TECHNICAL NOTE

Impact of NovaSeq™ flow cell lane edge integrity on sequencing performance

Spatial variability in primary metrics does not affect WGS variant calling

illumina®

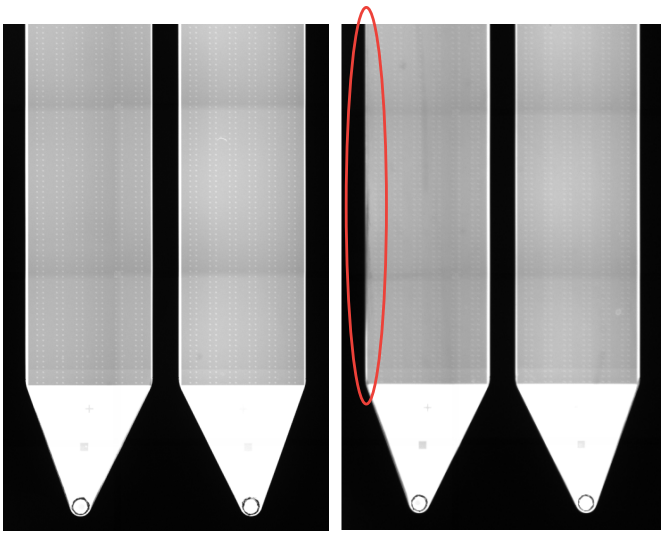
Introduction

Whole-genome sequencing (WGS) is a high-intensity application that provides a comprehensive view of the genome to enable analysis of a full range of variants, including single nucleotide variants (SNVs), insertions and deletions (indels), structural variants (SVs), and more. Critical metrics for variant calling include recall (number of correctly called variants relative to the expected number) and precision (accuracy of variant calling). Adequate sequencing coverage (number of reads covering any given locus in the genome) and high data quality scores (Q-scores)* are required to enable accurate variant calling.

Performing WGS on the NovaSeq 6000 System using the NovaSeq S4 flow cell delivers an output of 2400-3000 Gb or 8-10B read pairs with > 85% of bases higher than Q30,† averaged over the run. It is a known phenomenon that quality scores decrease over the course of a run as noise increases relative to signal; however, this variability is overwhelmed by the overall volume and quality of data to sequence 48 genomes in dual flow cell mode, or 24 genomes per flow cell, with primary metric specifications that meet a minimum threshold for consistent performance in variant calling.

Lane edge width variability

In addition to changes in read quality over the course of a sequencing run, there is normal expected variation in different areas across a flow cell. The imaging area of the NovaSeq S4 flow cell has been maximized to generate the most data possible and is very close to the total area of the flow cell lane. Therefore, the integrity of the narrow lane-edge buffer zone can have an effect on data in flow cell tiles immediately adjacent to the edge. Although the edge of a lane should show a uniform white buffer area, some variability in the width of the lane edge is part of normal expected variation in the manufacturing process (Figure 1).



FCID HW7HNDSXY FCID HW7C7DSXY

Figure 1: Lane-edge width variability—The flow cell on the left (FCID HW7HNDSXY) shows uniform bright white lane edges, while the flow cell on the right (FCID HW7C7DSXY) shows an area where the lane width is narrower, seen by the absence of the bright white lane edge (area circled in red).

This technical note compares sequencing runs performed on NovaSeq S4 flow cells with different lane edge widths and demonstrates that quality score variability does not impact performance in WGS variant calling.

Methods

Flow cell production

All NovaSeq S4 flow cells evaluated in this study were produced in an Illumina facility and met quality standards. At a critical step in the manufacturing process, images were captured for in-process analysis using a specially configured imaging system. All flow cells in the study were analyzed manually to determine lane edge uniformity.

Library preparation

A 24-plex pool of libraries was prepared for WGS using DNA from sample NA12878 with Illumina DNA PCR-Free Prep.

