Using a PhiX Control for HiSeq® Sequencing Runs

A low-concentration spike-in of Illumina PhiX Control v3 provides quality and calibration controls.

Highlights

- **For most libraries, Illumina recommends using PhiX Control v3 (Catalog # FC-110-3001) in a low-concentration spike-in (1%) for improved sequencing quality control**
- **For samples with bias base composition, use a dedicated PhiX control lane to improve cross-talk and phasing calculations**
- **For samples with low diversity, use a dedicated control lane and a high-concentration spike-in (40% or higher) of PhiX to create a more diverse set of clusters**

What are PhiX Control Libraries?

Control libraries generated from the PhiX virus serve as an effective control in sequencing runs. Characteristics of the PhiX genome provide several benefits:

- **Small**—PhiX is a small genome, which enables quick alignment and estimation of error rates.
- **Diverse**—The PhiX genome contains approximately 45% GC and 55% AT.
- **Well-Defined**—PhiX has a well-defined genome sequence.

Illumina cluster detection algorithms are optimized around a balanced representation of A, T, G, and C nucleotides. Illumina PhiX Control v3, with an insert size of 425–525 bp, is a balanced and diverse library that can help mitigate sequencing challenges in unbalanced and low diversity libraries.

Benefits of Using a PhiX Control

PhiX libraries provide a quality control for cluster generation, sequencing, and alignment, and a calibration control for cross-talk matrix generation, phasing, and prephasing.

Cluster Generation

Illumina PhiX Control v3 is shipped as a ready-to-use 10 nM library that can be used as a positive control in the clustering process. If a problem occurs in sample preparation, PhiX will still generate clusters. This helps you discern whether a lack of clusters is due to sample preparation failure or a failure in the cluster generation process.

Cross-Talk Matrix Generation

During an Illumina sequencing run, the cross-talk due to spectral overlap between the four fluorescently-labeled nucleotides is calculated during template generation (cluster identification) in cycles 1–4.

For proper cross-talk calculation, HiSeq Control Software (HCS) v1.3.8, or higher, requires approximately equal numbers and at least 50,000 clusters/mm^2 in each of the four bases. Therefore, it is imperative to have a balanced representation of bases at the beginning of each read, excluding the Index Read.

Phasing and Prephasing

During sequencing by synthesis, 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 jumping 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 in cycles 2–12. Any library that does not comprise a balanced base composition (e.g., initial cycle indexing, restriction enzyme libraries) should use a control lane.

High GC samples typically show higher phasing rates. However, if the sample has good diversity (for example, whole-genome sequencing libraries with 40–60% GC), it does not require a control lane.

Alignment

Because PhiX has a small, well-defined genome sequence, it is an excellent alignment control. If a lane containing PhiX (control lane or low-concentration spike-in) is designated for your run, Real Time Analysis (RTA) software aligns complete sequences to the PhiX reference beginning after the 25th cycle is accumulated and calculates error rates, providing an indication of sequencing success during the run.

Using a PhiX Control for Unbalanced Samples

Most mammalian genomic or whole-transcriptome RNA samples have a balanced genomic composition (approximately equal proportions of A, T, G, and C). These samples do not require a dedicated control lane to generate accurate matrix and phasing estimations. Unbalanced samples contain genomes with high AT or GC content (less than 40% or greater than 60%). Some examples are Arabidopsis, Plasmodium, some bacteria, and bisulfite conversion studies.

Use a dedicated control lane with unbalanced samples to improve cross-talk and phasing calculations.

Using a PhiX Control for Low Diversity Libraries

Low diversity libraries are libraries where a significant number of the reads have the same sequence. This shifts the base composition because the reads are no longer random.

Low diversity can occur with some expression studies with > 25%

one type of transcript, amplicon pools, adapter dimer, and initial cycle indexing, for example. Low diversity libraries require the use of a dedicated control lane for matrix and phasing/prephasing calculations ***and*** might require a high concentration spike-in to help balance the overall lack of sequence diversity.

Strategies for Low Diversity Libraries

Use the following strategies when sequencing low diversity libraries on the HiSeq:

- **Reduce Cluster Density**—The optimal cluster density specified for sequencing runs assumes that the lane contains a diverse library. Template generation (cluster identification) algorithms detect clusters across multiple sequencing cycles, which resolves overlapping clusters when their sequence differs between cycles. This advantage is lost when clusters have the same sequence throughout template generation cycles. Reducing the cluster density by 50–80% when sequencing low diversity libraries reduces the number of overlapping clusters. Use a PhiX control lane for matrix and phasing calculations to ensure the presence of a balanced genomic library.
- **High-Concentration PhiX Spike-In**—Spike-in a PhiX control at about 40–50% of the total library; the percentage must be empirically determined. This should be combined with a reduction of the amount of library used so that the total cluster density does not exceed Illumina recommendations. This method provides the appropriate clusters/mm^2 of each fluor required for matrix generation. Adding less PhiX (e.g. 1% spike-in) can determine an alignment error rate, but will not optimize diversity for matrix generation, phasing, and prephasing calculations.
- **Concatamerize and Fragment Amplicons**—When sequencing PCR amplicons or other samples where diversity is low for the first 12 cycles, concatamerize the amplicons and fragment to create sequencing diversity. Alternatively, you can increase the probability of an equal representation of A, T, G, and C nucleotides by adding custom primers with a random sequence onto the amplicons on both the 5' and 3' ends. Other options include indexing your libraries and sequencing them with other diverse libraries, such as PhiX, or to control the pooling of different amplicons such that the first few bases fit the balancing requirements.

How To Use a PhiX Control

You can apply a PhiX library to your HiSeq sequencing run in the following three ways:

- **Low-concentration spike-in of 1%**—For most libraries, Illumina recommends a low-concentration spike-in as an in-lane positive control for alignment calculations and quantification efficiency.
- **High-concentration spike-in of 40% or higher**—Reduce the cluster density of your library by 40–50%, and then use a high-concentration spike-in to create a more diverse set of clusters for matrix, phasing, and prephasing calculations.
- **Dedicated control lane**—In specific cases of an unbalanced genome or low-diversity libraries, use a dedicated control lane for matrix and phasing calculations. PhiX v3 provides an appropriate source for calculating phasing and prephasing values. Other libraries with high diversity and balanced genome composition (e.g. mammalian whole-genome sequencing) can also serve as a control lane.

Specifying PhiX Control Options Using HCS

You can specify PhiX for lane alignment and error rate calculations in the run setup steps. From the Scan screen in HCS, locate the control lane option. From the Control Lane drop-down list, select one of the following:

- Select the number of the lane containing the PhiX control, or the lane containing another balanced and diverse library. With this setting, the cross-talk matrix, phasing, and prephasing are calculated from the control lane and applied to the whole flow cell.
- Select **None** to indicate that you are not using a dedicated control lane. With this setting, calibration calculations are made on a per-lane basis from all clusters in each lane. This setting is appropriate if you have a diverse library in every lane, or if you are using a PhiX spike-in and plan to use PhiX to generate error rates.

For more information about using HCS to setup a sequencing run, see the *HiSeq 2000 User Guide*, Part # 15011190.

Downstream Analysis of PhiX v3 Library Reads

PhiX v3 is not indexed, but can be used with other indexed libraries. The non-indexed reads will be parsed into a folder named Unknown or Undetermined_Indices (depending on the CASAVA version) along with any indexed reads where the index could not be identified. The PhiX library can be further processed by downstream alignment.

Illumina, Inc. • Headquartered in San Diego, California, U.S.A. • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2012 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 770-2011-041 Current as of 01 February 2012

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA