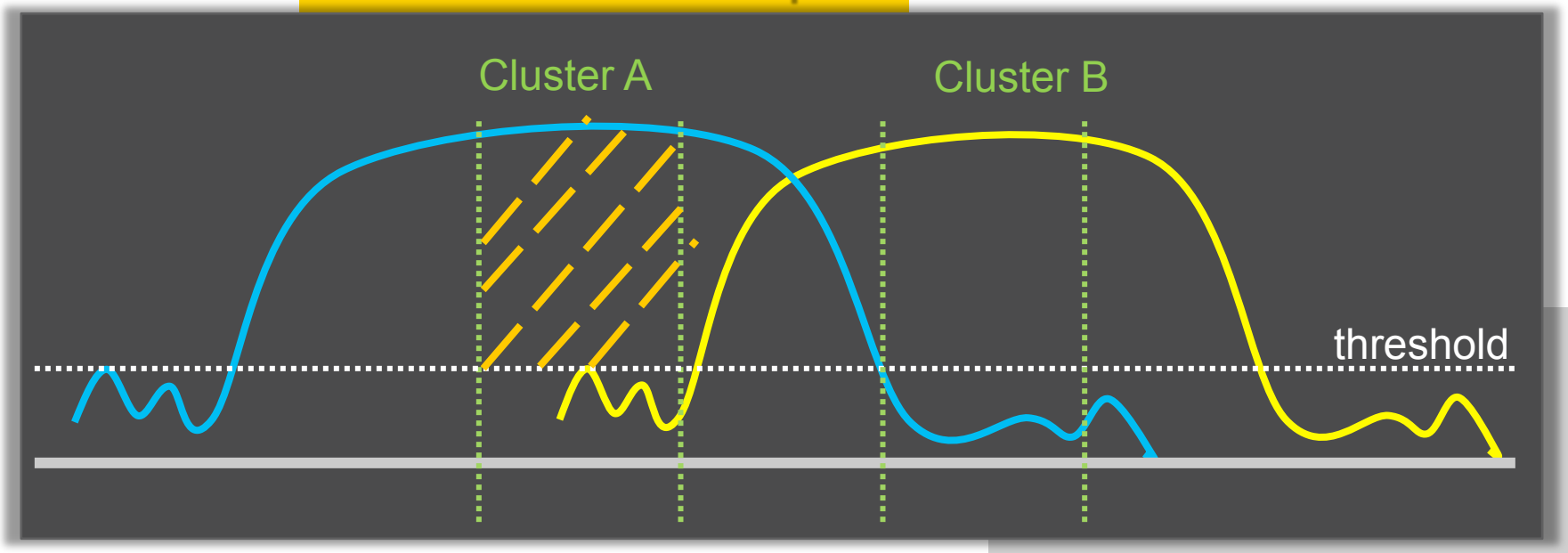
Offsets/offsets.txt

X	Y	dx	dy	
0.00	0.00	0.00000	0.00000	A
0.32	1.41	0.00069	0.00068	T
-0.01	1.82	-0.00123	-0.00125	C
0.14	1.59	-0.00097	-0.00092	G

Process of Extracting Intensity

Part 1

Only a small area of the signal is considered to avoid cluster overlap

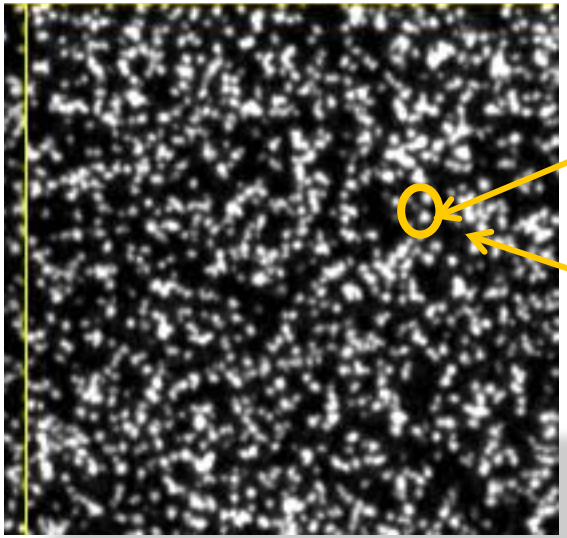


Values below threshold are not considered

Process of Extracting Intensity
Part 2

- Compute background for each cluster
- Compute signal for each cluster
- Subtract the background from each cluster
- Save the intensity values to the *.cif file