* The standard for Illumina sequencing by synthesis (SBS) chemistry is for a majority of bases to exceed a quality score of 30 (> Q30) or an expected error probability of 1:1000

† Expected output when using paired-end 150 base pair reads

Sequencing

Prepared libraries were sequenced in parallel on a NovaSeq 6000 System using NovaSeq S4 flow cells with narrow and wide lane edges at a read length of 2 × 150 bp.

Primary data analysis

Output data were displayed using the Illumina Sequencing Analysis Viewer (SAV v2.4.7). Several metrics were analyzed, including summary metrics, tile-level quality scores per cycle, and the spatial heat map display of tile-level quality scores on the flow cell.

Secondary data analysis

Base call output for each sequencing run was converted to FASTQ using bcl2fastq v2.20.0.422 on a high-performance computing cluster. The paired-end reads from each FASTQ were then uploaded to BaseSpace™ Sequence Hub. No quality trimming was performed. Alignment and small variant calling were completed using the DRAGEN™ Germline Pipeline app Version 3.2.8 using the default settings. The resulting VCF files were evaluated using the Hap.py Benchmarking app Version 1.0.0 using vcfeval. Alignments were generated using the GRCH37 + Decoy reference genome and small variants were evaluated against NA12878 Platinum Genome V2017.1.

Results

Variability in quality scores near lane edges

A NovaSeq S4 flow cell with a narrow lane edge (FCID HW7C7DSXY) showed reduced Q-scores in tiles located near the lane edges in a cycle near the end of the run, whereas a NovaSeq S4 flow cell with a wide lane edge (FCID HW7HNDSXY) showed all tiles Q30 > 70% during the same cycle (Figure 2). Although the difference is small in terms of the aggregate overall yield for the flow cell, it is easily seen using SAV. The increased variability in Q-scores was seen in flow cell FCID HW7C7DSXY throughout the sequencing run, and increased as the run progressed (Figure 3). Although these results indicate increased variability in flow cells with narrow lane edges, only a few tiles on one surface are affected, such that flow cells FCID HW7HNDSXY (wide lane edge) and FCID HW7C7DSXY (narrow lane edge) have nearly identical summary metrics for quality and yield (Table 1).

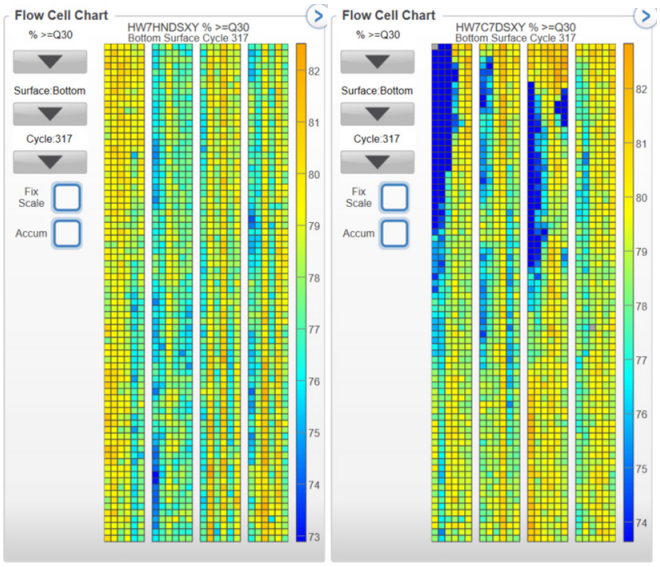


Figure 2: Variation in primary metrics adjacent to narrow lane edges—Flow cell FCID HW7HNDSXY (left) shows all tiles Q30 > 70 at cycle 317, while flow cell FCID HW7C7DSXY (right) shows an area where the tiles have a Q30 in the 60-70 range (blue-color tiles are below the autoscaling shown on the Y axis). This gradient is only visible on the bottom surface of the flow cell.

