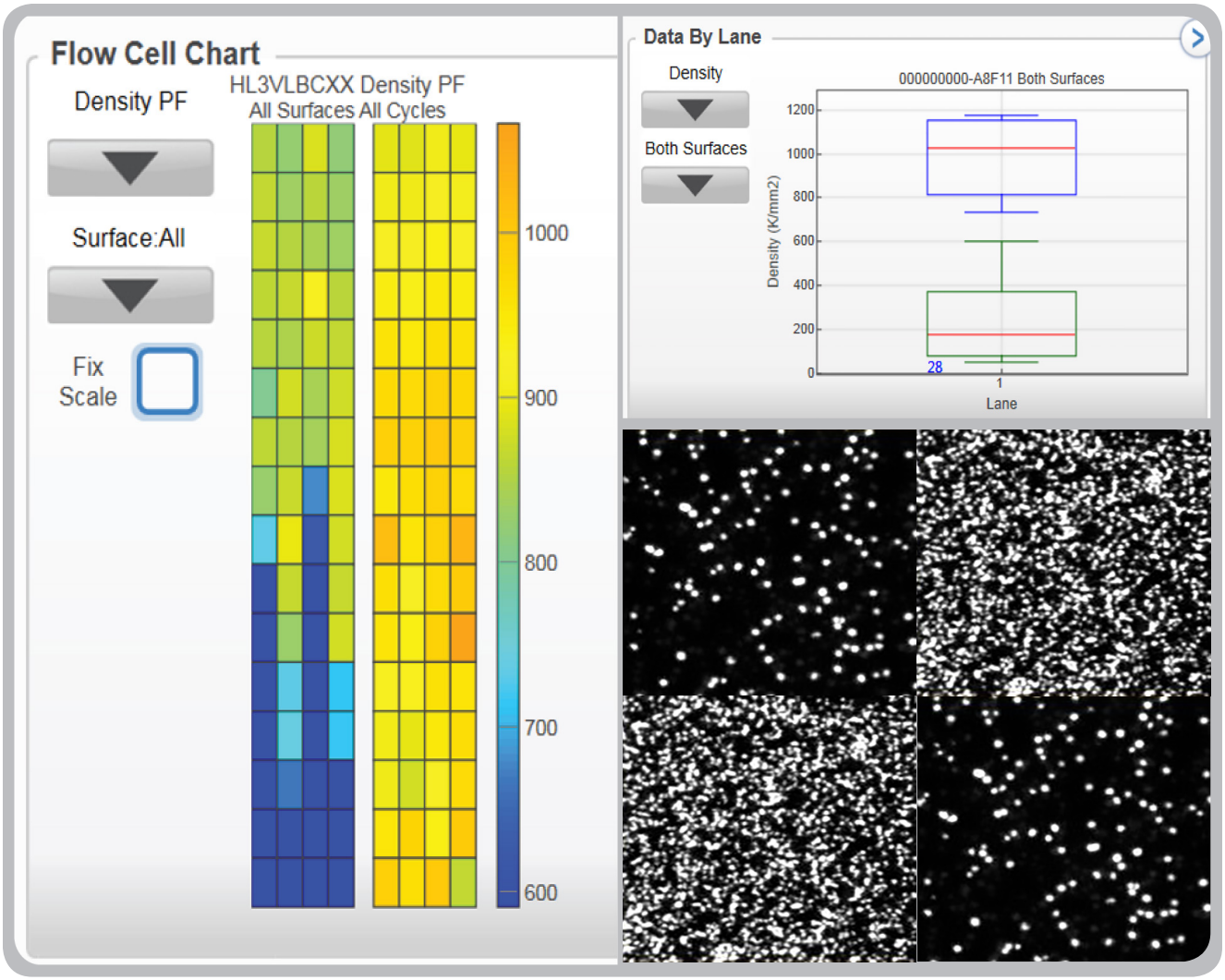

Optimizing Cluster Density on Illumina Sequencing Systems

Understanding cluster density limitations and strategies for preventing under- and overclustering.

illumina®

Table of Contents

I. Introduction

The Illumina sequencing workflow is based on 3 simple steps: libraries are prepared from virtually any nucleic acid sample, amplified to produce clonal clusters, and sequenced using massively parallel synthesis. The density of clonal clusters has a large impact on sequencing performance in terms of data quality and total data output. This primer discusses the principles of optimal cluster density, how to diagnose overclustering, and common causes and strategies for the prevention of overclustering on Illumina sequencing systems.

II. Understanding Optimal Cluster Density

Cluster density is a critically important metric that influences run quality, reads passing filter, Q30 scores, and total data output. While underclustering maintains high data quality, it results in lower data output. Alternatively, overclustering can lead to poor run performance, lower Q30 scores, the possible introduction of sequencing artifacts, and—counterintuitively—*lower total data output*.

Performing a run at optimal cluster density involves finding a balance between under- and overclustering. The goal is to sequence at high enough densities to maximize total data output, while maintaining low enough densities to avoid the negative effects of overclustering.

a. How Does Overclustering Affect Sequencing Data?

Overclustering creates image analysis problems, including loss of focus, poor template generation, and issues with cluster registration. The increased overall signal brightness of the flow cell makes it difficult for the sequencer to find the appropriate focal plane. Together these challenges act on sequencing data in the following ways:

- Lower Q30 Scores—Due to overloaded signal intensities, the ratio of base intensity to background for each base is decreased. This decrease often results in ambiguity during base calling, and leads to a decrease in data quality.
- Lower Clusters Passing Filter—The percentage of clusters passing filter (%PF) is an indication of signal purity from each cluster. Overclustered flow cells typically have higher numbers of overlapping clusters. This leads to poor template generation, which then causes a decrease in the %PF metric.
- Lower Data Output—Reduced yield (gigabases [Gb] per flow cell) is a byproduct of lower %PF.
- Inaccurate Demultiplexing—Index reads usually have low diversity by design, which can lead to poor base calling. Overclustering exacerbates the potential for poor base calling, which in turn, can lead to demultiplexing failures.
- Complete Run Failure—In cases of extreme overclustering, focusing can fail and the run may terminate at any cycle.

III. How to Diagnose Overclustering with Sequencing Analysis Viewer

Cluster density can influence many aspects of sequencing data. Therefore it is helpful to understand and recognize how overclustering can be detected using real-time run metrics. Several features in the Illumina Sequencing Analysis Viewer (SAV) can be used to diagnose overclustering.

a. The Analysis Tab

Data by Cycle: Intensity

If severe intensity drops occur in all channels early in the sequencing run, it can indicate poor template generation due to overclustering (Figure 1). The software would be unable to extract intensity information from subsequent images resulting in midrun failure.

Data by Cycle: % > Q30

Intensity drops and/or lower Q30 scores are the most common ways overclustering can be detected. Overclustering can affect either Read 1 or Read 2, but Read 2 is commonly more severely affected. This is because during paired-end (PE) chemistry, cluster sizes increase slightly due to extra cycles of amplification, which can lead to an increase in the number of overlapping clusters. With overclustered flow cells, this can affect run image registration and lead to poor Q30 scores and possible run failures (Figure 2).

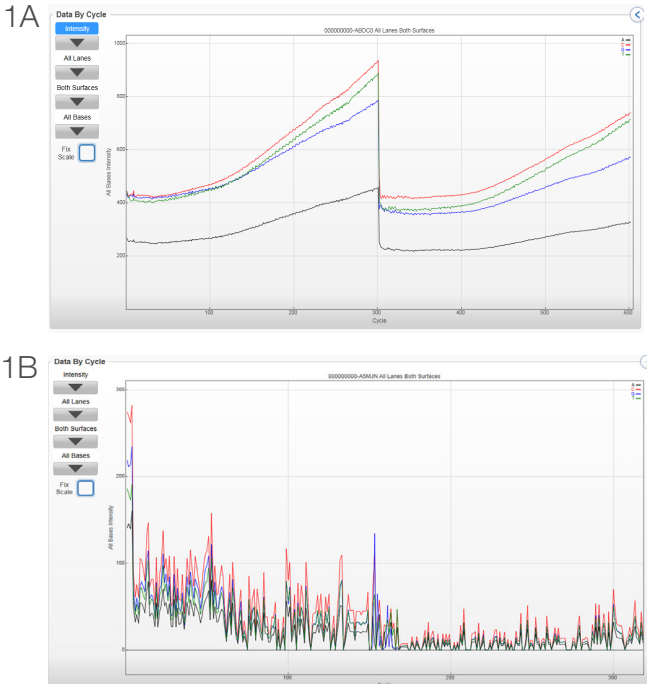

Figure 1: Data by Cycle: Intensity. A) Intensity profile from a normally clustered flow cell. B) Intensity plot shows midrun failure due to an overclustered flow cell.

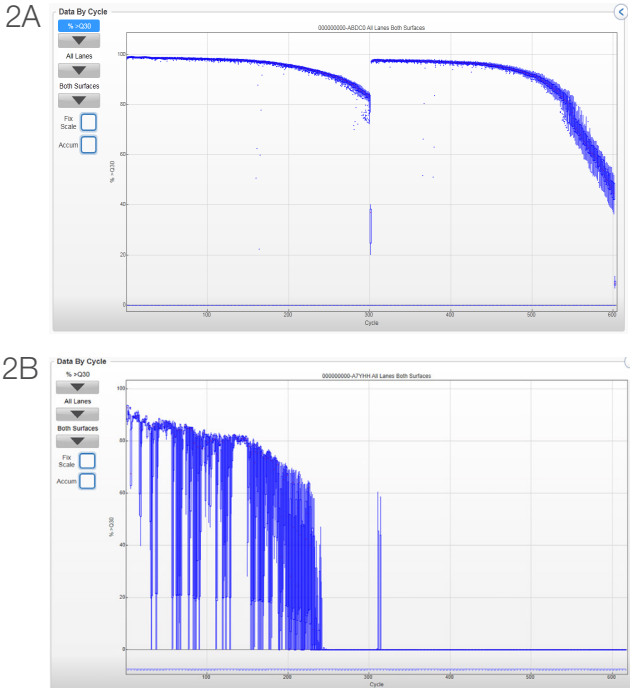

Figure 2: Data by Cycle: % > Q30. A) % > Q30 profile from a normally clustered flow cell. B) % > Q30 profile shows large standard deviations leading to a run failure due to an overclustered flow cell.

Data by Lane: Density

Data by Lane: Density box plots compare the raw cluster density to the %PF cluster density (Figure 3). With optimal density, the raw cluster density and %PF box plots appear close to one another (Figure 3A). As the cluster density increases beyond optimal density, the %PF decreases and the box plots appear further apart (Figure 3B). Also, clusters will not be identified correctly, which can result in underestimation of the raw cluster density. With severe overclustering, no clusters passed filter and the %PF plot is displayed as a green line at 0 density (Figure 3C).

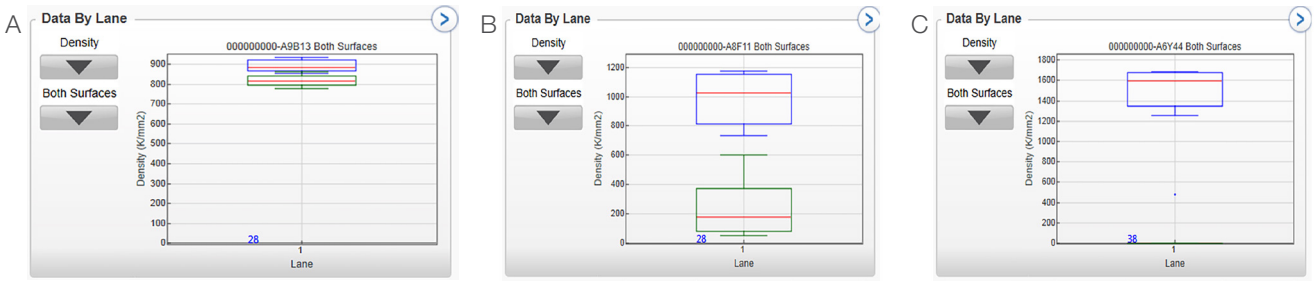

Figure 3: Data by Lane: Density. The blue boxes illustrate the raw cluster density range, the green boxes illustrate the %PF cluster density range, and the red lines indicate the median cluster density values. A) Optimal density. B) Overclustered. C) Severely overclustered.

Flow Cell Chart

Flow cell charts are useful for visualizing data per tile across the flow cell. Setting the first drop-down field for **Density PF** shows the range of cluster densities across all tiles of the flow cell. The figure legend (ie color scale) provides color coding for easy identification of density levels per tile. The color scale will change from run to run; therefore, it is important to reference the values in the scale when assessing cluster density. With optimal density, the legend displays cluster density values within the recommended range (Figure 4A, lane 2). Overclustered flow cell charts have tiles at the higher end of the color range and can include blue tiles (Figure 4A, lane 1). The blue color represents tiles that have low intensity or tiles that have "dropped out" (intensities = 0) due to image extraction failure. Setting the first drop-down field for **Intensity** is also helpful for evaluation of overclustering. Example flow cell charts indicating severe overclustering by intensity on the HiSeq® 2500 in rapid run mode (Figure 4B), the HiSeq 2500 in high output mode (Figure 4C), and the MiSeq® (Figure 4D) are shown. The blue or black tiles represent tiles that have intensities = 0 due to high cluster density. This is an indication that the flow cell is severely impacted.

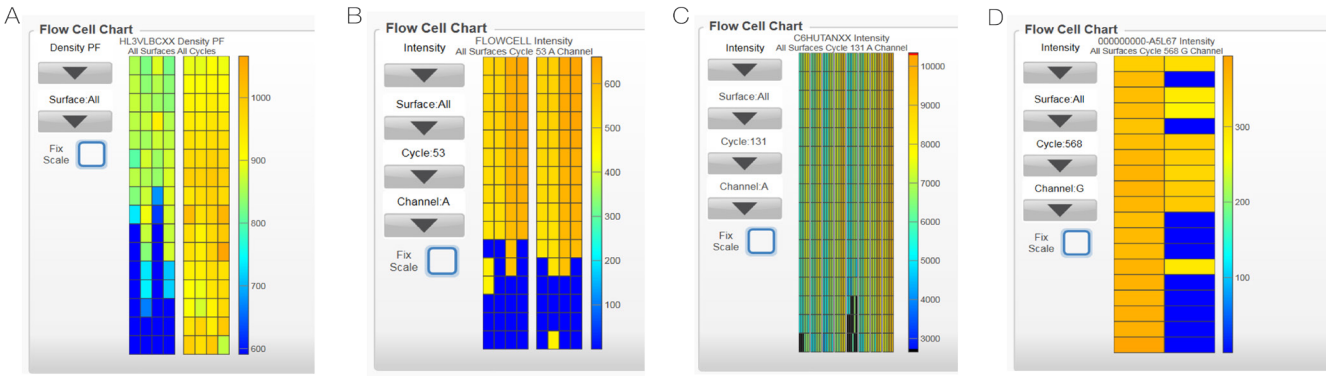

Figure 4: Flow Cell Charts. A) Density PF evaluation shows severe overclustering on the HiSeq 2500 in rapid run mode. Intensity evaluation shows severe overclustering on the HiSeq 2500 in rapid run mode (B), the HiSeq 2500 in high-output mode (C), and the MiSeq (D).

b. The Imaging Tab

Thumbnail Images

Thumbnail images are a powerful tool for troubleshooting run performance issues. A quick review of thumbnail images can reveal whether the flow cell is under- or overclustered (Figure 5).

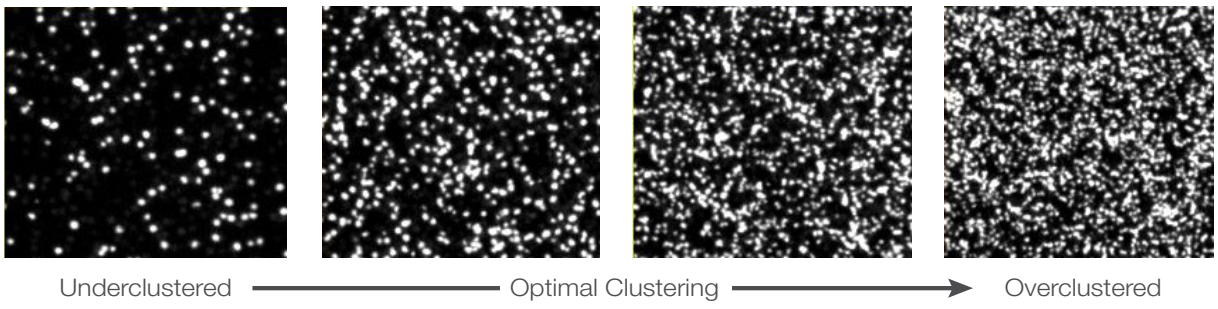

Figure 5: Thumbnail Images. Thumbnail images showing cluster densities ranging from underclustered to overclustered.

Imaging Tab Metrics Table

Issues with cluster registration can also be diagnosed with the **Imaging Tab Metrics Table.** For each tile, SAV generates a report of run metrics. The P90 A, C, G, and T metrics show the intensity values extracted from each cluster. With optimal clustering, they will reflect numeric intensity values. With overclustering, 0s and/or NaNs (not a number) can be reported in these fields even though clusters are visible in the thumbnail images. This is an indication that image extraction failed due to overclustering. Example run metrics tables for the MiSeq (Figure 6A) and the HiSeq 2500 (Figure 6B) are shown.

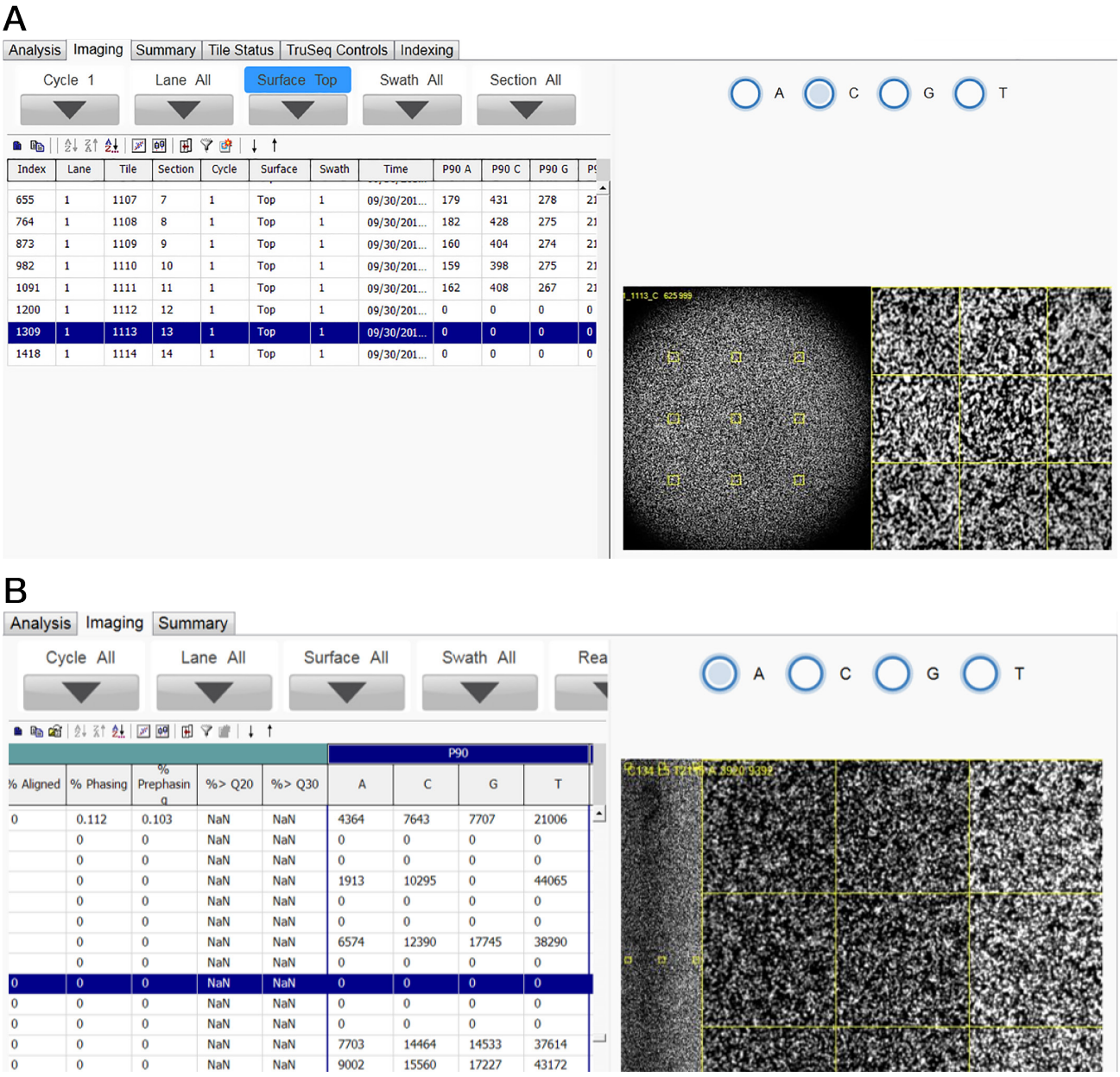

Figure 6: Run Metrics. (A) An example run metrics table for the MiSeq is shown. (B) An example run metrics table for the HiSeq 2500 is shown. "0" and "NaN" values indicate that the software was unable to extract intensity values.

c. The Summary Tab

After cycle 25 of Read 1, the following run metrics in the **Summary** tab can be used to check if the run is overclustered:

- **Clusters PF**—Reported percentages are below 80%. This value is not an official specification for this metric from Illumina; however, it can be used as a general guideline.
- **Raw Density**—Reported raw density value exceeds cluster density range recommended by Illumina.* Alternatively, reported raw density can be an underestimation of the true density due to saturation.
- **% Aligned**—If the reported percentage of reads aligning to PhiX is significantly lower compared to what was spiked-in, it is possible that the starting concentration of the libraries were underestimated, leading to overclustering.

Learn More

For more information, consult the "Sequencing Analysis Viewer (SAV) Software Guide."[1]

IV. Common Causes of Under- and Overclustering and Strategies for Prevention

a. Library Quality

Insufficient library clean-up steps can lead to the presence of library preparation contaminants such as adapter dimers, primer dimers, or partial library constructs that can greatly affect library quantification and subsequent clustering efficiency. Verification of all libraries for quality and purity using the Bioanalyzer or similar technology will check for library integrity, average insert size, and the presence of contamination. Detailed instructions regarding quality assessment with the Bioanalyzer are included in the Illumina library preparation user guides.

Sample Denaturation Considerations:

Preparing DNA samples for cluster generation requires denaturation with sodium hydroxide (NaOH). Special consideration should be given, as 2 factors can adversely affect cluster generation and density:

- NaOH used for sample denaturation must be freshly diluted with a pH > 12.5. NaOH that is not freshly diluted can acidify; the resulting decrease in pH will impair DNA denaturation and result in lower cluster density.
- Excess NaOH concentrations (> 1 mM) in diluted samples inhibits the formation of clusters.

For optimal cluster density, denature samples with freshly diluted NaOH (pH > 12.5) and make sure that the final concentration of NaOH in diluted samples is < 1 mM. If the diluted library must have a final concentration of NaOH > 1 mM, use Tris-HCl to neutralize the pH, as indicated in system specific denature and dilution guid[2].[2-5]

b. Library Quantification

The most common cause of under- and overclustering is inaccurate library quantification. While this section provides several recommendations for accurate library quantification, some library preparation protocols include a final bead-based normalization step and do not require additional quantification steps. The choice of library quantification methods will depend on the type of Illumina library preparation kit used and access to quantification resources.

Recommended Quantification Methods:

- **qPCR**—qPCR is the most effective method of library quantification when paired with a standard of similar size range. This technique is powerful as it measures only functional library fragments rather than all DNA species within a sample (eg primer dimers, free nucleotides, library fragments). However, this method may not be appropriate for quantifying libraries that have a broad insert size. For more information on library quantification with qPCR, see the "Sequencing Library qPCR Quantification Guide and Nextera® Library Validation and Cluster Density Optimization Technical Note."[6,7]

* Cluster density recommendations are listed on page 8.

- PicoGreen/Qubit System—Fluorometric systems such as PicroGreen or Qubit only measure dsDNA and are generally accurate. PicoGreen or Qubit Systems are the best option for libraries with a broad fragment size range. However, there is a risk of overestimating library concentration because this method measures all dsDNA in the pool, including partially constructed and adapter-dimer library contaminants. Assuming the Bioanalyzer quality assessment results show low levels of library contamination, PicoGreen/Qubit systems are highly accurate.
- Bioanalyzer— While the Bioanalyzer is recommended for quality control purposes, it should only be used for quantifying the following 3 types of libraries: TruSeq® Small RNA, TruSight® Tumor 26, and TruSeq Targeted RNA Expression libraries. The Bioanalyzer is not optimal for quantifying other libraries due to decreasing accuracy with increasing library fragment size range.

Quantification Methods to Avoid:

- NanoDrop/Spectrophotometry—Library quantification using spectrophotometry-based methods are subject to overestimation of library concentration and should be avoided, because single-stranded nucleic acids and free nucleotides are included in the library quantification.

c. Flow Cell Loading

The optimal raw cluster density specifications for balanced libraries differ by sequencer and reagent chemistry version. Table 1 lists the optimal raw density for each Illumina sequencing system.

Table 1: Optimal Raw Densities for Illumina sequencing systems

	MiniSeq™	MiSeq		NextSeq®	HiSeq 2500 Rapid Run (RR)	HiSeq 2500 High Output (HO)	
Versions	High and Mid Output	v2	v3	v2 High and Mid Output	v1 and v2	v3	v4
Raw Density (K/mm^2)	170–220	1000–1200	1200–1400	170–220	850–1000	750–850	950–1050

Considerations for Cluster Density on Patterned Flow Cells

The HiSeq X™ Ten and HiSeq 3000/HiSeq 4000 Systems use patterned flow cells, which consist of a nanowell substrate with billions of ordered wells. The uniform cluster sizes enable optimal spacing and increased cluster density. In fact, overclustering is not possible on patterned flow cells. However, special consideration must be given to loading patterned flow cells, as suboptimal loading will negatively influence %PF and resulting sequencing data:

- Underloading a patterned flow cell will result in low %PF and *increased* duplicate reads (Figure 7, green triangles).
- Overloading a patterned flow cell will result in mixed clusters that will not pass quality filters during chastity analysis, resulting in lower %PF (Figure 7, purple circles).

Learn More

For more information, consult the “Calculating Percent Passing Filter for Patterned and Non-Patterned Flow Cells Technical Note.”[8]

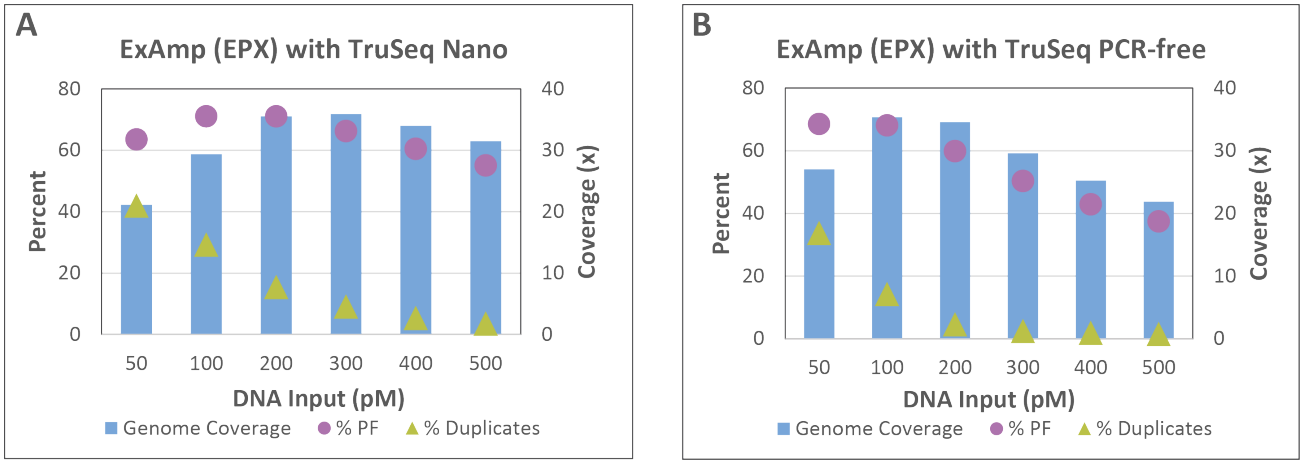

Figure 7: Sample Loading on Patterned Flow Cells. Genome coverage (blue bars), %PF (purple dots), and % duplicates (green triangles) are plotted as a function of increasing DNA input (pM) for a TruSeq Nano Library (A) and a TruSeq PCR-Free Library (B). Low input causes underclustering that results in low coverage, low %PF, and high % duplicates. High input reduces % duplicates and increases coverage, but causes reductions in %PF due to mixed clustering.

d. Library Nucleotide Diversity

Overlooking the nucleotide diversity of a library is another common cause of flow cell overloading. This section provides a general description of nucleotide diversity and recommendations for mitigating low diversity issues.

What is nucleotide diversity?

The term "nucleotide diversity" describes the proportion of each nucleotide (A, T, C, and G) at each position in a sequencing library. A "balanced" or "diverse" library has *equal proportions* of A, C, G, and T nucleotides at each base position in a sequencing library. Cluster images and SAV data by cycle plots from diverse/balanced, low-diversity, and unbalanced libraries are shown (Figure 8).

Nucleotide diversity is required for effective template generation on Illumina sequencing systems and is important for the generation of high-quality data. In particular, nucleotide diversity is key during the first 4–7 cycles of Read 1 for all Illumina sequencing systems. The sequencing software uses images from these early cycles to identify the location of each cluster in a process called template generation. Nucleotide diversity is also important for the first 25 cycles for %PF evaluation. These factors are used in base calling and quality score calculations.

Mitigations for Low Nucleotide Diversity Libraries

It is important to keep in mind that the cluster density recommendations provided by Illumina assume the library is evenly balanced. Illumina recommends applying the following mitigations to low-diversity libraries to improve software performance and sequencing data accuracy:

- Reduce template input when running an unbalanced library. The amount of reduction required must be empirically determined.
- Spike in a high-diversity library, such as PhiX, to increase the overall nucleotide diversity of the sample. The amount of spike-in required must be empirically determined, but Illumina recommends the following:
  - For current versions of HiSeq Systems, spike in at least 10% PhiX
  - For current versions of MiSeq Systems, spike in at least 5% PhiX
  - For current versions of NextSeq and MiniSeq Systems, spike in 10–50% PhiX

High Base Diversity Libraries

While there is no need to worry about providing a balanced signal with high-diversity libraries, Illumina recommends spiking in 1% PhiX as a positive sequencing quality control.

Sequencing a New or Unknown Library

If a library is new or has unknown nucleotide diversity, choose a more conservative loading concentration (ie target the lower end of the recommended cluster density range). Spiking in a system-specific amount of PhiX is also good practice with libraries of unknown nucleotide diversity.

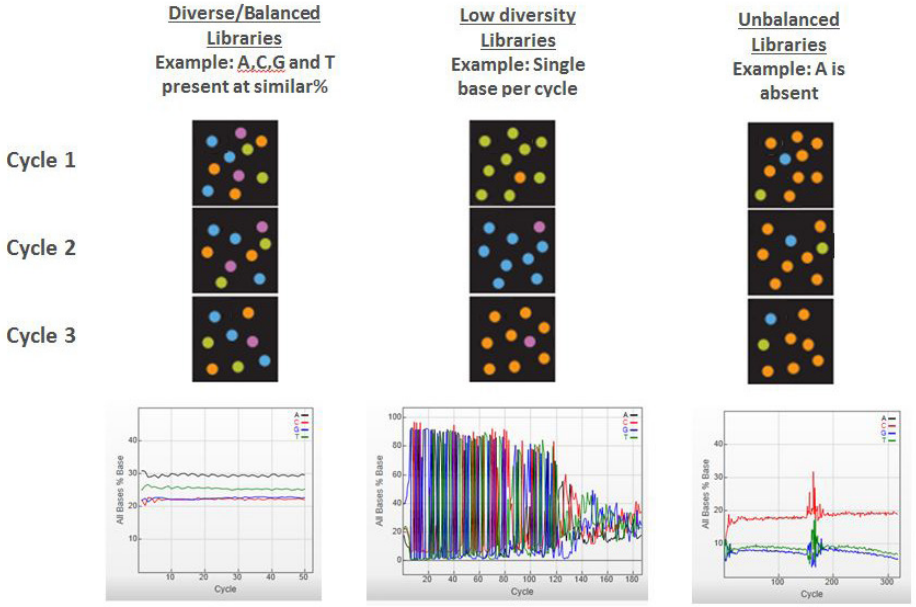

Figure 8: Nucleotide Diversity and Data by Cycle: % Base Plots. A) Diverse/balanced libraries contains equal proportions of A, T, C, and G nucleotides. The Data by Cycle: % Base plot shows even, horizontal curves centered on 25%. B) Low-diversity libraries, such as PCR amplicon, metagenomic, enriched/targeted, or ChIP libraries, have an uneven proportion of nucleotides across the flow cell from one cycle to the next. The Data by Cycle: % Bases plot shows large intensity spikes at each cycle. C) Unbalanced libraries, such as bisulfite converted libraries, have one base at a much lower percentage than the others. This example shows a library with a low percentage of the "A" nucleotide.

V. Summary

Cluster density optimization is a critical step for the generation of high-quality data and high yield runs. Under- and overclustering can be largely avoided by taking precautions early on, such as careful library quantification (using recommended quantification methods), assessing the library quality, and performing mitigations for low diversity libraries. Empirical testing may be needed to determine the best library concentration range to meet the Illumina recommended cluster density range. When the run is in progress, understanding how overclustering affects SAV run metrics, thumbnail images, and summary data allows real-time monitoring of run quality. Diagnosing overclustering early in the run, aborting runs when necessary, or quickly identifying the root cause of run failures will save valuable time and effort, allowing researchers to meet their goals with greater speed and efficiency.

VI. Glossary

Image Registration: Intensity values are extracted during a process called “image registration.” During image registration, fluorescence is converted to a numeric intensity value and assigned to an X, Y-position. This “map” of cluster positions is applied to the entire run.

Passing Filter: During cycles 1–25 of Read 1, the chastity filter removes the least reliable clusters from the image extraction results. Clusters “pass filter” if no more than 1 base call has a chastity value below 0.6 in the first 25 cycles. Chastity is defined as the ratio of the brightest base intensity divided by the sum of the brightest and the second brightest base intensities.[4]

Template Generation: Template generation involves analyzing images from the first 4–5 cycles of a run to map the location of each individual cluster. This is referred to as the “cluster template” or “cluster map.” By detecting clusters from multiple cycles, there is a better chance of resolving overlapped clusters.

VII. References

1. Illumina (2016) Sequence Analysis Viewer v1.11 software guide. (support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav/documentation.html)
2. Illumina (2016) MiniSeq system denature and dilute libraries guide. (support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miniseq/miniseq-denature-dilute-libraries-guide-1000000002697-00.pdf)
3. Illumina (2016) MiSeq system denature and dilute libraries guide. (support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/miseq-denature-dilute-libraries-guide-15039740-01.pdf)
4. Illumina (2015) NextSeq system denature and dilute libraries guide. (support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/nextseq/nextseq-denature-dilute-libraries-guide-15048776-02.pdf)
5. Illumina (2016) HiSeq and GA$_{IIx}$ systems denature and dilute libraries guide. (support.illumina.com/downloads/hiseq-denature-dilute-libraries-guide-15050107.html)
6. Illumina (2011) Sequencing library qPCR quantification guide. (support.illumina.com/downloads/sequencing_library_qpcr_quantification_guide_11322363.html).
7. Illumina (2013) Nextera library validation and cluster density optimization technical note. (www.illumina.com/documents/products/technotes/technote_nextera_library_validation.pdf).
8. Illumina (2015) Calculating Percent passing filter for patterned and nonpatterned flow cells technical note. (support.illumina.com/content/dam/illumina-marketing/documents/products/technotes/hiseq-x-percent-pf-technical-note-770-2014-043.pdf)

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

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

illumina®