Illumina Sequencing and Data Analysis

1 Hiseq First Run Report at PICB Omics Core

The first and latest run at PICB Omics Core hiseq platform has been done successfully. The data summary information listed as below.

Lane	Level	Reads(M)	Reads PF(M)	Reads PF(%）	% >= Q30	Error Rate 35 cycle(%)	Error Rate 100 cycle(%)	Aligned (%)
Lane8（C）	1*100	217.04	204.17	94.07	94.1	0.14	0.47	98.7
Lane1	1*100	235.78	221.05	93.75	90.9	0	0	0
Lane2	1*100	271.69	249.27	91.75	89.7	0	0	0
Lane3	1*100	303.18	238.18	78.56	87.1	0	0	0
Lane4	1*100	307.15	204	66.42	85.6	0	0	0
Lane5	1*100	312.07	178.38	57.16	84.8	0	0	0
Lane6	1*100	310.21	165.81	53.45	84.6	0	0	0
Lane7	1*100	306.71	192.14	62.65	85.1	0	0	0

Table 1: Data summary information for eight lanes on our first run (PhiX as control lane)

Our first single read 100 nt sequencing run generated about 200M 1x100 nt reads for PhiX control itself, which is about 20G bases. The error rate is also very low, at 0.14% and 0.47% for 35 and 100 cycles respectively. Read quality is very high, over 94% passed Q30 for PhiX control lane and about 85% for all the eight lanes.

For eight lanes shown as Figure 1 as below, we got from 160 to 250 M reads (which passed filter) meaning 16 to 25G bases.

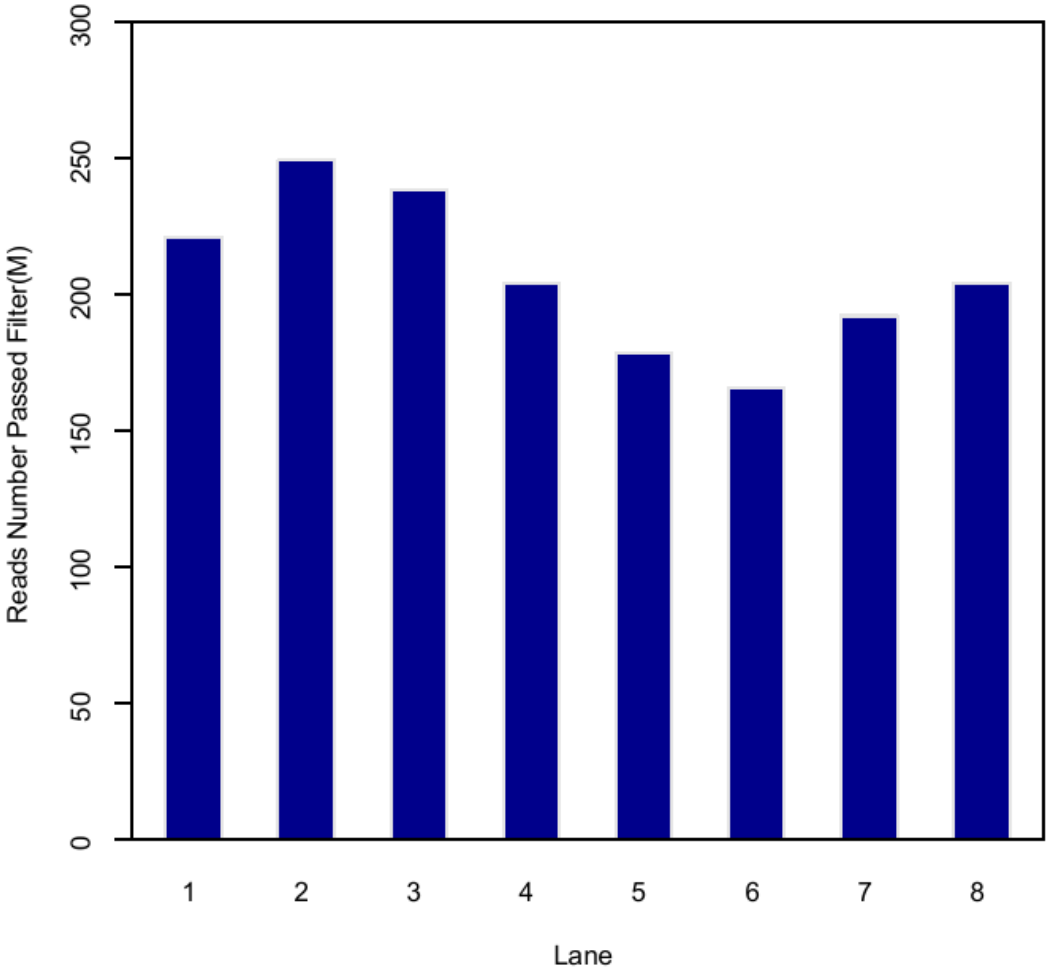

Figure 1: Reads passed filter for each of eight lanes on our first run

2 Illumina Sequencing Overview

2.1 Illumina Sequencing Workflow

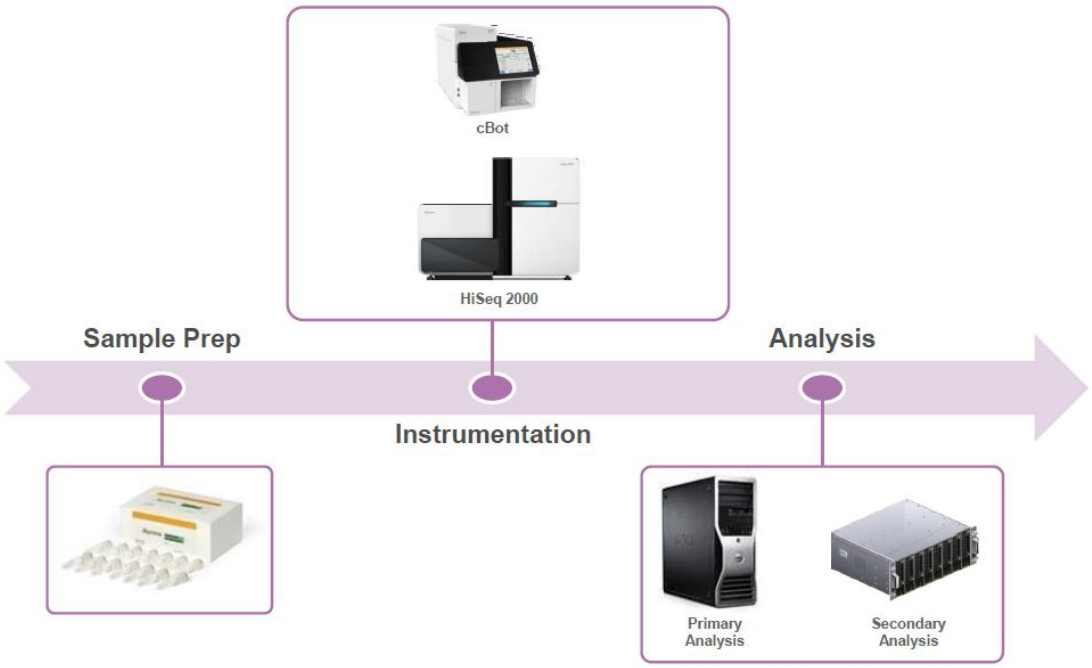

Figure 2: Illumina Sequencing Workflow

2.2 Illumina Sequencing Workflow Outcomes

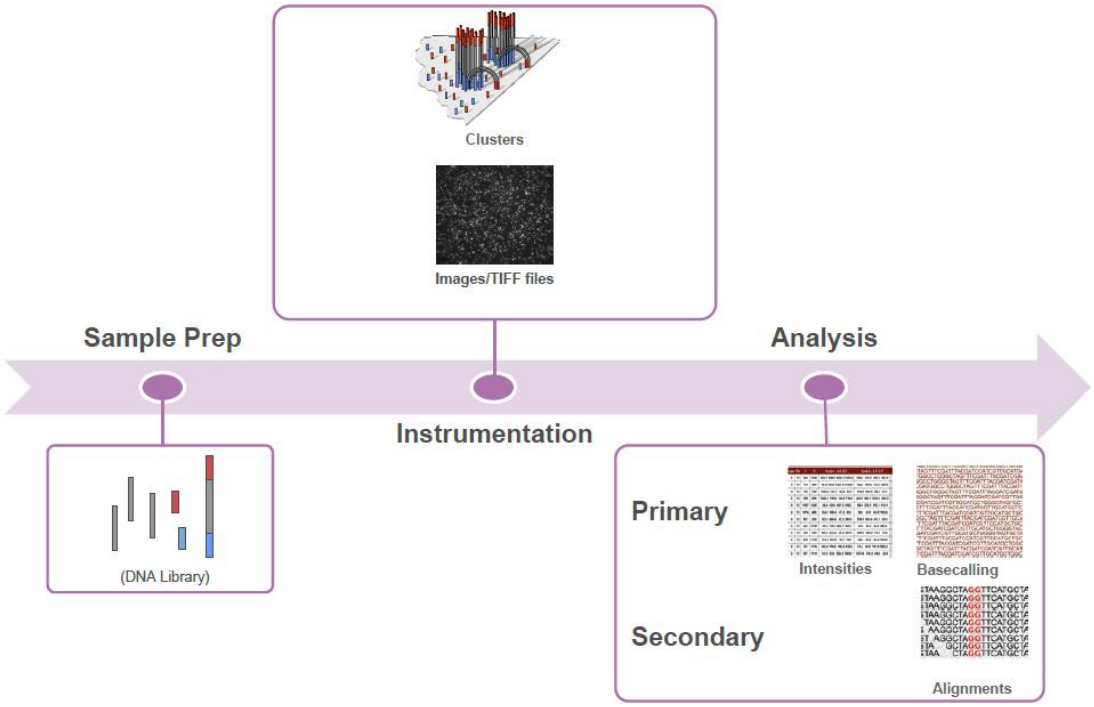

Figure 3: Illumina Sequencing Workflow Outcomes

3 Analysis of Sequencing Data

3.1 Analysis of Sequencing Data Overview

After the sequencing platform generates the sequencing images, the data are analyzed in five steps: image analysis, base calling, bcl conversion, sequence alignment, and variant analysis and counting. CASAVA performs the bcl conversion, sequence alignment, andvariant analysis and counting steps, demultiplexes multiplexed samples during the bcl conversion step.

1 **Image analysis**—Uses the raw images to locate clusters, and outputs the cluster intensity, X,Y positions, and an estimate of the noise for each cluster. The output from image analysis provides the input for base calling. Image analysis is performed by the instrument control software.

2 **Base calling**—Uses cluster intensities and noise estimates to output the sequence of bases read from each cluster, a confidence level for each base, and whether the read passes filtering. Base calling is performed by the instrument control software's Real Time Analysis (RTA) or the Off-Line Basecaller (OLB).

3 **Bcl conversion**—Converts *.bcl files into *.fastq.gz files (compressed FASTQ files) in
CASAVA. Multiplexed samples are demultiplexed during this step.

4 **Sequence alignment**—Aligns samples to a reference sequence using the compressed FASTQ files.

5 **Variant analysis and counting**—Calls Single Nucleotide Polymorphisms (SNPs) and indels, and performs read counting (for RNA sequencing).

After variant analysis and counting are finished, the results can be viewed and analyzed further in the GenomeStudio® software, or the result files can be analyzed using third-party software.

Figure 4: Sequencing Data Analysis Workflow