This is a primary reason why over clustering is problematic



Base Calling

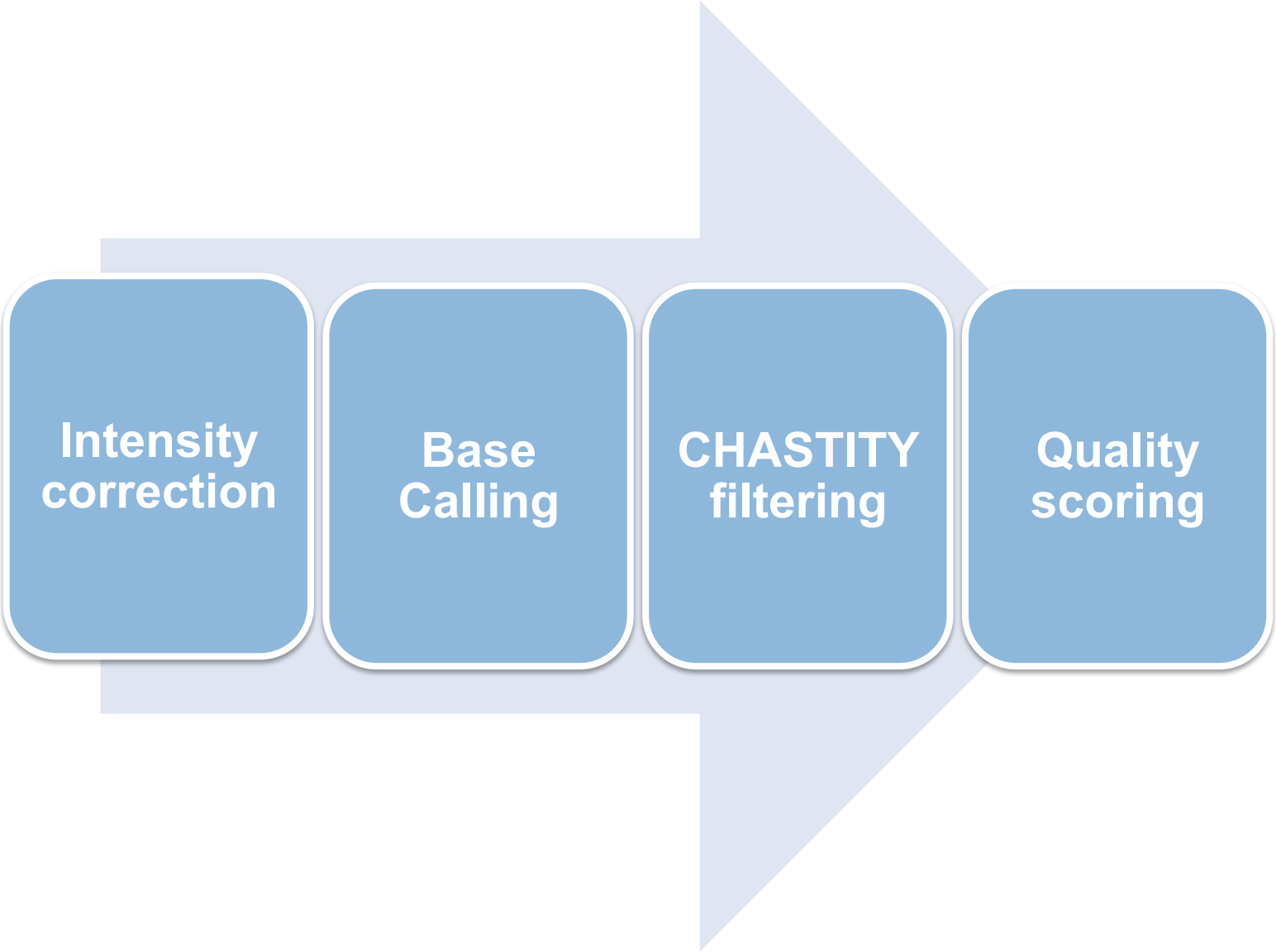
Base Calling Input and Output

Base calling
- Performed by RTA
- Starts automatically after Image analysis

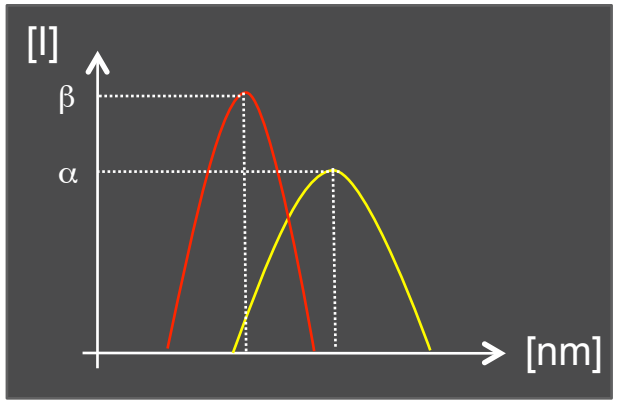
***.fastq file**
- an industry standard

fastq format contains:
- information about the read
- sequence of the read
- Q scores for each base in the read sequence

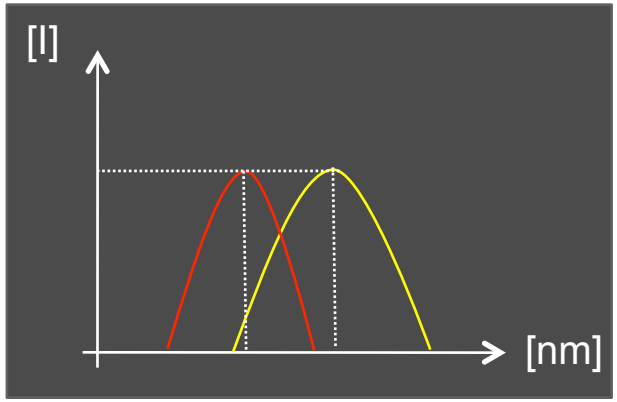
Base Calling Workflow



Intensity Correction



Raw – different peak intensities



Normalized – same peak intensities

The fluors used to represent G, A, T, C bases have difference intensities per molecule

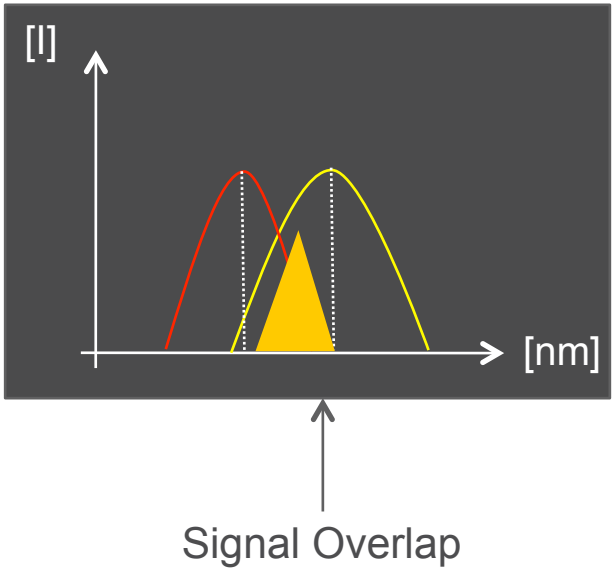
Normalizing the intensities of all four bases to a common mean reduces this effect

Cross-talk Estimation

Crosstalk
For overlapping emission spectra, it is possible to define and remove overlaps in signal to achieve a purer read

Parameters can be estimated and placed in a matrix

The matrix can then be used to correct for the dye crosstalk



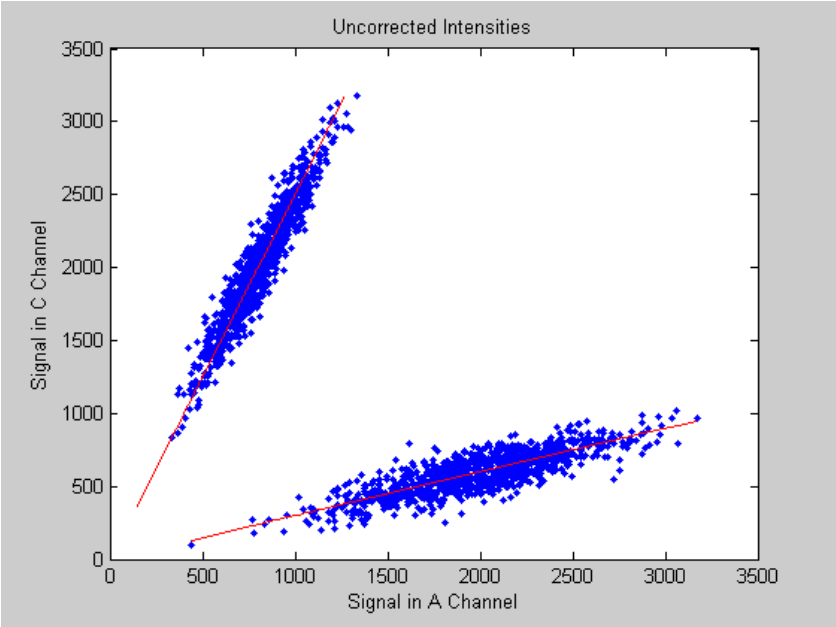
	A	C	G	T
A	0.65	0.14	0.01	0.02
C	0.81	1.04	0.02	0.03
G	0.02	0.02	1.18	0.04
T	0.03	0.03	1.10	1.56

