

Cluster Optimization Overview

This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document.

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY, AND WILL VOID ANY WARRANTY APPLICABLE TO THE PRODUCT(S).

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE).

© 2021 Illumina, Inc. All rights reserved.

All trademarks are the property of Illumina, Inc. or their respective owners. For specific trademark information, see www.illumina.com/company/legal.html.

Revision History

Document	Date	Description of Change
Document # 1000000071511 v01	April 2021	Added HTML format.
Document # 1000000071511 v00	March 2019	Initial release.

Table of Contents

Revision History	iii
Introduction	1
Optimal Cluster Density	3
Common Clustering Issues and Prevention	5
Library Quality	5
Library Quantification	5
Loading Concentration	6
Nucleotide Diversity	7
ExAmp Reagent Preparation	9
Diagnosing Suboptimal Clustering (Patterned Flow Cells)	10
Diagnosing Suboptimal Clustering (Nonpatterned Flow Cells)	12
Summary Metrics	12
Imaging Metrics	12
Run Charts	14

Introduction

This documentation describes strategies for optimizing cluster density and preventing and diagnosing clustering issues on Illumina flow cells. Use this guide as a reference when preparing and sequencing libraries.

Related Videos

Several topics in this guide have complementary videos. Visit the [Training](#) page of the Illumina website to watch the videos.

Topic	Video
Cluster generation and sequencing by synthesis (SBS)	<i>Sequencing: Illumina Technology</i>
Library quantification	<i>How do I achieve consistent quantitation? Part 1</i> <i>How do I achieve consistent quantitation? Part 2</i>
Base calling, clusters passing filter, and nucleotide diversity	<i>How do I optimize amplicon sequencing data? Part 1</i> <i>How do I optimize amplicon sequencing data? Part 2</i>
Overclustering nonpatterned flow cells	<i>Is my HiSeq or MiSeq run overclustered?</i>

Clustering Overview

A cluster is a clonal group of library fragments on a flow cell. Each cluster produces one single read or one paired-end read. For example, a flow cell with 10,000 clusters produces 10,000 single reads or 20,000 paired-end reads.

i | A paired-end read sequences both ends of a DNA fragment in the same run, while a single read sequences only one end. For more information, see [Indexed Sequencing on Illumina Systems](#).

During clustering, each fragment binds to the flow cell and seeds a template that is amplified until the cluster consists of hundreds or thousands of copies. The number of clusters and the location of each cluster is fixed throughout a run. An incorporation mix flows through the flow cell, tagging each fragment with a fluorescent-labeled nucleotide. Base calls are made from the resulting signal (intensity) that each cluster emits.

Base Calling and Passing Filter

The Real-Time Analysis software runs on the instrument control computer. During a sequencing run, it extracts intensities from images to perform base calling, and then assigns a quality score to the base call.

Cluster density on a flow cell impacts the following steps in the Real-Time Analysis workflow:

- **Passing filter**—During cycles 1–25 of Read 1, a filter removes unreliable clusters from the image extraction results. Clusters pass filter when certain quality specifications are met. For more information, see *Calculating Percent Passing Filter for Patterned and Nonpatterned Flow Cells (Pub. No. 770-2014-043)*.
- **Registration and intensity extraction**—For each cluster on the flow cell, the software records a cluster location and calculates an intensity value.
- **Template generation**—The software analyzes images from the first 5–7 cycles of a run to map the location of each cluster on a nonpatterned flow cell. (Cluster locations on a patterned flow cell are predetermined.) The resulting template is input for the subsequent registration step.

The implementation of Real-Time Analysis, including workflow steps, varies by system. For system-specific information, see the system guide for your instrument.

Optimal Cluster Density

The density of clusters on a flow cell significantly impacts data quality and yield from a run, and is a critical metric for measuring sequencing performance. It influences run quality, reads passing filter, Q30 scores, and total data output.

Performing a run at optimal cluster density involves finding a balance between underclustering and overclustering. The goal is to sequence at a high enough density to maximize total data output, while maintaining a low enough density to avoid overclustering.

