Low-Diversity Sequencing on the Illumina HiSeq® Platform

HCS v2.2.38 facilitates low-diversity template generation and data analysis.

Introduction

Nucleotide diversity—having equal proportions of A, C, G, and T nucleotides at each base position in a sequencing library—is required for effective template generation on Illumina sequencing platforms. For this reason, most Illumina library preparation workflows include a random fragmentation step to create the necessary sequence diversity at each base position in the library. Libraries generated without random fragmentation, or produced through amplicon generation, have traditionally been challenging to sequence. Data from these libraries tend to have lower yields and quality scores compared to more nucleotide-balanced libraries constructed from genomic DNA. To increase nucleotide diversity and produce high-quality data, low-diversity libraries often require library preparation modifications that include pooling together a number of uniquely bar-coded samples (multiplexing), spiking in a high-diversity library such as PhiX, and/or designating a control lane with the HiSeq Control Software (HCS). See *Using a PhiX Control for HiSeq Sequencing Runs (Pub. No. 770-2011-041).*

This technical note describes improvements to HCS that eliminate the need for library preparation modifications and the need for a designated control lane. Additionally, results from the Illumina internal validation testing are provided, supporting a method for generating high-quality performance scores from low-diversity libraries.

Improvements in RTA v1.18.61 Software

Recently, Illumina released HCS v2.2.38, which includes updates to the real-time analysis (RTA) software v1.18.61. These updates are modeled after changes to the MiSeq® RTA March 2013 release that significantly improved performance with low-diversity samples. See *Low-Diversity Sequencing on the Illumina MiSeq Platform (Pub. No. 770-2013-013).* HCS v2.2.38 includes optimization of the color normalization matrix estimation, phasing, and prephasing rates. These updates improve the ability of RTA to analyze samples with low diversity such as in amplicon or bisulfite converted samples, and those with unbalanced genomic composition such as AT- or GC-rich genomes (Figure 1).

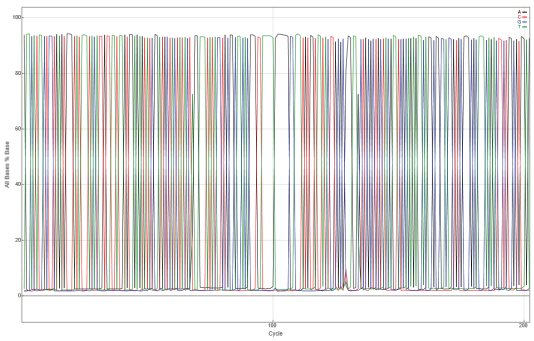
The following features have been added to HCS v2.2.38 and applied to all HiSeq run modes.

Color Matrix Estimation

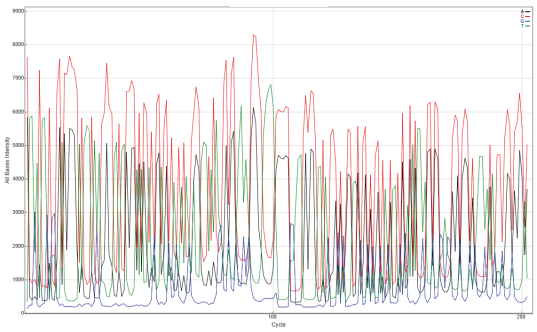
RTA v1.18.61 now uses the first 11 cycles of sequencing data for color matrix estimation, and the same matrix is used to correct all reads. Because nucleotide distribution is critically important during the first few sequencing cycles for cluster identification and template generation, HCS will continue to use the first four cycles of data (or five in the case of v4 runs) to calculate the color matrix. After template

Figure 1: SAV Data from Low-Diversity Samples

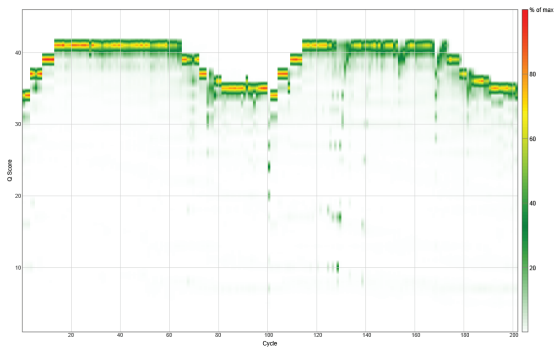
A. Percent Base



B. Intensity



C. Q-Score Plot



Standard quality metrics generated from an internal run of a single amplicon sample demonstrate improved ability to call unbalanced samples with high quality.

