Poster
CPHM-289
6/18/2023

Evaluation of Early Access NextSeq 2000 2x300 Cycle Sequencing Chemistry Utilizing Datasets of Importance to Food Safety

Christopher Grim[1], Phillip Curry[1], Ai Kataoka[1], Jose Roberto Guzman[2], Padmini Ramachandran[1], Tina Y. Huang[3], Samuel S. Hunter[3], Adrienne Hall[3], Kelly Hoon[3], Sandra M. Tallent[1]

[1]Center for Food Safety and Applied Nutrition, Food and Drug Administration, USA; [2]Goldbelt C6, LLC., Chesapeake, VA, USA, [3]Illumina, Inc., San Diego, CA

*christopher.grim@fda.hhs.gov, Tel +240-402-3582

Abstract

Background: Illumina-based short-read chemistry has become the community standard for whole genome sequencing (WGS) and the MiSeq™ (MS) benchtop sequencer has become a "workhorse" in this sector. The Illumina NextSeq™ (NS) 1000/2000 suits a number of mid-throughput applications; however, until recently only up to 2x150 bp paired-end read sequencing was available.

Methods: To assess the application of 600 cycle sequencing on the NextSeq 2000 platform, both bacterial isolate WGS and shotgun metagenomics libraries were used.

The WGS sample set (strains of Shiga toxin-producing *E. coli*, *Salmonella*, *Vibrio parahaemolyticus* and a diverse validation foodborne strain set) were run on NS 2000 P1 & P2 flow cells (FCs) with 600 cycles (NS600), compared to NS 2000 P1 300 cycles (NS300) and MiSeq 600 cycles (MiSeq) and analyzed for read1 and read2 quality scores, assembly number of contigs and N50.

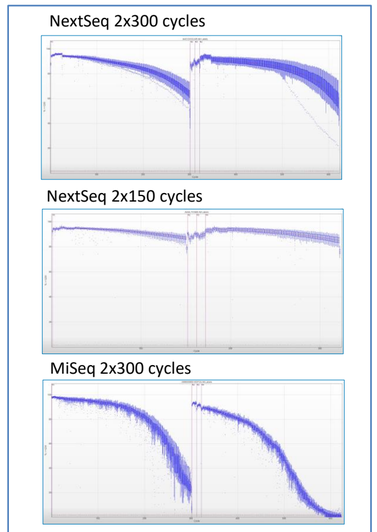
The metagenomics samples (scat and wastewater samples) were run on NS 2000 P2 with 600 cycles and compared with 300 Cycles (NS300). Metagenomic datasets were compared according to % classified reads and % reads identified as a target foodborne pathogen serovar.

Results: The mode of the average read quality scores for read 1 and read 2 were Q32 and Q31 for NS600 cycle, Q35 and Q31 for MiSeq, and Q33 and Q32 for NS300 cycle, respectively; however, MiSeq runs were restricted to 2 x 250 cycles to achieve this read quality. Regardless of the organism, the N50 value was higher and the contig numbers were lower for the draft assemblies from the NS600 flow cell runs, compared to NS300 and MiSeq, irrespective to coverage. For the metagenomic sequencing trials, we observed improved resolution at the read level for the NS600 flow cells. For a scat metagenome dataset (n=72), we were able to determine the same *Salmonella* serovar for all 26 samples identified with NS300 cycle flow cell, with only 25% of read sequencing depth. Further, an additional 5 samples were correctly identified, as determined by culture results, that were not with NS300 cycle sequencing. As expected, the longer sequencing cycling condition also resulted in a higher percentage of sequencing reads being identified using Kraken2 and an in-house kmer taxonomic classification tool, Bactikmer.

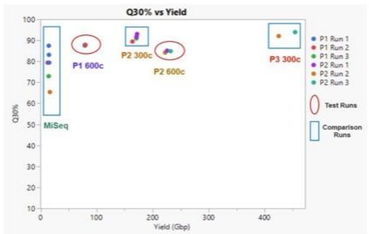
Conclusions: The NextSeq 2000 P1 and P2 600 cycle flow cells produced 2x longer reads with no reduction is sequencing quality, which allowed better isolate WGS assemblies and better discrimination with lower read depth for metagenomic datasets. The improved resolution enables higher throughput sequencing applications for food safety.