Effects of Underclustering and Overclustering

Overclustering increases signal brightness, which makes finding the focal plane difficult and causes poor template generation, poor cluster registration, and other image analysis issues. These issues negatively affect sequencing data in the following ways:

- **Lower Q30 scores**—Overloaded signal intensities decrease the ratio of base intensity to background, creating ambiguity during base calling and decreasing data quality.
- **Lower clusters passing filter (lower data output)**—Overclustered flow cells typically have more overlapping clusters, which cause poor template generation and a decrease in percent of clusters passing filter (%PF). The %PF metric indicates signal purity from each cluster. Lower %PF reduces yield (the number of bases in gigabases [Gb]) called for a run.
- **Inaccurate demultiplexing**—Index reads typically have lower diversity, which can cause poor base calling. Overclustering exacerbates the potential for poor base calling, leading to demultiplexing failure.
- **Run failure**—When overclustering is extreme, image focusing can fail and terminate the run at any cycle.

Underclustering maintains high data quality, but lowers data output. In general, underclustering is preferable to overclustering because the effects are less severe.

Recommended Cluster Densities

When targeting optimal cluster density for nonpatterned flow cells, use the raw cluster density range for your system and reagent kit as a guideline.

System	Reagent Kit	Raw Cluster Density (K/mm ²)
HiSeq 2500 (High Output)	HiSeq v4	950–1050
	TruSeq v3	750–850

System	Reagent Kit	Raw Cluster Density (K/mm ²)
HiSeq 2500 (Rapid Run)	HiSeq v2, TruSeq (v1), and Rapid Duo	850–1000
MiniSeq	MiniSeq High Output and Mid Output	170–220
MiSeq	MiSeq v3	1200–1400
	MiSeq v2	1000–1200
NextSeq	High Output and Mid Output (v2.5 and v2)	170–220

Density is measured as 1000 (K) clusters per square millimeter (mm²). Raw cluster density indicates how many clusters are on the flow cell, regardless of whether they passed filter.

Raw cluster density is not a useful metric for patterned flow cells because the ordered arrangement of nanowells ensures uniform cluster density.

Common Clustering Issues and Prevention

Clustering issues occur when a patterned flow cell is loaded with too high a concentration (overloading) or too low a concentration (underloading) for effective sequencing. For nonpatterned flow cells, too many clusters (overclustering) or too few clusters (underclustering) create these issues.

Inconsistent clustering from run to run or across a flow cell can indicate incompatibility between the library and system.

Library Quality

Library prep contaminants such as adapter dimers, primer dimers, or partial library constructs can impact quantification and clustering. Insufficient library cleanup can cause the presence of these contaminants.

Verify the quality and purity of all libraries using the method described in the library prep documentation, such as Bioanalyzer or Fragment Analyzer. Check for library integrity, average insert size, and contaminants. The average insert size is necessary for calculating library molarity.

Library Quantification

Inaccurate library quantification often leads to suboptimal clustering. Quantification is necessary for protocols that do not include a final bead-based normalization step.

Recommended Methods

The following table describes recommended library quantification methods. See the library prep documentation for product-specific recommendations and instructions.

Quantification Method	Description
qPCR ¹	qPCR is the most effective method of library quantification when paired with a standard of similar size range. This method measures only functional library fragments instead of all DNA species in a library (such as primer dimers, free nucleotides, library fragments). ²

Quantification Method	Description
PicoGreen or Qubit ¹	PicoGreen, Qubit, and other fluorometric methods measure dsDNA only and are best for libraries with a broad fragment size range. These methods can overestimate library concentration because they measure all dsDNA in a pool, including partially constructed and adapter-dimer contaminants. However, PicoGreen and Qubit are highly accurate when the Bioanalyzer quality assessment indicates low levels of library contamination.
Bioanalyzer	Although recommended for quality control purposes, use the Bioanalyzer to quantify three types of libraries only: TruSeq Small RNA, TruSight Tumor 26, and TruSeq Targeted RNA Expression. Due to decreasing accuracy with increasing library fragment size range, the Bioanalyzer is not optimal for quantifying other library types.

¹ Quantifying libraries is distinct from checking library quality, which can require use of the Bioanalyzer. See [Library Quality on page 5](#).

² For more information, see the *Sequencing Library qPCR Quantification Guide (document # 11322363)* and *Nextera Library Validation and Cluster Density Optimization (Pub. No. 770-2013-003)*.

Methods to Avoid

Illumina does not recommend a NanoDrop or spectrophotometry method to quantify libraries. The quantification includes single-stranded nucleic acids and free nucleotides, so these methods can overestimate library concentration.

Loading Concentration

Loading concentration or final loading concentration is the ultimate concentration of a library loaded onto an instrument for sequencing. After library prep, libraries are diluted to the loading concentration appropriate for the library type, sequencing system, and reagent kit.

- Loading libraries at a concentration that is too high results in overloading or overclustering, which reduces %PF and can cause run failure.
- Loading libraries at a concentration that is too low results in underloading or underclustering, which reduces data output and accuracy.


To determine the optimal loading concentration, ***adjust the loading concentration in small increments***. As a starting point, see the recommended loading concentration in the denature and dilute instructions for your system. Depending on the system, instructions are in the denature and dilute guide or the system guide.

The relationship between cluster density and loading concentration is nonlinear, so small, empirical adjustments are necessary to determine the optimum loading concentration. For example: If a 10 pM loading concentration results in 70% of optimal cluster density, do not assume that loading another 30% achieves optimal cluster density. Achieving optimal cluster density likely requires a smaller increase in loading concentration than 30%.

Denaturation Considerations

Sodium hydroxide (NaOH) denatures libraries normalized using standard quantification and quality control procedures. When denaturing libraries with NaOH, use freshly diluted NaOH with a pH > 12.5. Make sure that the final concentration of NaOH in diluted libraries is < 1 mM.

- NaOH that is not freshly diluted can acidify. The resulting decrease in pH impairs denaturation and reduces cluster density.
- Excess NaOH concentration in diluted libraries inhibits cluster formation. Dilute NaOH to the concentration indicated in the denaturation instructions for your system, using Tris-HCl as needed to neutralize the pH.

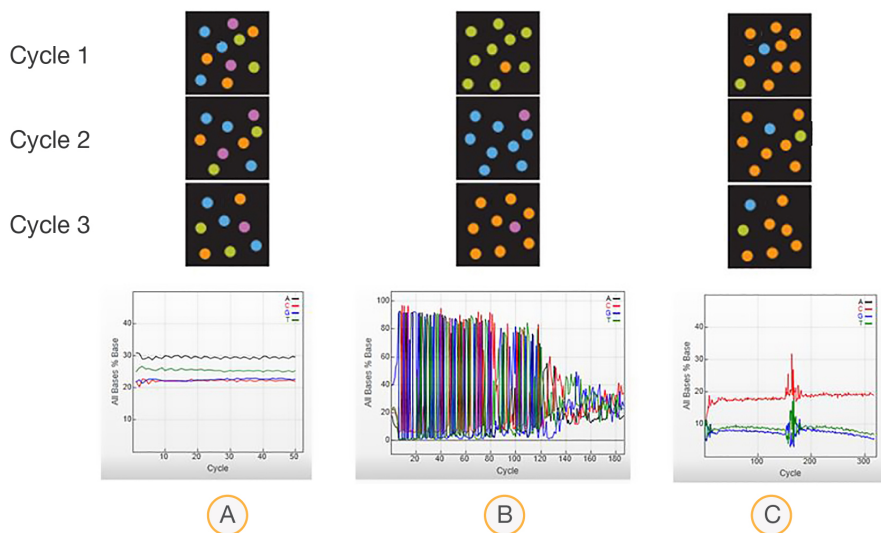
 | The iSeq 100 System denatures libraries automatically onboard the instrument, so these considerations do not apply.

Nucleotide Diversity

Illumina bases cluster density recommendations on diverse libraries. Diversity describes the proportion of each nucleotide (A, T, C, and G) at each position in a library. Low-diversity libraries compromise software performance and data accuracy.

A balanced or diverse library has equal proportions of A, C, G, and T. Low-diversity libraries, such as PCR amplicon, metagenomic, and ChIP, have an uneven proportion of nucleotides across the flow cell from one cycle to the next. Unbalanced libraries, such as bisulfite-converted libraries, have one base at a much lower percentage than the others.

Figure 1 Cluster Images and Data by Cycle



- A. **Diverse libraries**—Equal proportions of A, C, T, and G with even, horizontal curves centered on 25%.
- B. **Low-diversity libraries**—Uneven proportions of A, C, T, and G with large intensity spikes at each cycle.
- C. **Unbalanced libraries**—A low percentage of A and a high percentage of C.