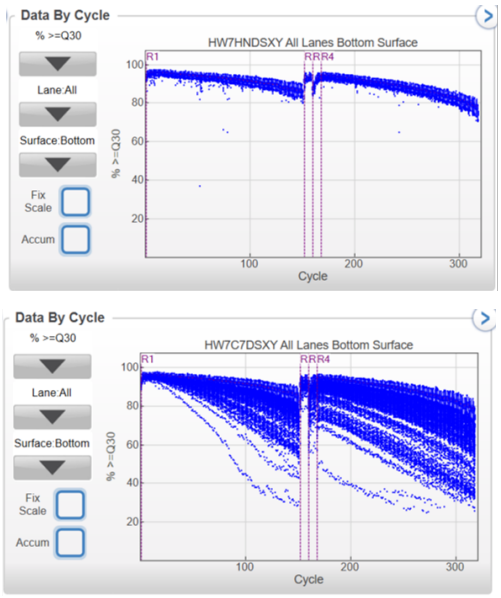


Figure 3: Spatial variation in primary metrics increase with cycles—Flow cell FCID HW7HNDSXY (top) shows a tight range of Q-scores with little decrease as the run progresses, while flow cell FCID HW7C7DSXY (bottom) shows more variability, which increases with additional cycles.

Table 1: Overall metrics for flow cells with different lane-edge widths

	Yield total (G)	Projected total yield (G)	% aligned	Error rate (%)	Intensity cycle 1	% ≥ Q30	% ≥ Q30 (last 10 cycles)
FCID HW7HNDSXY (wide lane edge)							
Read 1	1829.27	1829.27	4.81	0.17	1360	93.09	55.58
Read 2 (i)	85.36	85.36	0.00	NaN	1404	93.79	93.79
Read 3 (i)	85.35	85.35	0.00	NaN	1375	92.03	92.03
Read 4	1829.26	1829.26	4.79	0.19	967	89.89	82.70
Non-indexed total	3658.53	3658.53	4.80	0.18	1164	91.49	85.64
Total	3829.24	3829.24	4.80	0.18	1277	91.55	88.63
FCID HW7C7DSXY (narrow lane edge)							
Read 1	1795.16	1795.16	5.04	0.22	1376	93.31	88.25
Read 2 (i)	83.77	83.77	0.00	NaN	1445	94.00	94.00
Read 3 (i)	83.77	83.77	0.00	NaN	1406	92.37	92.37
Read 4	1795.05	1795.05	5.00	0.27	976	89.88	82.43
Non-indexed total	3590.20	3590.20	5.02	0.25	1176	91.60	85.34
Total	3757.74	3757.74	5.02	0.25	1301	91.67	88.57

Evaluating additional flow cells

Primary sequencing metrics

Prepared libraries were sequenced on a total of 10 flow cells, six with narrow lane edges and four with wide lane edges. Analysis of primary sequencing metrics showed that all narrow lane edge flow cells had a Q30 gradient in tiles on the affected lane edge (data not shown). However, this variability did not affect overall metrics, as all 10 flow cells exceeded the system specifications for overall quality (% > Q30 greater than 85%) and yield (8-10B read pairs passing filter(PF)) (Table 2).

Secondary sequencing metrics

Sequencing data from the prepared libraries run on the 10 flow cells were aligned to a reference human genome and assessed for secondary quality metrics. No significant differences were seen in coverage (Figure 4A), percent alignment (Figure 4B), or Q40 MAP score (Figure 4C).

Table 2: Summary of primary sequencing metrics

FCID	Lane edge	% > Q30	Last 10% > Q30	Read pairs PF (M)
HTMLVDSXY	Narrow	89.76	79.50	12145
HTMM5DSXY	Narrow	90.25	80.59	12091
HTMM7DSXY	Narrow	90.99	81.02	12240
HTMMCDSXY	Narrow	89.96	79.51	12243
HW7C5DSXY	Narrow	92.09	82.93	11467
HW7C7DSXY	Narrow	88.57	82.43	11967
HTMLNDSXY	Wide	90.63	81.24	12313
HTMLTDSXY	Wide	90.2	81.22	12078
HW257DSXY	Wide	91.55	82.98	12203
HW7HNDSXY	Wide	91.55	82.70	12194

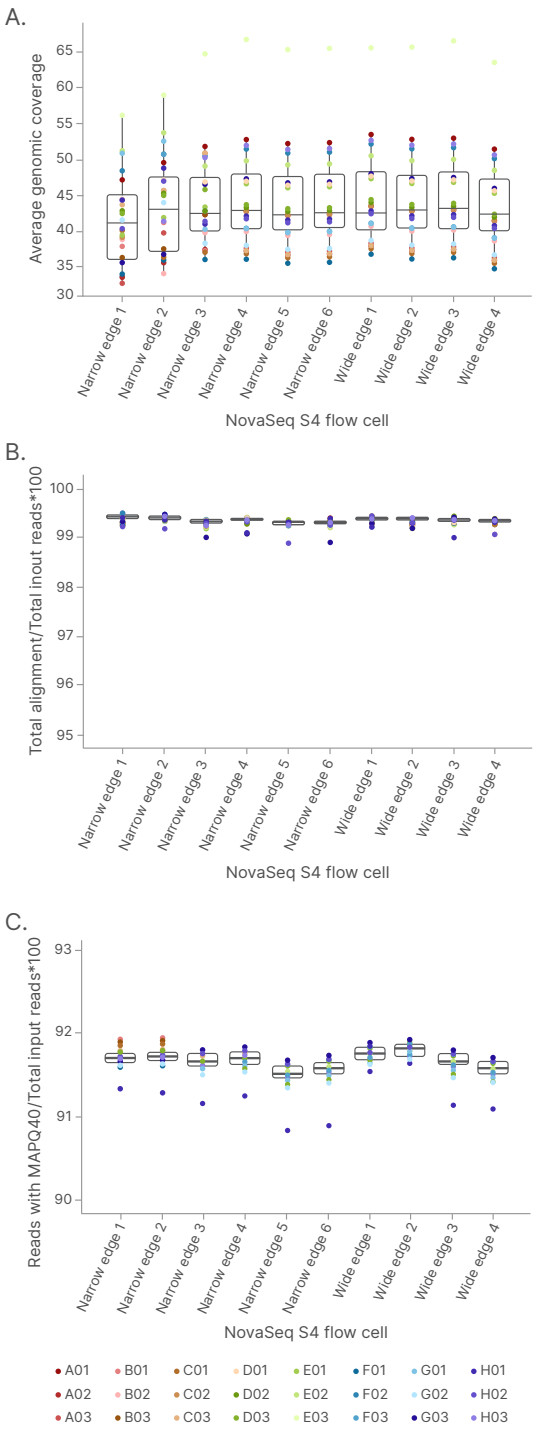


Figure 4: Secondary sequencing metrics—No significant differences were seen in secondary quality metrics, including (A) coverage, (B) percent alignment, or (C) Q40 MAP scores using flow cells with narrow or wide lane edges.

Variant calling performance

Flow cells with narrow and wide lane edges were evaluated for precision and recall for SNPs and indels present. NA12878 is fully sequenced with all known variants cataloged. Variant calling for SNPs was consistent across all flow cells, regardless of lane edge width, as measured by precision (Figure 5A) and recall (Figure 5B). Similarly, indel variant calling showed consistent precision (Figure 6A) and recall (Figure 6B) for flow cells with either narrow or wide lane edges. All metrics were equivalent to previously published results for germline variant calling using the DRAGEN Pipeline.[1]