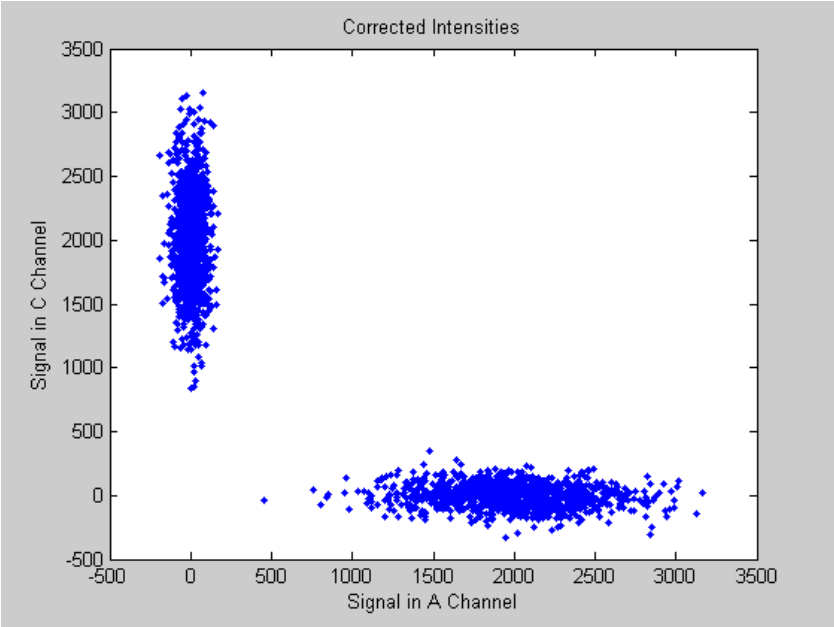
Matrix

Graphical View of Crosstalk Correction

Signal in one channel affects Signal in the other channel

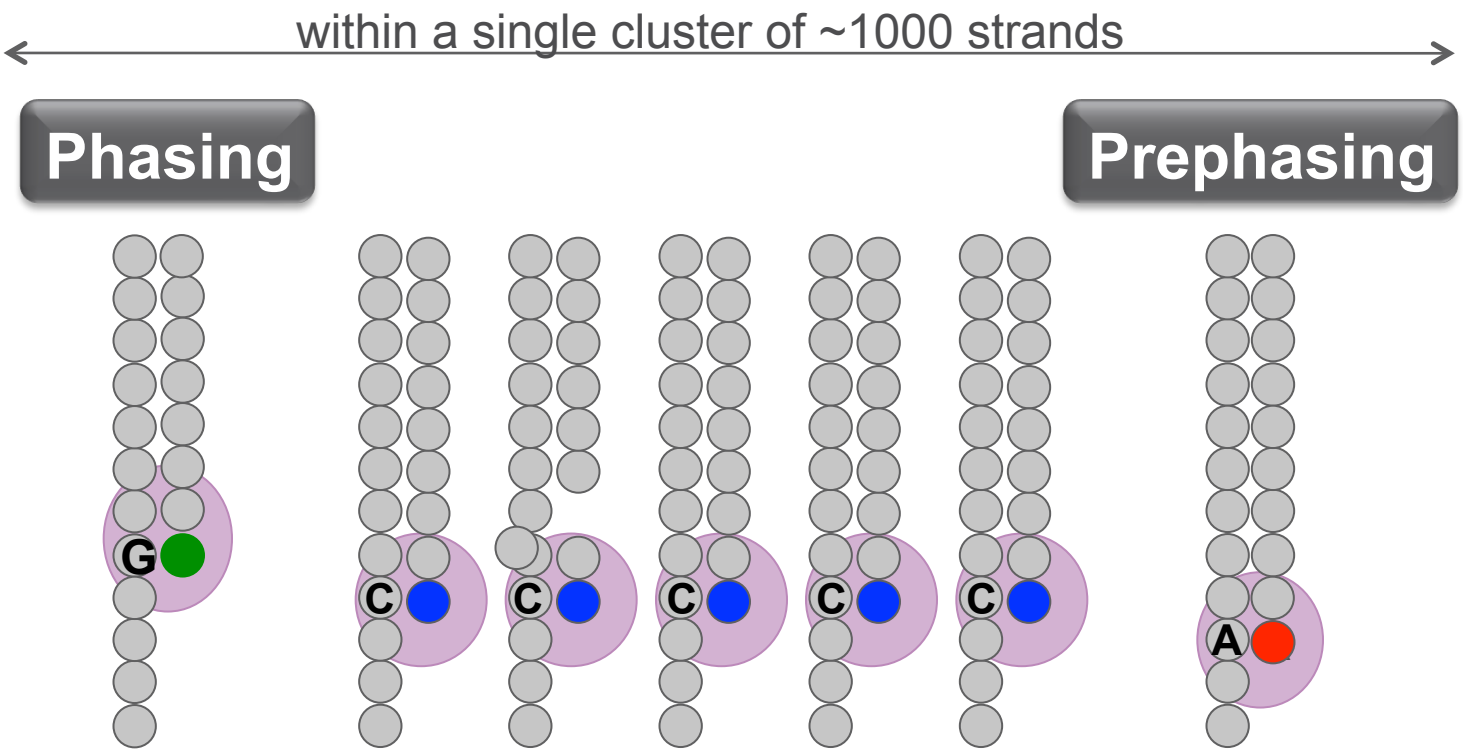
Signal in one channel is independent of signal in the other channel





Correction for spectral overlap requires unbiased base content of template that is used for the calculation

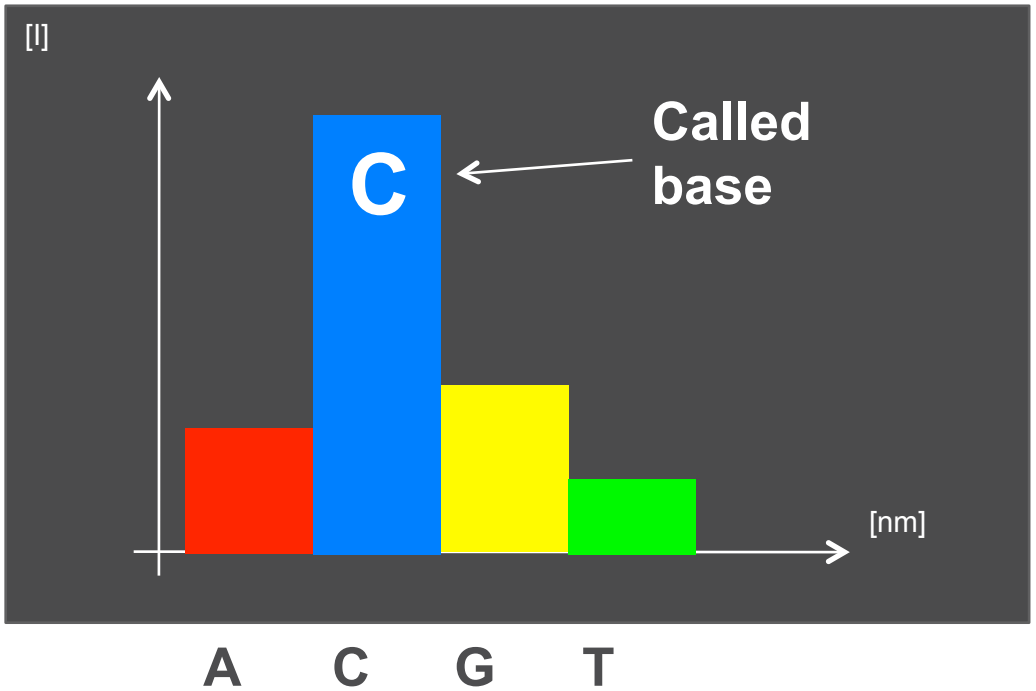
Base Calling Phasing Correction



Base Calling

After intensity correction, the base with the highest intensity is the one called

For base positions where all bases are very low intensity, no base may be called



Quality Filtering of Clusters

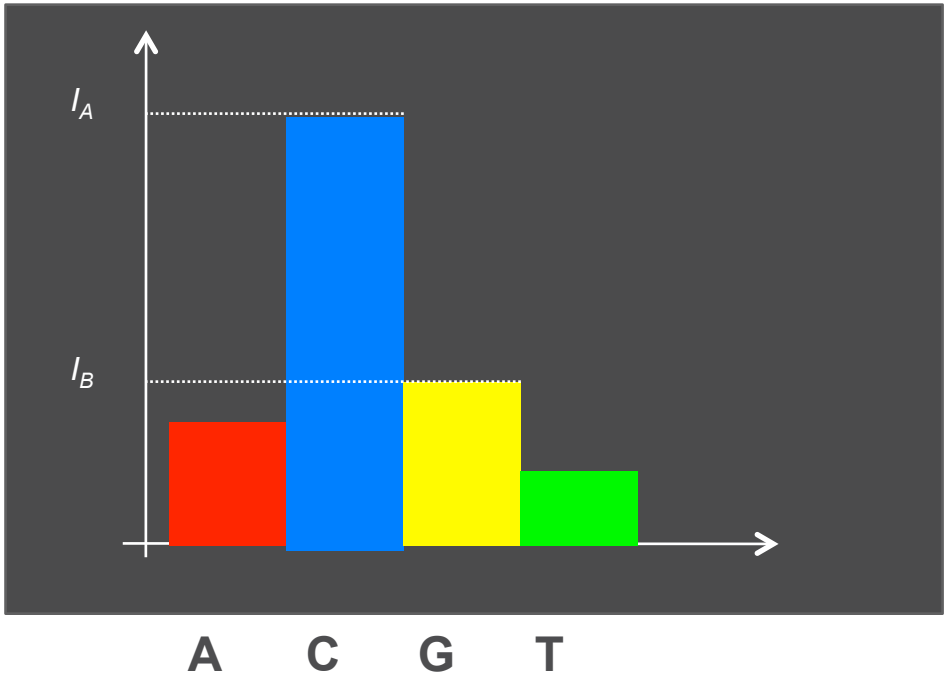
CHASTITY
the ratio of the highest intensity to the sum of the highest and second highest intensities

CHASTITIY filter is calculated for each cluster over the first 25 bases of the sequence

Filters cluster by signal purity
Removes overlapping and low-intensity clusters

$$C = \frac{I_A}{I_A + I_B}$$

CHASTITY formula



Quality Scoring

Quality Scores

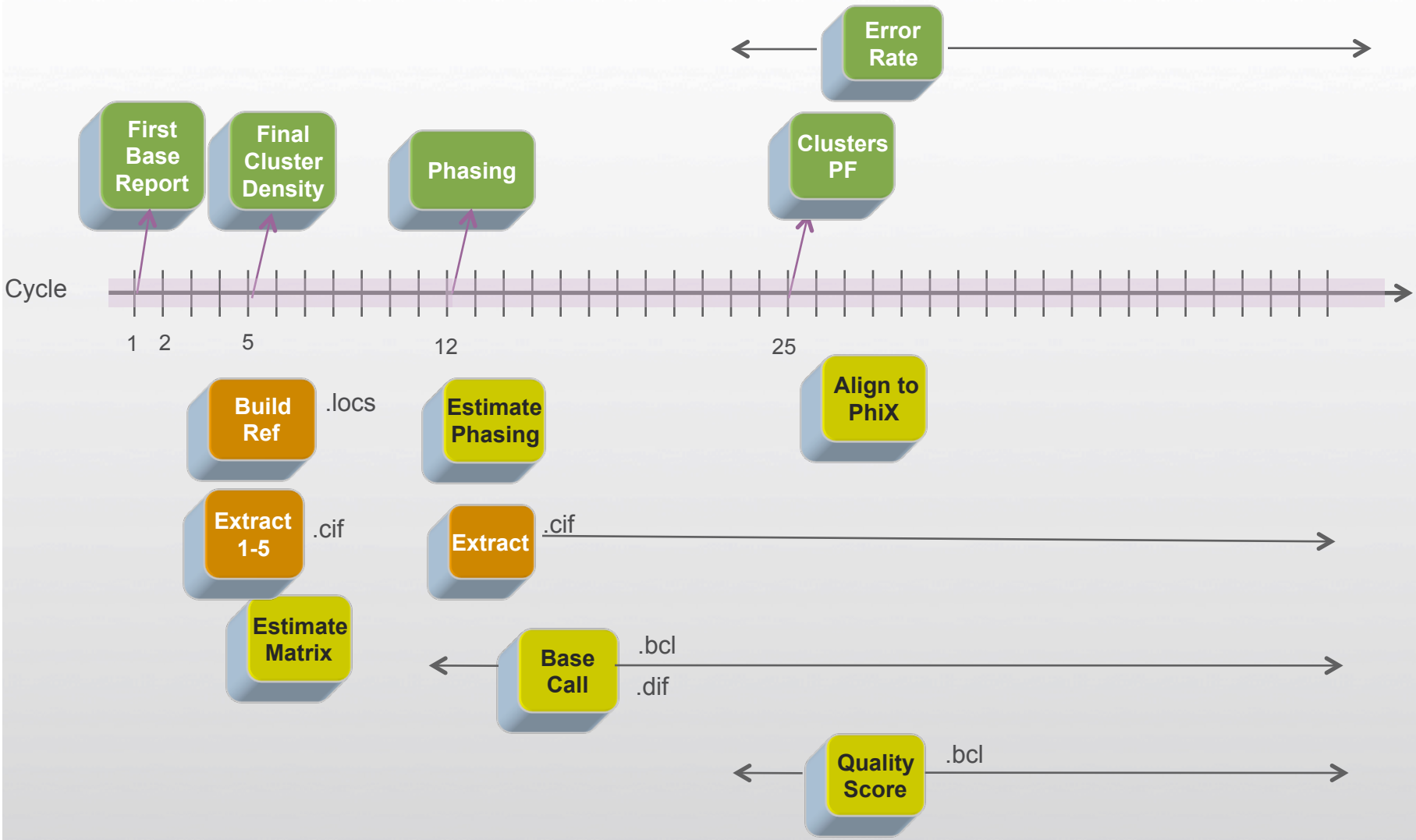
- A prediction of the probability of an error in base calling
- A method for assigning quality scores to sequencing data using numerical predictors of base calling

Produced by a model that uses quality predictors as inputs and produces Q-scores as outputs

Quality scores are calculated after quality filtering, starting at cycle 24

Phred Quality Score	Probability of Incorrect Based Call	Base Call Accuracy	Q-score
10	1 in 10	90%	Q10
20	1 in 100	99%	Q20
30	1 in 1000	99.9%	Q30
40	1 in 10000	99.99%	Q40

Detailed Analysis Workflow



Illumina Data Analysis Workflow

1 Primary Analysis

2 Secondary Analysis

MiSeq Reporter

CASAVA

3 Data Visualization

illumina®

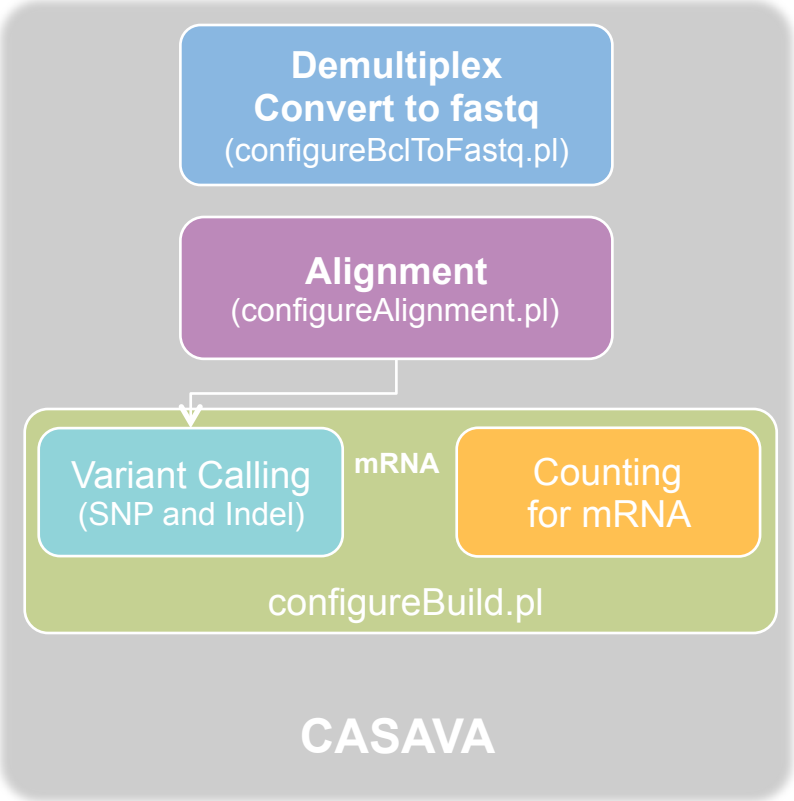
Options for secondary analysis

CASAVA	• GA IIx • HiSeq
MiSeq Reporter (MSR)	• MiSeq
BaseSpace	• MiSeq
Third Party Software	• GA IIx • HiSeq • MiSeq

CASAVA Overview

CASAVA
- Consensus Assessment of Sequence and Variation
- Linux-based application

CASAVA Build
- CASAVAS's output folder structure
- Ready for import into Genome Studio for visualization and further analysis



MiSeq Reporter Overview

What is MSR?

- A Windows-based application that runs on the MiSeq computer
- Does not have a software icon

What does MSR do?

- Performs on-instrument secondary analysis
- Processes base calls generated during primary analysis

When does Secondary Analysis begin?

- Secondary analysis begins immediately following completion of primary analysis

How long does the Analysis take?

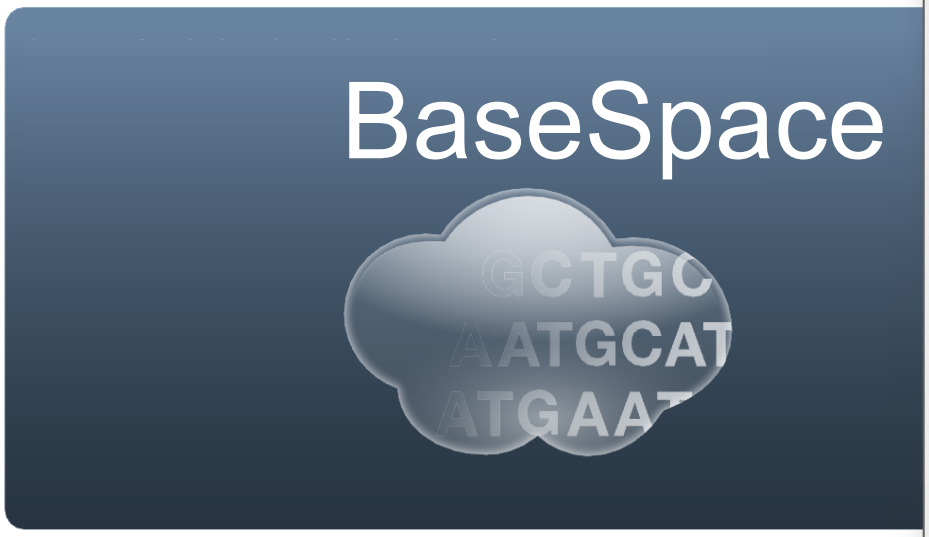
- Analysis times vary depending upon the experimental design and setup

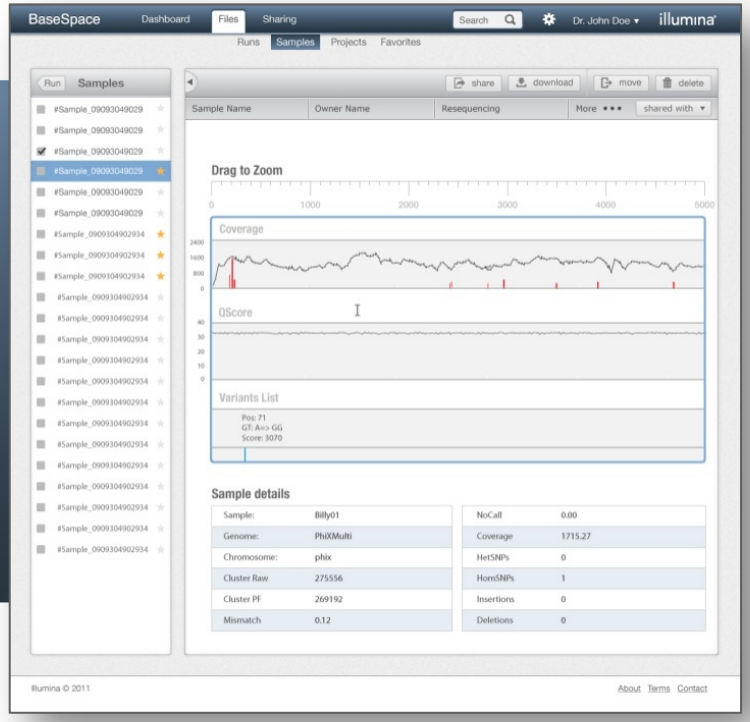
How can I view my Data?

- Use any web browser on another computer connected to the same network as your MiSeq to view your data

BaseSpace Overview

- BaseSpace is Illumina's genomic cloud computing environment
  - Eliminates need for onsite storage and computing
  - Enables Web-based data management and analysis
  - Provides tools for collaboration and sharing
  - Available GLOBALLY for Illumina and non-Illumina customers





Illumina Data Analysis Workflow

1 Primary Analysis

2 Secondary Analysis

MiSeq Reporter

CASAVA

3 Data Visualization

Options for data visualization

Genome Studio	• GA IIx • HiSeq
MiSeq Reporter (MSR)	• MiSeq
BaseSpace	• MiSeq
Third Party Software	• GA IIx • HiSeq • MiSeq

Questions?