Sequencing Low- and High-Diversity Libraries

When sequencing low-diversity libraries, use the following strategies to increase nucleotide diversity and provide a balanced signal:

- Reduce the loading concentration of the library. Empirically determine the reduction amount.
- Spike in PhiX or other high-diversity library. Empirically determine the spike-in amount, using the following percentages as a general guideline. Start with a higher spike-in percentage and reduce based on run performance.

System	PhiX Spike-In for Low Diversity (%)
HiSeq 2500	≥ 10
HiSeq 3000 and HiSeq 4000	5–20
HiSeq X	5–20
iSeq 100	≥ 5
MiniSeq	10–50
MiSeq	≥ 5
NextSeq 500 and NextSeq 550	10–50
NovaSeq 6000	~5

Although providing a balanced signal is not a concern for high-diversity libraries, Illumina recommends a 1–2% PhiX spike-in as a positive control for sequencing.

Sequencing New or Unknown Libraries

If a library is new or the nucleotide diversity is unknown, target a conservative loading concentration. A conservative loading concentration is at the lower end of the recommended cluster density range.

A 1–2% PhiX spike-in is also recommended.

ExAmp Reagent Preparation

For systems and workflows that require manually mixing ExAmp reagents, improper ExAmp preparation can impact metrics for percent occupancy and clusters passing filter. Percent occupancy indicates the percentage of wells on a patterned flow cell that contain at least one cluster, regardless of whether the clusters passed filter.

Follow instructions for preparing ExAmp reagents carefully, using the specified volumes and durations.

- **ExAmp mixing**—ExAmp reagents are viscous and must be mixed carefully. Pipette and dispense slowly so that the entire volume is expelled from the tip.
- **ExAmp staging time**—When too much time elapses between combining the ExAmp reagents and loading them onto the flow cell, the reagents start to degrade.
- **Flow cell staging time (NovaSeq Xp workflow only)**—When too much time elapses between loading the ExAmp/library mix onto the flow cell and starting the run, clusters form prematurely.

Diagnosing Suboptimal Clustering (Patterned Flow Cells)

Patterned flow cells consist of a nanowell substrate with millions or billions of ordered wells. The uniform cluster sizes and spacing increase cluster density and prevent overclustering. Although overclustering is not possible, loading a library with a suboptimal concentration negatively impacts data.

Run Metrics

During a run, review the following combination of run metrics to determine whether a patterned flow cell is underloaded, optimal, or overloaded. Depending on the system, metrics are available in BaseSpace Sequence Hub or Run Metrics Software. Some also appear on the instrument monitor after cycle 25.

These metrics can vary by library type and system. The low, medium, and high designations are relative to typical metrics and intended as a general guideline.

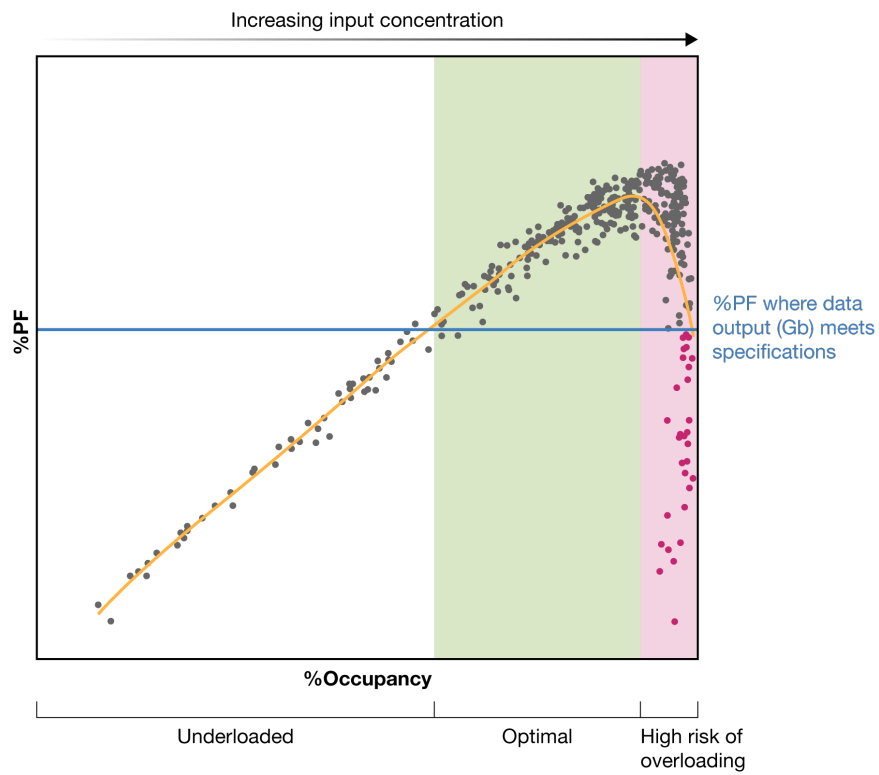
Metric	Underloaded	Optimal	Overloaded
%Occupancy¹	Low	High	High
%PF	Low	High	Low
% ≥ Q30	High	High	Variable
%Duplicates²	High	Medium	Low

¹ Available for the iSeq 100 and NovaSeq 6000 Systems only. Some software labels this metric %Occupied.

² When available from your bioinformatics pipeline.

The following figure plots example data points to illustrate the %PF and %Occupancy metrics for underloaded, optimal, and potentially overloaded flow cells. Red dots indicate overloaded flow cells. When reviewing run metrics, reference this figure to help determine optimum loading concentration. Illumina software does not generate plots of this type.

Figure 2 Relationship Between %PF and %Occupancy



Diagnosing Suboptimal Clustering (Nonpatterned Flow Cells)

During a run, monitor certain run metrics and thumbnail images to diagnose overclustering of a nonpatterned flow cell. Diagnosing overclustering early, canceling runs when necessary, or quickly identifying the root cause of a run failure improves sequencing efficiency.

Use Run Metrics Software (Run Metrics Software) or BaseSpace Sequence Hub to monitor run metrics. Run Metrics Software divides run metrics between several tabs, three of which are helpful for diagnosing overclustering. In BaseSpace Sequence Hub, the Charts and Metrics tabs under the Runs tab display the same metrics.

This section specifically describes how to use the software to diagnose overclustering. For more information on either application, see the *Sequencing Analysis Viewer Software Guide (document # 15066069)* or *BaseSpace Sequence Hub Online Help*.

Summary Metrics

After cycle 25 of Read 1, review the following run metrics to determine whether the flow cell is overclustered. Metrics are available from the Summary tab in Run Metrics Software or the Metrics tab in BaseSpace Sequence Hub.

- **Aligned**—Check whether the percent of reads that aligned to PhiX is close to the percent spiked in, which is ideal. A mismatch can indicate that the starting concentration of the library was misestimated. For example, for a 10% PhiX spike-in:
 - A 1% PhiX alignment indicates an unexpectedly high library concentration, generally with high cluster density or overclustering.
 - A 50% PhiX alignment indicates an unexpectedly low library concentration, generally with low cluster density or underclustering.
- **Clusters PF**—Check whether the percentage of clusters passing filter is sufficiently high.
- **Density**—Check whether the density value exceeds the range Illumina recommends for the system and reagent kit version. See [Recommended Cluster Densities on page 3](#).

Imaging Metrics

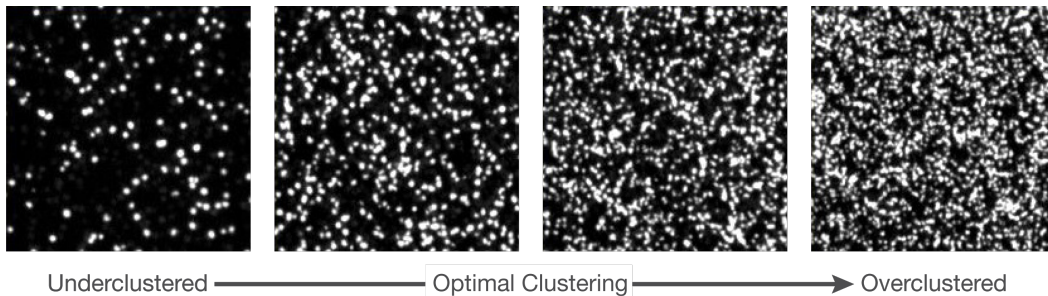
The Imaging tab in Run Metrics Software includes thumbnail images of tiles and detailed run metrics. Imaging metrics are not visible in BaseSpace Sequence Hub.

Thumbnail Images

Although not indicative of run quality, thumbnail images are useful for diagnosing clustering issues. Review thumbnail images to determine visually whether a flow cell is underclustered or overclustered. Because one thumbnail image shows one channel, review thumbnails from all channels for a comprehensive representation of the flow cell surface.

The following figure shows example thumbnail images for a range of cluster densities. The actual appearance of cluster density varies by system.

Figure 3 Thumbnail Images of Clusters



By default, the MiniSeq, NextSeq 550, and NextSeq 500 systems do not save thumbnail images. For help turning on this feature, contact Illumina Technical Support.

Metrics Table

The metrics table, which reports run metrics for each tile, is useful for diagnosing registration issues. The P90 (90th percentile of signal) A, C, G, and T table cells show the intensity values extracted from each cluster:

- With optimal clustering, these cells display numeric intensity values > 0 .
- With overclustering, these cells display 0 or NaN (not a number) even though the thumbnail images show clusters. This situation indicates that overclustering prevented the software from extracting intensity values.

Figure 4 Example P90 Values for an Overclustered Flow Cell (MiSeq System)

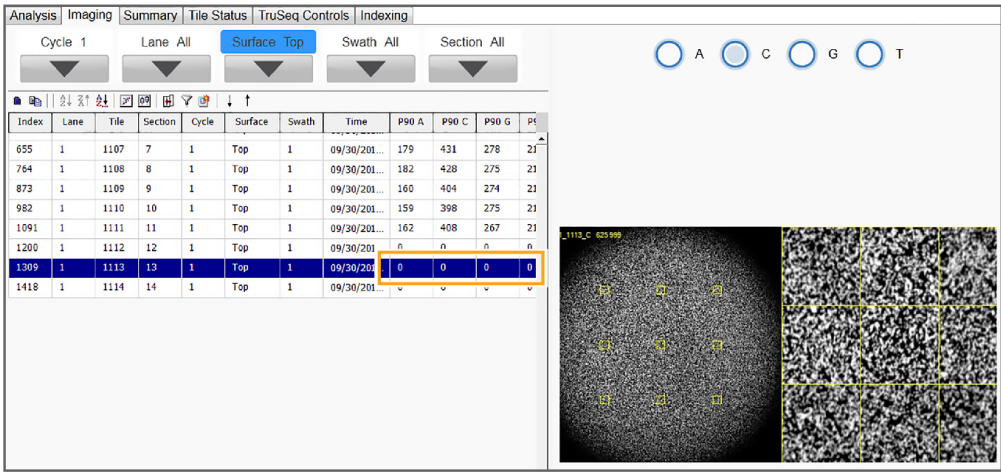
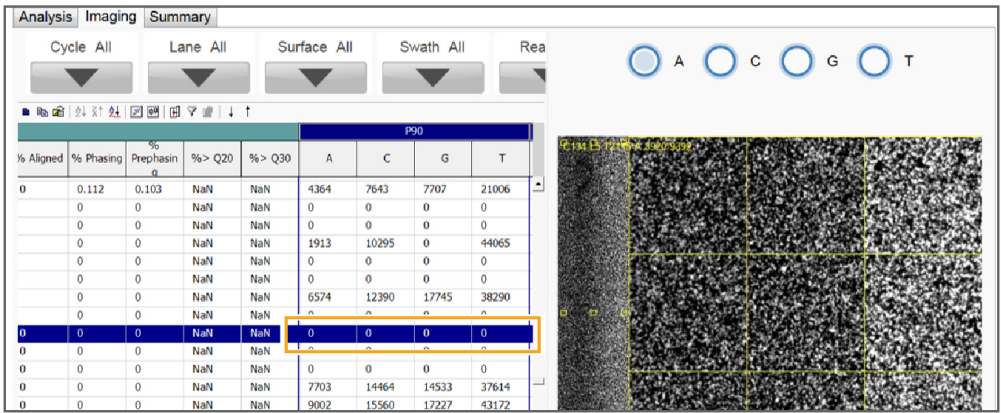


Figure 5 Example P90 Values for an Overclustered Flow Cell (HiSeq 2500 System)



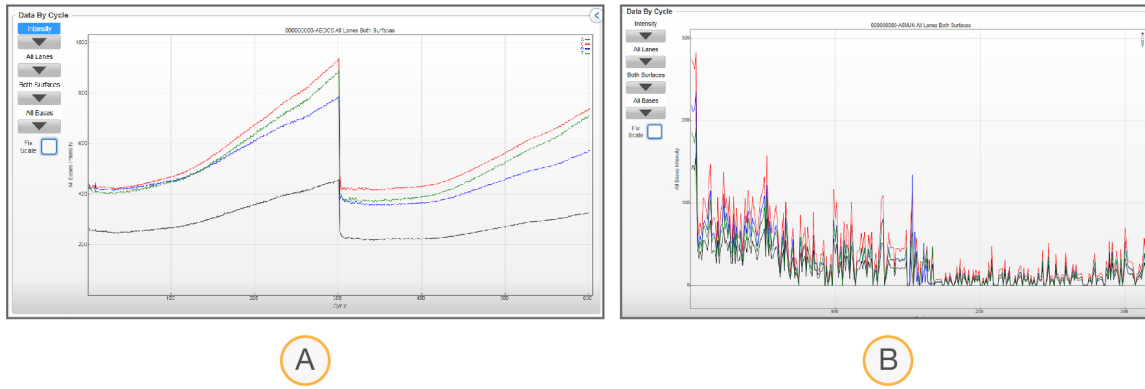
Run Charts

Several views from the Analysis tab in Run Metrics Software or the Charts tab in BaseSpace Sequence Hub provide metrics that are useful for diagnosing overclustering.

Data by Cycle: Intensity

Severe intensity drops in all channels early in the run can indicate poor template generation due to overclustering. When these drops occur, the software cannot extract intensity information from subsequent images so quality can be poor and the run might fail.

Figure 6 Comparison of Intensity Profiles in Run Metrics Software

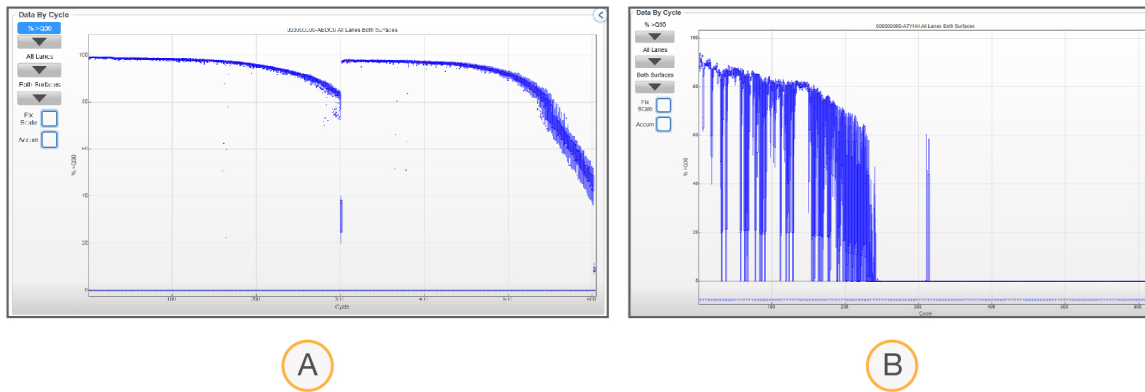


- A. Normally clustered flow cell
- B. Midrun failure due to an overclustered flow cell

Data by Cycle: % ≥ Q30

Overclustering can affect Read 1 or Read 2, but Read 2 is typically more affected. Extra amplification cycles during paired-end resynthesis slightly increase cluster sizes, which can increase the number of overlapping clusters. Overlapping clusters on an overclustered flow cell can affect image registration, causing poor Q30 scores and possible run failure.

Figure 7 Comparison of % ≥ Q30 Profiles in Run Metrics Software



- A. Normally clustered flow cell
- B. Large standard deviations preceding run failure due to an overclustered flow cell

Data by Lane: Density

Density box plots compare raw cluster density to %PF cluster density. Raw cluster density indicates how many clusters are on the flow cell, while %PF cluster density indicates how many of those clusters passed filter.

With optimal density, the raw cluster density and %PF box plots appear close together. As density increases beyond optimum, the %PF decreases and the box plots appear further apart. Also, clusters might be misidentified so raw cluster density is underestimated. With severe overclustering, no clusters pass filter and the %PF plot is displayed as a horizontal green line at zero density.