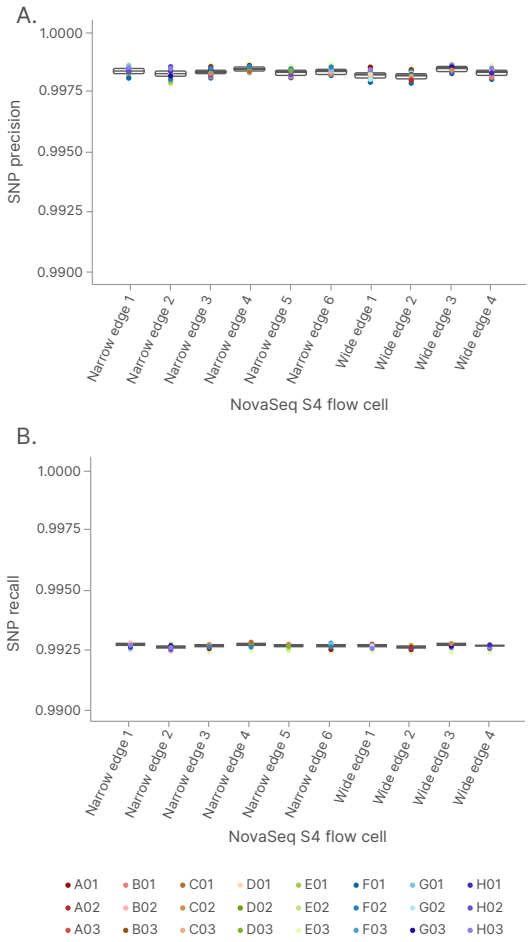


Figure 5: SNP variant calling precision and recall—SNP variant calling was consistent across 24 libraries sequenced on flow cells with narrow and wide lane edges, as measured by (A) precision and (B) recall.

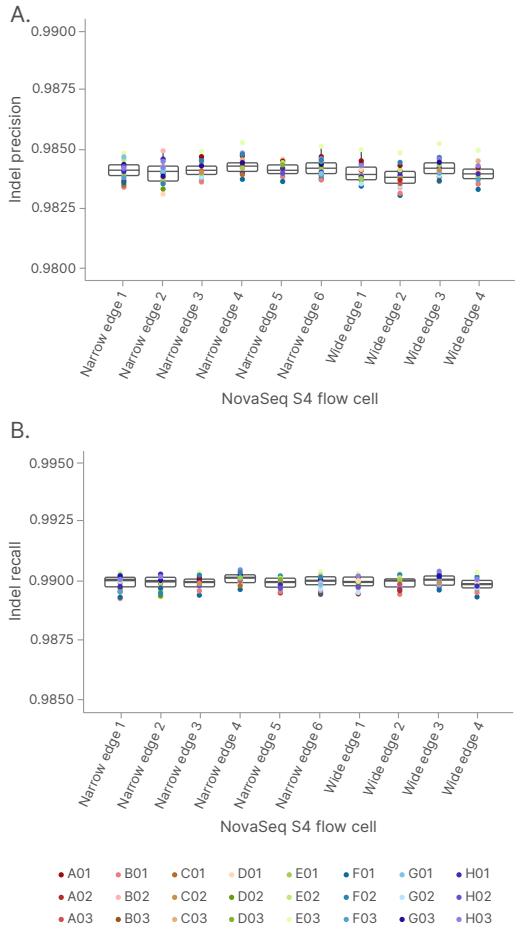


Figure 6: Indel variant calling precision and recall—Indel variant calling was consistent across 24 libraries sequenced on flow cells with narrow and wide lane edges, as measured by (A) precision and (B) recall.

Summary

As a result of the manufacturing process, NovaSeq S4 flow cells can have a known source of variability in the width of flow cell lane edges. This technical note evaluates flow cells with differing lane edge widths to determine their suitability for WGS variant calling. While minor differences are seen in primary sequencing and alignment metrics, all flow cells exceed all performance requirements, and the differences do not impact variant calling.

Learn more

NovaSeq 6000 System, www.illumina.com/systems/sequencing-platforms/novaseq.html

References

1. Zhao S, Agafonov O, Azab A, Stokowy T, Hovig E. Accuracy and efficiency of germline variant calling pipelines for human genome data. *Sci Rep*. 2020;10(1):20222. doi: 10.1038/s41598-020-77218-4.

illumina®

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-AMR-00104 v1.0