Sequencing Read Quality

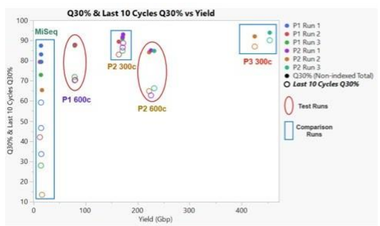
Representative distribution plots of %Q30 base-call quality score by cycle number for NextSeq and MiSeq platforms. Plots enclosed in boxes represent runs in which the same DNA Prep libraries were run on each instrument and flow cell.



Base-call quality score, %Q30, plotted versus sequencing yield for all trials. For the same libraries sequenced on P1 and P2 600c kits, total %Q30 has a much tighter distribution compared to MiSeq runs.

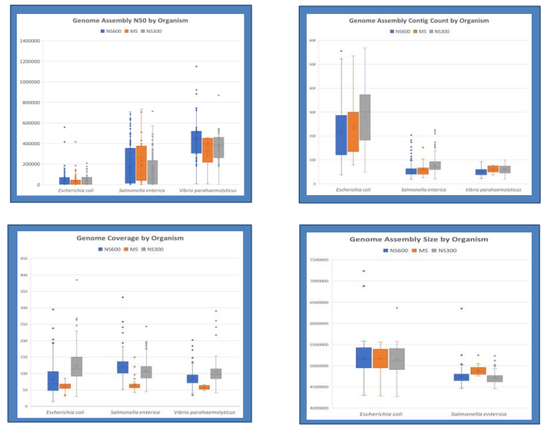


Last 10 cycles %Q30 on P1 and P2 600c kits are consistently higher with less spread vs. previous MiSeq runs.

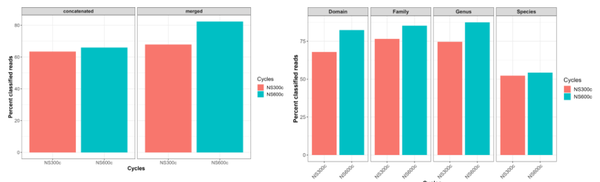


WGS Assembly Data Quality

Isolate whole genome sequencing *de novo* assembly metrics by organism for each platform and cycle combination



Percent of sequencing reads classified by Kraken2 from metagenomic trials by NextSeq flow cell kit. Analysis was performed on concatenated and merged read datasets, and then further divided by taxonomic level.



Shannon diversity box plots for metagenomic trials (5 and 6) by NextSeq flow cell kit. Analysis was performed on the merged read datasets.

Salmonella target (serotype(s)) resolution – NS 600c achieve equivalent or better resolution compared to NS 300c with, on average, 28% sequencing read depth. NS600c also detected 5 additional serovar strains: Dublin (x2), Infantis, Newport, and Poona (second serovar present in sample). Four samples, with multiple enrichment timepoints, shown.

Salmonella Newport

Flowcell	Sample	1° enrich	RV	TT
P2-600	F762	ND	ND	23582 (0.63%)
P3-300		ND	ND	110620 (0.8%)
P2-600	F798	49444 (1.34%)	1686617 (30.39%)	167536 (1.96%)
P3-300		249754 (1.77%)	7240153 (33.32%)	629114 (2.0%)

Salmonella Braenderup

Flowcell	Sample	1° enrich	RV	TT
P2-600	F1000	8556 (0.26%)	778856 (14.52%)	818856 (20.55%)
P3-300		-	2980947 (14.78%)	3251692 (20.84%)

Salmonella Infantis

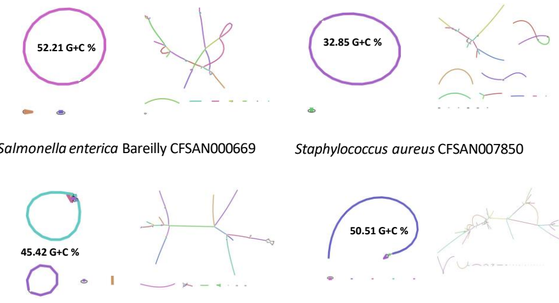
Flowcell	Sample	RV	TT
P2-600	F1006	60661 (1.32%)	7105 (0.17%)
P3-300		228711 (1.34%)	

Coming Soon - ICLR

https://www.illumina.com/science/technology/next-generation-sequencing/long-read-sequencing.html

FastQC plots of base-call quality scores by cycle number for MiSeq™, NextSeq™, and Illumina Complete Long Read (ICLR) for the CVSS (CFSAN Verification Strain Set). For NextSeq™ and MiSeq™, the same DNA prep libraries were run on each instrument and flow cell. For ICLR, initial libraries were run on NovaSeq ™ 6000.