Table 1: Standard Quality Metrics with HCS v2.2.38

Sample Type	Run Length (bp)	Cluster Density[b] (K/mm²)	% PF	% Q30	% PhiX
TruSeq SBS v3					
Bisulfite	2 × 101	587	92.8	89.3	5.3
Bisulfite	2 × 101	602	92.6	89.8	7.3
Internal Constant Bases[a]	2 × 101	461	92.1	86.1	0.0
Single Amplicon	2 × 101	557	92.7	94.2	7.0
HiSeq SBS v4					
Bacillus cereus (65% A/T)	2 × 126	1,031	95.9	93.8	0.0
Rhodobacter sphaeroides (69% G/C)	2 × 126	896	94.1	85.9	0.0

Illumina internal sequencing runs using HCS v2.2.38 with both TruSeq® SBS v3 and HiSeq SBS v4 chemistries demonstrate that low-diversity samples yield high-quality scores with a spike-in of 10% PhiX.

a. Diverse library with internal five base-pair monosequence after nine cycles.

b. For optimal cluster densities, see *Denaturing and Diluting Libraries for the HiSeq and GAIIx (part #15050107).*

generation is complete, the initial matrix is discarded and the first 11 cycles of intensity data are used for final matrix estimation.

Empirical Phasing Correction Algorithm

During sequencing by synthesis (SBS), each DNA strand in a cluster extends by one base per cycle. A small proportion of strands may become out of phase with the current cycle, either falling a base behind (phasing) or running a base ahead (prephasing). The phasing and prephasing rates define the fraction of molecules that become phased or prephased per cycle. Calculation of these rates requires a balanced and random base composition. A new empirical phasing correction is now included in the software to accommodate low-diversity samples. For each cycle of sequencing, unique phasing corrections are calculated to maximize data quality. The reported phasing shown in Sequence Analysis Viewer (SAV) is the tile median slope of the observed phasing corrections for cycles 1–25.

Control Lane Eliminated

With these HCS software improvements, a dedicated control lane is no longer required to estimate matrix and phasing. Internal testing demonstrated excellent performance using a 10% PhiX spike-in with a range of genomes and several single amplicon libraries (Figure 1c). Therefore, the option to designate a control lane has been removed from HCS v2.2.38 in all run modes including TruSeq v3, HiSeq v4, and Rapid Run mode, allowing the lane to be used for additional samples.

Evaluating Low-Diversity Samples

While HCS v2.2.38 offers significant improvements, Illumina recommends spiking in a percentage of PhiX control DNA with low-diversity samples to increase the library nucleotide balance and to make clusters easier for the software to identify.

Note: This option will not work for *de novo* sequencing or sequencing organisms with high homology to PhiX.

To validate software changes in RTA primary analysis and to assess the performance of low-diversity runs, Illumina carried out rigorous internal testing. Table 1 summarizes the sequencing data generated with RTA v1.18.61, demonstrating that low-diversity samples yield high-quality results and quality scores with a spike-in of ≥ 10% PhiX. These results were accomplished without modifying the library preparation protocol or software. Higher amounts of PhiX do not negatively affect the sequencing quality, and may be necessary for some libraries.

Summary

Using RTA v1.18.61, Illumina has successfully demonstrated that the HiSeq platform can support sequencing of low-diversity samples without a control lane or modified library preparation. Experimental quality was validated by alignment and variant calling of known samples. Runs that generated poor data quality with earlier versions of RTA, delivered high-quality results with RTA v1.18.61. When running at supported cluster densities and with PhiX spike-in, bisulfite, single-amplicon, and high and low GC-content samples of varying run lengths generated high Q-scores. The best results were achieved with a PhiX spike-in of at least 10%, improving both software performance and sequencing accuracy. Although the amount of PhiX spike-in can vary with sample and library type, Illumina recommends 10% as a starting point when sequencing low-diversity samples using the HiSeq Systems.

Learn More

To make sure your HiSeq system meets the requirements for the HCS v2.2.38 and RTA v.1.18.61 updates, review the software release and install notes at www.support.illumina.com/downloads/hcs-2-2-38-software.html before installation.

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

FOR RESEARCH USE ONLY

Pub. No. 770-2014-035 Current as of 20 August 2014

illumina®