In the following figure: blue boxes illustrate raw cluster density range, green boxes illustrate %PF cluster density range, and red lines indicate median cluster density values.

Figure 8 Comparison of Density Box Plots in Run Metrics Software



- A. Optimal density
- B. Overclustered
- C. Severely overclustered

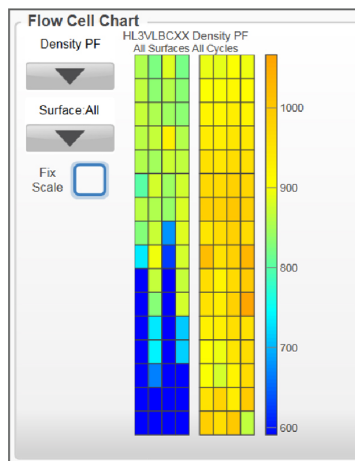
When the percentage of clusters passing filter drops to zero, raw cluster density becomes meaningless. Whether you loaded double the amount of library or 10 times the amount of library, the metric remains the same. When zero clusters pass filter, Illumina recommends rechecking libraries (quality and quantity) and incrementally adjusting the loading concentration.

Flow Cell Chart: Density PF

The Flow Cell Chart visualizes metrics for each tile across the entire flow cell. The Density PF view shows a range of cluster densities across all flow cell tiles. The legend (color scale) indicates which values the colors represent and dynamically changes with each run.

- With optimal density, the legend displays cluster density values within the recommended range.
- With overclustering, the chart has tiles at the higher end of the color range and can include blue tiles. Blue represents low-density tiles or tiles with zero density due to image extraction failure.

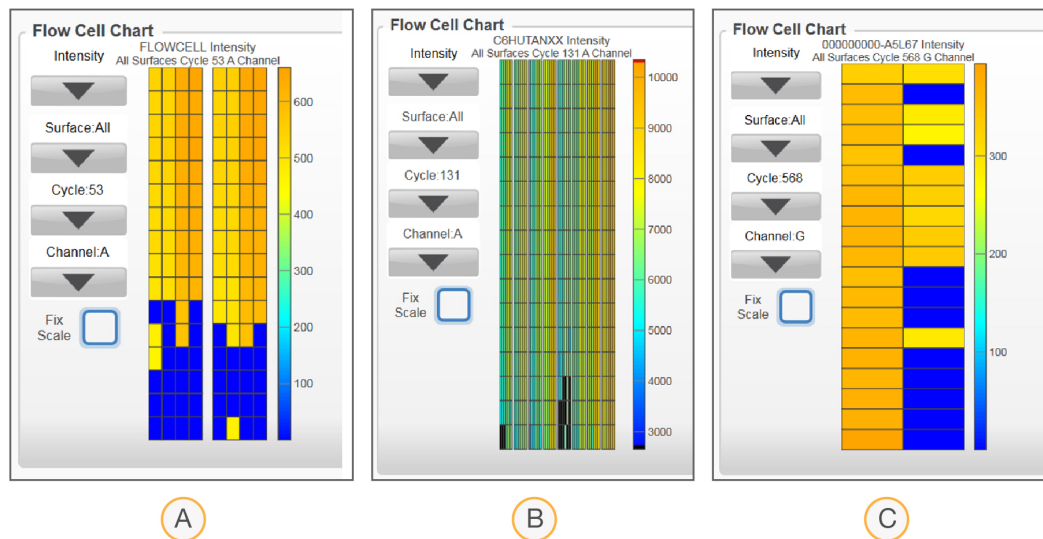
Figure 9 Density PF in Run Metrics Software Showing Severe Overclustering (HiSeq 2500 System in Rapid Run Mode)



Flow Cell Chart: Intensity

The Intensity view on the Flow Cell Chart is also helpful for evaluating overclustering. Blue or black tiles represent tiles with intensities lower than other tiles due to high cluster density.

Figure 10 Intensity in Run Metrics Software Showing Severe Overclustering



- A. HiSeq 2500 System in Rapid Run mode
- B. HiSeq 2500 System in High Output mode
- C. MiSeq System



Illumina

5200 Illumina Way

San Diego, California 92122 U.S.A.

+1.800.809.ILMN (4566)

+1.858.202.4566 (outside North America)

techsupport@illumina.com

www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

© 2021 Illumina, Inc. All rights reserved.

illumina[®]