Bandage visualizations of ICLR and short-read primary sequencing *de novo* assemblies of CVSS examples.



Salmonella enterica Bareilly CFSAN000669

Staphylococcus aureus CFSAN007850

Vibrio parahaemolyticus CFSAN123154

Escherichia coli O157:H7 CFSAN076620

Experimental Design

- Illumina provided 3 - P1 600 cycle kits and 3 - P2 600 cycle kits for Early Access testing on NextSeq 2000 (NS2K)
- Objective(s) – To evaluate NS2K 600 cycle kit performance compared to NS2K 300 cycle kit and MiSeq (2x250 or 2x 300) performance
- 6 trials (and evaluation parameters):
  - ➢Isolate WGS - assembly N50 and # contigs
    1. P1 run1 – 81 STEC, 15 *Salmonella*; NS600 and MS500
    2. P1 run2 – 114 *Salmonella*, 75 *V. para*; NS600, NS300, MS500
    3. P1 run3 – 192 STEC; NS600, NS300, MS500
    4. P2 run1 – 196 Salmonella, CVSS; NS600, NS300, MS600
  - ➢ Shotgun Metagenomics – taxonomic classification, select target resolution
    5. P2 run2 – 72 scat and soil samples; unenriched, and primary and selective (RV or TT) enrichments for *Salmonella*
    6. P2 run3 – 92 wastewater samples

CFSAN Verification Strain Set (CVSS)

CFSAN #	Organism	Purpose
CFSAN000189	*Salmonella enterica* Bareilly	SNP recovery
CFSAN000318	*Salmonella enterica* Heidelberg	plasmid detection
CFSAN000661	*Salmonella enterica* Bareilly	SNP recovery
CFSAN000669	*Salmonella enterica* Bareilly	SNP recovery
CFSAN000752	*Salmonella enterica* Bareilly	SNP recovery
CFSAN002349	*Listeria monocytogenes*	SNP recovery
CFSAN007850	*Staphylococcus aureus*	AMR, toxin detection, low G+C%
CFSAN007894	*Staphylococcus aureus*	toxin detection, low G+C%
CFSAN008100	*Listeria monocytogenes*	well-characterized strain
CFSAN008585	*Salmonella enterica* Derby	organism diversity
CFSAN023464	*Listeria monocytogenes*	SNP recovery
CFSAN023465	*Listeria monocytogenes*	SNP recovery
CFSAN023468	*Listeria monocytogenes*	SNP recovery
CFSAN023469	*Listeria monocytogenes*	SNP recovery
CFSAN029786	*Shigella dysenteriae* serotype 3	AMR
CFSAN030807	*Shigella sonnei*	plasmid detection
CFSAN032805	*Campylobacter coli*	organism diversity, low G+C%
CFSAN032806	*Campylobacter jejuni*	organism diversity, low G+C%
CFSAN044836	*Listeria innocua*	organism diversity
CFSAN051458	*Escherichia coli* O121	toxin detection
CFSAN068773	*Cronobacter sakazakii*	organism diversity
CFSAN068816	*Bacillus cereus*	organism diversity
CFSAN076620	*Escherichia coli* O157:H7	toxin detection
CFSAN084950	*Pseudomonas aeruginosa*	common background, High G+C%
CFSAN084952	*Pseudomonas fluorescens*	common background, High G+C%
CFSAN086180	*Klebsiella variicola*	common background
CFSAN086181	*Klebsiella pneumoniae*	common background
CFSAN086182	*Citrobacter braakii*	common background
CFSAN086183	*Enterobacter cancerogenus*	common background
CFSAN122995	*Salmonella enterica diarizonae*	organism diversity
CFSAN123154	*Vibrio parahaemolyticus*	organism diversity
CFSAN123155	*Vibrio parahaemolyticus*	organism diversity

Conclusions

- NextSeq 600 cycle kits yielded sequencing accuracy quality scores equivalent or better than MiSeq 600 cycle V3 chemistry, especially in late R1 and R2 cycles, with less variability.
- NextSeq 600 cycle kits yielded genome assemblies that were equivalent or better than MiSeq 600 cycle V3 chemistry and were a considerable improvement over NS 300 cycle kits.
- For metagenomic samples, longer reads improve taxonomic classification - fewer unclassified reads, and improve diagnostic power – more specific with lower read depth.
- ICLR generates long reads for complete or near-complete de novo genome assembly.

 SP-00369