Sequencing Library QC on the MiSeq™ System

Pre-configured, automated QC workflows for cost-effective measurement of library quality prior to a large-scale sequencing study.

Introduction

Increasing yields from the HiSeq® 2000 sequencing system have enabled the largest and most complex genomic studies to date. To maximize the efficiency of high-throughput sequencing, it is important to know the quality of the starting library. A poor quality library can undermine the success of large-scale sequencing, and lead to costly and time-consuming repeat experiments. Performing library quality control (QC) using the MiSeq system before committing it to a full-scale HiSeq run can save time and money, while leading to better sequencing results.

Using a single sample prep method and taking only a single day, detailed QC parameters, including cluster density, library complexity, percent duplication, GC bias, and index representation can be generated on the MiSeq system. The MiSeq system has the unique ability to do paired-end (PE) sequencing for accurately assessing insert size. Library cluster density can also be determined and used to predict HiSeq cluster density, maximizing yield and reducing rework. This application note describes a library QC run performed on the MiSeq system.

QC Metrics

Using the Illumina Experiment Manager, samples within the library were defined in the sample sheet, and the samples loaded into the MiSeq reagent cartridge for automated cluster generation and sequencing. After a 2 × 26 bp run, an automated Library QC report was generated on the instrument using the MiSeq Reporter software. The Library QC workflow does not perform variant calling, but provides a range of important QC metrics that pertain to library quality and sequencing performance (Table 1 and Figure 1A). A Summary tab (Figure 1B) provides information about the run parameters, depicting graphs of cluster data, high and low values, and mismatches. In addition, a sample tab provides coverage statistics showing the number of aligned reads that cover each position in the reference sequence, and quality score data for each sample (data not shown). Users can regularly monitor these metrics and develop a standard set of criteria that can guide future high-throughput sequencing runs.

Summary

Assessing the quality of a sequencing library before committing it to a full-scale sequencing run ensures maximum sequencing efficiency, leading to accurate sequencing data with more even coverage. Because the MiSeq system uses the same sample prep methods and proven sequencing by synthesis chemistry as the HiSeq system, it is ideal for analyzing prepared libraries in a single day prior to performing high-throughput sequencing. With automated reports available through the MiSeq Reporter, all quality metrics are easily accessible and exportable. Data generated by the MiSeq system is comparable to other Illumina next-generation sequencing platforms, ensuring a smooth transition from one instrument to another. Based on your individual experimental requirements, metrics obtained from performing simple QC can be used to streamline and improve your sequencing projects.

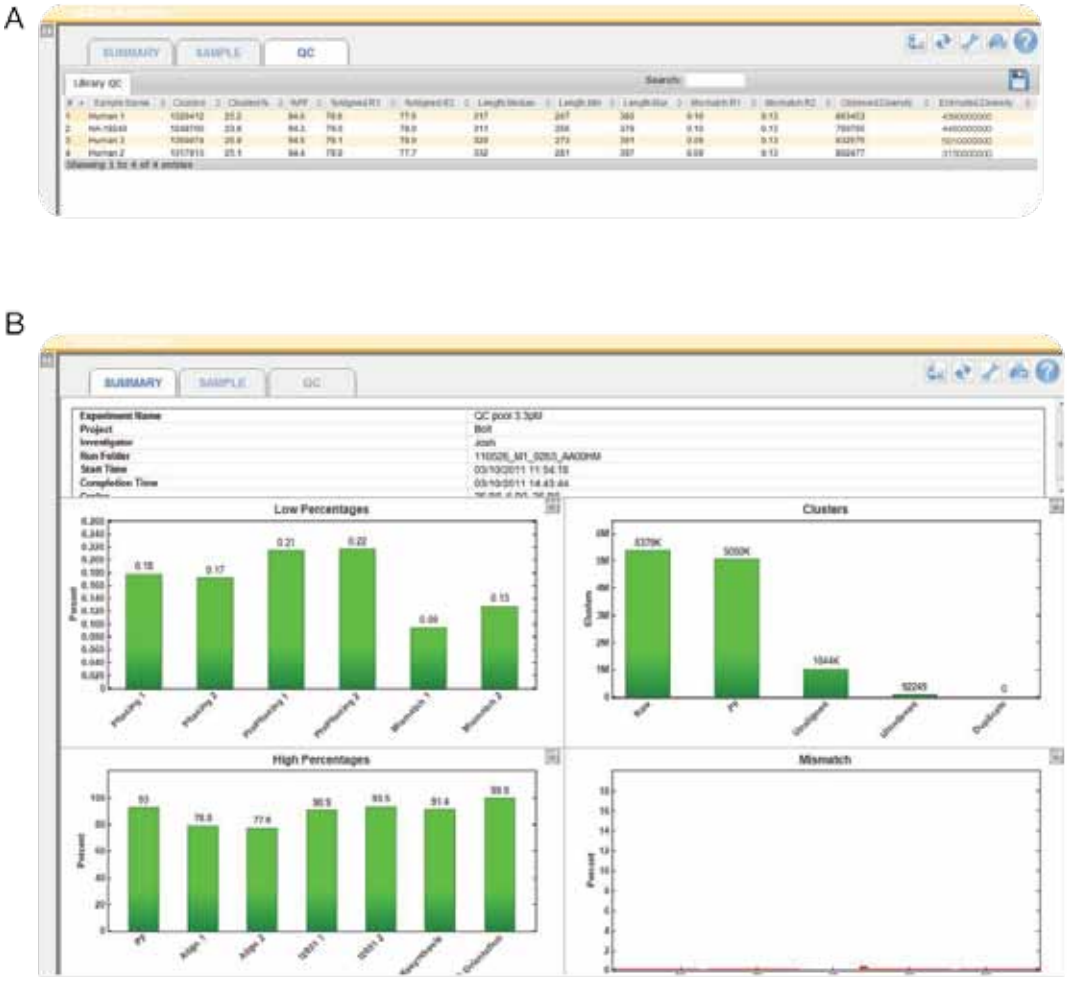
Table 1: Sequencing Library QC Metrics

Metric	Definition
Sample Name	Sample name from the sample sheet
Clusters	Number of clusters sequenced for this sample
Clusters %	Percentage of successfully indexed clusters from this sample
% PF	Percentage of clusters for this sample that passed filters
% Aligned R1	Percentage of clusters for which read 1 successfully aligned
% Aligned R2	Percentage of clusters for which read 2 successfully aligned
Length Median	Median fragment length for this sample
Length Min	Low percentile (corresponding to 3 standard deviations from the median) of fragment lengths for this sample
Length Max	High percentile (corresponding to 3 standard deviations from the median) of fragment lengths for this sample
Mismatch R1	Mismatch rate for this sample in read 1
Mismatch R2	Mismatch rate for this sample in read 2
Observed Diversity	Number of distinct aligned positions (x, y) for read 1 and read 2. Reads with the same aligned positions may represent PCR duplicates of the same source oligonucleotide.
Estimated Diversity	Estimate of the total library diversity derived from the observed diversity and the number of apparent PCR duplicates

Learn More

Go to www.illumina.com/miseq to learn more about the next revolution in personal sequencing.

Figure 1: Automated Library QC Reports Generated on MiSeq Reporter



A. Quality Control (QC) Tab showing quality metrics for the sequencing library, including sample name (column one), abundance (columns two through six), fragment length (columns seven through nine), mismatch rate (columns ten through twelve), and library diversity (last two columns).
B. Summary Tab showing high-level run information including low run statistics (upper left graph), cluster performance (upper right), high run statistics (lower left), and mismatches between reads and the reference sequence (lower right).

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2011 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc.
All other brands and names contained herein are the property of their respective owners.
Pub. No. 770-2011-014 Current as of 06 October 2011

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA