New Whole-Genome Sequencing Workflows from Illumina FastTrack® Services

New whole-genome sequencing workflows offer low-input DNA options and improved variant calling.

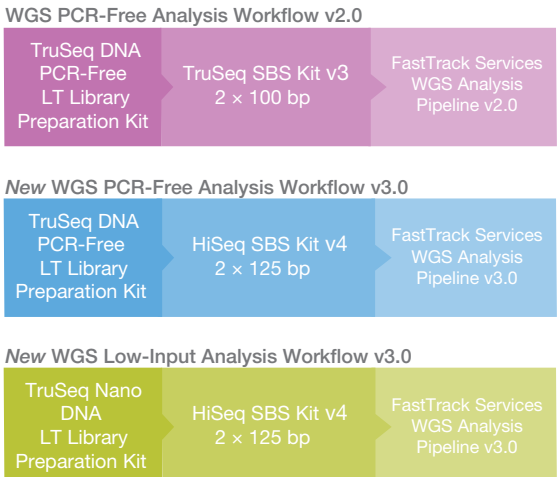
Introduction

The Illumina Genome Network™ (IGN)[1] is a group of leading institutions worldwide, with experience and expertise in Illumina technology, that offers sequencing services to individual researchers. As part of the IGN, Illumina FastTrack Sequencing Services has offered whole-genome sequencing services to customers worldwide since 2010. With proven Illumina sequencing technology and years of experience, FastTrack Services is committed to providing the highest data quality to its customers. To continue offering best-in-class sequencing services, the IGN and FastTrack Services have incorporated the latest advances in library preparation, instrument chemistry, and data analysis into their standard whole-genome sequencing (WGS) workflows (Figure 1).

The new FastTrack WGS workflows provide a number of significant improvements and benefits. With TruSeq® DNA PCR-Free and TruSeq Nano DNA Library Preparation Kits, FastTrack Services now delivers high-quality data using lower DNA input requirements. Lower DNA input requirements make whole-genome sequencing more accessible to difficult sample types or limited quantity DNA samples, such as tumor DNA. Furthermore, TruSeq Nano provides good coverage of traditionally challenging genomic regions, such as AT-rich and GC-rich areas, promotors, and repetitive regions. For data analysis, the updated WGS Analysis Pipeline v3.0 has an enhanced structural variant (SV) calling algorithm that increases SV detection and accuracy. Combined with extended read lengths and HiSeq® SBS v4 chemistry, these updates provide customers with more powerful tools to investigate novel links between genetic variation and human disease.

This technical note describes a number of validation experiments designed to evaluate the performance of the new FastTrack Services WGS Analysis Workflows in comparison to the WGS PCR-Free Analysis Workflow v2.0. To assess the differences and similarities between the WGS workflow data sets, coverage uniformity, single nucleotide variant (SNV) and insertion-deletion (indel) variant calling statistics, sensitivity, specificity, and array concordance were compared.

Figure 1: Updated FastTrack Services WGS Workflows



Improvements to FastTrack Services WGS workflows include the introduction of low-input library preparation, HiSeq SBS v4 chemistry, longer read lengths, and a new version of the FastTrack Services Analysis Pipeline.

Materials and Methods

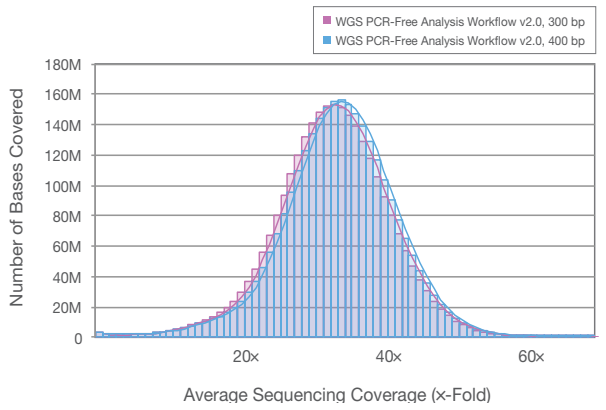
Insert Size Comparison

DNA samples from a familial CEPH trio were obtained from the Coriell Institute for Medical Research[2]. The father (NA12891), mother (NA12892), and daughter (NA12878) DNA samples were processed using the TruSeq DNA PCR-Free LT Library Preparation Kit (Illumina, catalog # FC-121-3001). For each sample, two libraries with two different insert sizes were prepared (300 bp insert size from the previous workflow and 400 bp from the new PCR-free workflow). Three lanes of HiSeq sequencing data from each sample were analyzed using the FastTrack Services WGS Analysis Pipeline v2.0[3]. Within this software package, alignment and variant calling were performed with the Isaac™ Aligner and Variant Caller[4]. Illumina FastTrack Services currently uses hg19 from the UCSC Genome Bioinformatics Group[5] as a reference genome.

Workflow Comparison

To assess the genome coverage quality and variant calling performance between the previous WGS workflow and the two new workflows, a three-way comparison was performed. All three CEPH samples were processed using the TruSeq DNA PCR-Free LT Library Preparation Kit or the TruSeq Nano DNA LT Library Preparation Kit (Illumina, catalog # FC-121-4001). Sequencing was performed with the TruSeq SBS Kit v3 Kit (Illumina, catalog # FC-401-3001) or the HiSeq SBS Reagent Kit v4 (Illumina, catalog # FC-401-4003). Libraries

Figure 2: Coverage Uniformity vs. Insert Size



Coverage uniformity was compared between the small insert size (300 bp) and the large insert size (400 bp). Insert sizes are specific to the experimental conditions under which studies are conducted and insert sizes may vary. Sequencing results depend on the quality of the DNA sample provided.

sequenced with TruSeq SBS v3 kits were analyzed with FastTrack Services WGS Analysis Pipeline v2.0. Samples sequenced with TruSeq SBS v4 kits were analized with the FastTrack Services WGS Analysis Pipeline v3.0.

Results

Insert Size and Coverage Uniformity

As part of the new standard workflows, FastTrack Services now constructs libraries with larger insert sizes. HiSeq v4 sequencing reagents and larger insert sizes enable longer read lengths and provide access to the latest improvements in sequencing chemistry. To assess whether this increased insert size affects coverage uniformity, the coverage uniformity plots of sample libraries with the smaller and larger insert sizes were compared. A sequencing run will have high coverage uniformity if the sequence reads are distributed randomly across the genome and if the ability to detect overlapping regions between reads is consistent within a sequencing run. High coverage uniformity will result in a plot with a Poisson-like distribution and a small standard deviation. Here, mapped depth refers to the total number of bases sequenced and aligned at a given reference base position. As shown in Figure 2, the genome coverage for the large-insert library showed high coverage uniformity and had a comparable coverage distribution plot compared to the small-insert library sample.

Insert Size and Variant Calling

To determine if larger insert size affects variant calling, various metrics were computed for SNVs and indels. These included call rates across all reference positions, the total number of variant calls, the ratio of heterozygous to homozygous variants, the fraction of variants not found in dbSNP, and the transition to transversion ratio (Table 1). A side-by-side review of the data from libraries with small and large insert sizes reveals comparable variant calling metrics.

Table 1: SNV and Indel Call Quality Statistics by Insert Size

Variant Class	Quality Metrics	WGS PCR-Free Analysis Workflow v2.0, 300 bp	WGS PCR-Free Analysis Workflow v2.0, 400 bp
	Average Coverage Depth	34.5×	35.3×
SNV	Call Rate (%)	95.08	95.47
	Total SNVs	3,465,345	3,510,092
	Ti/Tv	2.08	2.08
	Novelty Rate	3.21	3.28
	Het/Hom	1.61	1.60
Indel	Total Indels	365,616	368,466
	Novelty Rate	6.28	6.24
	Het/Hom	1.94	1.95

Metrics calculated as an average across the CEPH trio NA12891, NA12892, and NA12878. Total quality results depend on the average coverage depth.

Call Rate: Percent of non-N reference genome in which a reference or non-reference call was made for both alleles

Total SNVs: Total number of SNVs that have 'Pass' value in the FILTER key of the VCF file

Ti/Tv: Transition to Transversion ratio of SNV calls

Novelty Rate: Percent of SNVs or indels not found in dbSNP132

Het/Hom Rate: Heterozygous to Homozygous ratio of SNV or indel calls

Total Indels: Total number of indels that have 'Pass' value in the FILTER key of the VCF file

Table 2: Specificity, Sensitivity, and Array Concordance by Insert Size

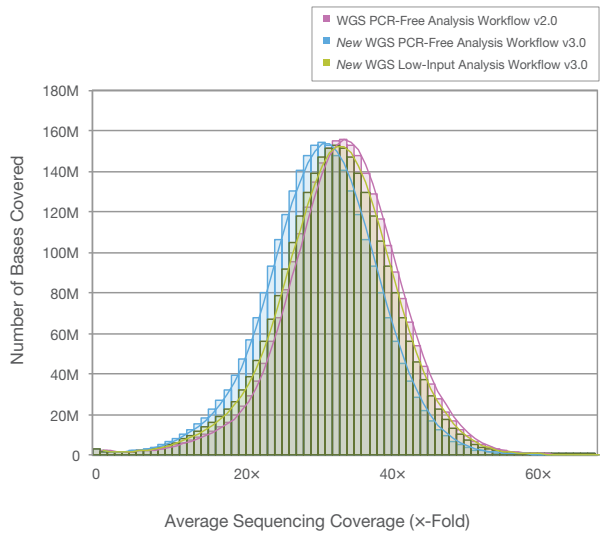
Detection Metric	Variant Class	WGS PCR-Free Analysis Workflow v2.0, 300 bp	WGS PCR-Free Analysis Workflow v2.0, 400 bp
Array Concordance	SNV	99.43	99.42
Specificity	SNV	99.87	99.87
	Indel	98.6	98.55
Sensitivity	SNV	92.38	92.87
	Indel	41.65	41.72

Array Concordance: Genome concordance with calls from HumanOmni2.5M array calculated as an average across the CEPH trio.

Specificity: Mendelian non-conflict rate for the variants called in the CEPH trio.

Sensitivity: Recovery rate of NA12878 variants previously reported[7]. (95,005 SNVs and 11,403 indels)

Figure 3: Coverage Uniformity by Workflow



Comparison of the sequencing depth versus total number of bases covered demonstrates that coverage uniformity is consistent across all three workflows.

To examine variant calling performance more closely, specificity, sensitivity, and array concordance to the Infinium® HumanOmni2.5M BeadChip[6] were measured for both insert size conditions (Table 2). Variant calling specificity measurements assume inheritance patterns based on the samples forming a familial trio (two parents and a child). Except for *de novo* mutations in the child, any variant identified in the child should be also called in a minimum of one parent. Conflict rate is based on comparison to expected Mendelian inheritance patterns and is an established proxy for the false positive rate. Sensitivity was measured as the ability to detect a set of well-characterized variants for NA12878, as reported previously[7]. Array concordance was measured as the agreement between SNV calls in the sequencing data versus a curated set of high-confidence calls made using a high-density microarray. These metrics indicate that variant calling is comparable for both insert size libraries.

Workflow Effects on Coverage Uniformity and Variant Calling

To examine the impact of the new WGS workflows on coverage uniformity, sequencing data from all three workflows (WGS PCR-Free Analysis Workflow 2.0, the new WGS PCR-Free Analysis Workflow 3.0, and the new WGS Low-Input Analysis Workflow v3.0) were compared. By plotting the sequencing depth versus the total number of bases covered for each data set, the data show that uniformity is comparable across all workflows (Figure 3).

Workflow-related differences in variant calling were determined by comparing various metrics for SNVs and indels, including call rate across all reference positions, the total number of variant calls, the ratio of heterozygous to homozygous variants, the fraction of variants not found in dbSNP, and the transition to transversion ratio (Table 3).

Table 3: SNV and Indel Call Quality Statistics by Workflow

Variant Class	Quality Metric	WGS PCR-Free Analysis Workflow v2.0	*New* WGS PCR-Free Analysis Workflow v3.0	*New* WGS Low-Input Analysis Workflow v3.0
	Average Coverage Depth	35.3×	32.5×	30.9×
SNV	Call Rate (%)	95.47	94.78	94.24
	Total SNVs	3,510,092	3,496,305	3,459,373
	Ti/Tv	2.08	2.08	2.09
	Novelty Rate	3.28	3.29	3.19
	Het/Hom	1.6	1.62	1.63
Indel	Total Indels	368,466	523,681	487,965
	Novelty Rate	6.24	6.72	6.32
	Het/Hom	1.95	2.03	2.45

Metrics calculated as an average across the CEPH trio NA12891, NA12892, and NA12878. Total quality results depend on the average coverage depth.

Call Rate: Percent of non-N reference genome in which a reference or non-reference call was made for both alleles

Total SNVs: Total number of SNVs that have 'Pass' value in the FILTER key of the VCF file

Ti/Tv: Transition to Transversion ratio of SNV calls

Novelty Rate: Percent of SNVs or indels not found in dbSNP132

Het/Hom Rate: Heterozygous to Homozygous ratio of SNV or indel calls

Total Indels: Total number of indels that have 'Pass' value in the FILTER key of the VCF file

Table 4: Specificity, Sensitivity, and Array Concordance by Workflow

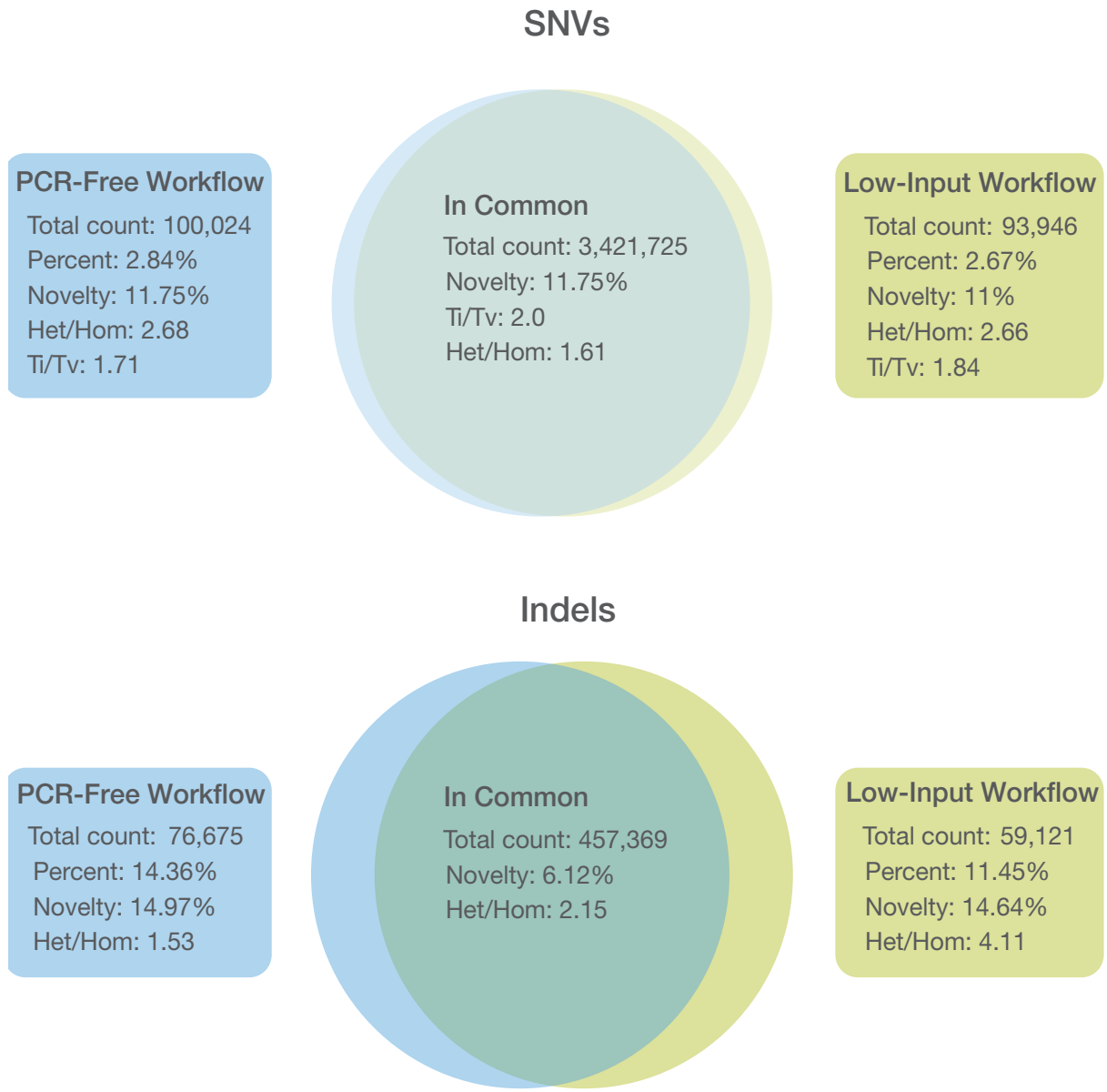
Detection Metric	Variant Class	WGS PCR-Free Analysis- Workflow v2.0	*New* WGS PCR-Free Analysis Workflow v3.0	*New* WGS Low-Input Analysis Workflow v3.0
Array Concordance	SNV	99.42%	99.43%	99.43%
Sensitivity	SNV	92.87%	93.55%	93.58%
	Indel	41.72%	52.92%	50.58%
Specificity	SNV	99.87%	99.88%	99.88%
	Indel	98.55%	98.19%	97.62%

Array Concordance: Genome concordance with calls from HumanOmni2.5M array calculated as an average across the CEPH trio

Sensitivity: Recovery rate of NA12878 variants previously reported[7]. (95,005 SNVs and 11,403 indels)

Specificity: Mendelian non-conflict rate for the variants called in the CEPH trio.

Figure 4: Overlap Between PCR-Free and Low-Input Data Sets



Comparison of results from the PCR-Free v3.0 and Low-Input v3.0 workflows using HiSeq v4 chemistry with 2 × 125 bp read lengths and WGS Analysis Pipeline v3.0 for the CEPH sample, NA12878.

Sensitivity, specificity, and array concordance were also compared between the three workflow data sets (Table 4). These data show no significant differences in variant calling metrics, specificity, or array concordance. However, with the new workflows, sensitivity significantly increased for variant calling, particularly for indels (Table 4).

PCR-Free vs Low-Input Analysis Workflows

In addition to computing summary metrics for variant calls, the overlap of variant calls between the new PCR-Free and Low-Input library preparation methods was measured (Figure 4). For SNVs, a call is considered overlapping if both workflows make a non-reference call at a genomic position. For indels, a call is considered overlapping if the genomic interval of the indel identified by both workflows overlaps. In addition to measuring the extent of the overlap, summary statistics are reported for the unique calls made by each workflow. The data show a high level of agreement between both workflows (99.8% agreement; total SNVs 3,421,725 and 96.7% agreement; total indels 457,369).

WGS Analysis Pipeline v3.0

To improve sequencing data quality, FastTrack Services offers the latest technologies for library preparation and sequencing chemistry. Significant upgrades to the WGS Analysis Pipeline include the incorporation of two new algorithms: Manta and Canvas. These new algorithms provide increased variant detection and measurable quality metrics, giving customers greater confidence in historically difficult-to-call variant types, such as SVs and CNVs.

The Manta algorithm calls all classes and sizes of structural variant types, including translocations, inversions, deletions/insertions (≥ 50 bp), and tandem duplications, without requiring copy number analysis or large-scale assembly. Manta combines paired-end and split read evidence during SV discovery and scoring to improve performance. It also provides genotype and quality scores for variants in single diploid samples.

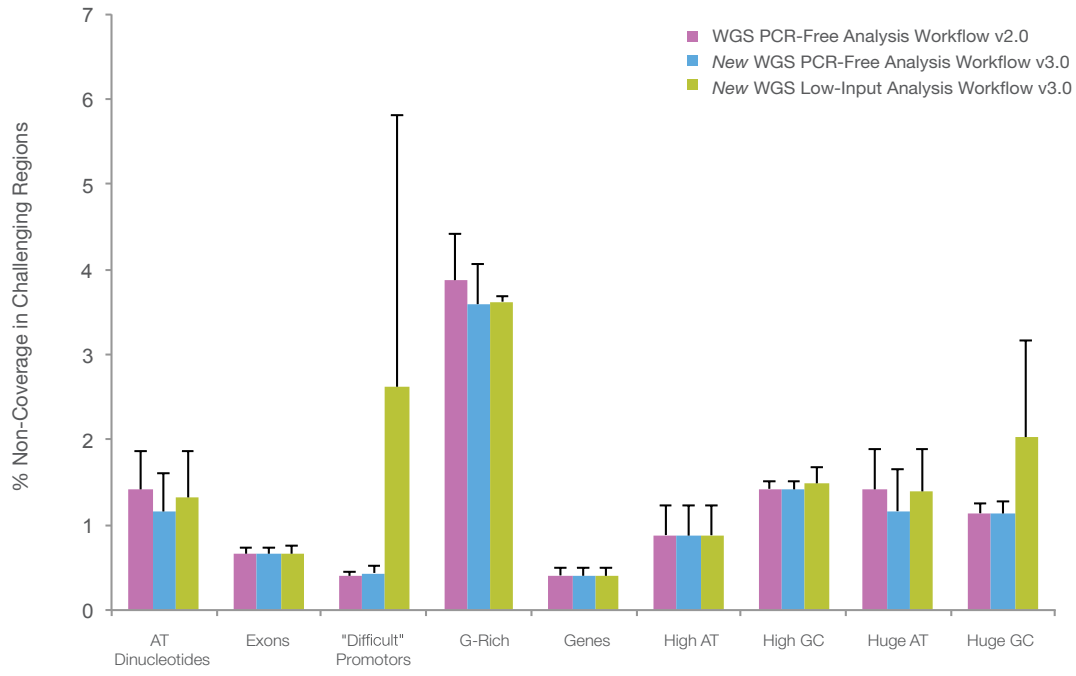
The Canvas script calls copy number variants from predominantly diploid samples. It identifies regions of the sample genome that are present zero, one, or more than two times. In brief, this is achieved by scanning the genome for regions having an unexpected number of short-read alignments. Regions with fewer than the expected number of alignments are classified as losses, while regions having more than the expected number are classified as gains. Canvas should be applied to low-depth cytogenetics experiments, low-depth single-cell experiments, or to whole-genome sequencing experiments. Canvas is not appropriate for whole-exome experiments or cancer studies.

Coverage of Challenging Regions

Sequence coverage, the average number of bases that align to known reference bases, is critical in determining confidence levels in variant discovery. To analyze performance of the new WGS workflows in traditionally difficult-to-sequence regions, the percent of non-coverage was plotted for various types of challenging regions (Figure 5). "Challenging regions" are defined as follows:

- **AT dinucleotide repeats** — 30 bases of repeated AT dinucleotides
- **Exons** — known human coding and non-coding exons defined in the RefSeq Genes track in UCSC Genome Browser[5]
- **"Difficult" promoters** — a set of 100 promoter regions, typically insufficiently covered, which are empirically defined by the Broad Institute of MIT and Harvard[8]
- **G-rich regions** — 30 bases with ≥ 80% G
- **Genes** — known human coding and non-coding genes defined in the RefSeq Genes track in UCSC Genome Browser[5]
- **High AT regions** — 100 bases with ≥ 75% AT content
- **High GC regions** — 100 bases with ≥ 75% GC content
- **Huge AT regions** — 100 bases with ≥ 85% AT content
- **Huge GC regions** — 100 bases with ≥ 85% GC content

Figure 5: Improved Coverage of Challenging Regions



Due to a PCR step during library preparation, the low-input workflow demonstrates increased variability in certain traditionally challenging regions when compared to other methods.

For well-characterized regions, such as exons and genes, the new workflows perform on par with the WGS PCR-Free Analysis Workflow v2.0. For some challenging regions, such as difficult promoters and GC-rich regions, the low-input workflow shows decreased coverage levels.

Discussion

The new FastTrack Services WGS workflows provide high-quality data and improved variant calling. In side-by-side comparisons of variant calling statistics, specificity, sensitivity, and array concordance, the new WGS workflows were comparable to WGS PCR-Free Analysis Workflow v2.0. Furthermore, several metrics show marked improvements with the new workflows, such as increased total indel detection (Table 3) and sensitivity for both SNVs and indels (Table 4).

Aside from the overall coverage uniformity and variant calling overlap between library preparation methods, there are trade-offs when working with the new low-input workflow. The low-input workflow shows increased variability between samples and sequencing coverage is reduced in several classes of challenging regions (Figure 5). The difference in coverage can be attributed to the inclusion of a PCR step in the low-input workflow, which leads to increased variance.

The new WGS workflows leverage many advantages from library preparation to data analysis. Reduced DNA requirements for TruSeq PCR-Free and TruSeq Nano Library Preparation Kits preserve limited or difficult sample types such as tumor samples. Furthermore, the TruSeq Nano Kit provides good coverage of traditionally challenging genomic sequences such as AT or GC-rich areas, promotors, and repetitive regions. The new workflows take advantage of the latest advances in sequencing chemistry, such as the HiSeq SBS v4 Kit, longer read lengths up to 2 × 125 bps, and data analysis improvements with the WGS Analysis Pipeline v3.0. The new data analysis pipeline includes the Manta and Canvas algorithms, which enhance copy number and structural variant calling for these difficult-to-call targets.

By providing the most up-to-date WGS workflows and data analysis methods currently available, FastTrack Services is empowering researchers with the tools to succeed in genomics research.

Learn More

To learn more about FastTrack Services visit www.illumina.com/services

To compare TruSeq Library Preparation Kits, see www.illumina.com/documents/products/technotes/technote_truseq_comparison.pdf

References

1. Illumina Genome Network (www.illumina.com/services/whole-genome-sequencing-services/sequencing-service-providers-ign.ilmn).
2. Coriell Institute for Medical Research (ccr.coriell.org).
3. Illumina Technical Support Note (2013) New algorithms increase computing efficiency for IGN whole-genome analysis. PN 770-2013-016.
4. Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, et al. (2013) Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics 29(16): 2041–2043.
5. UCSC Genome Browser (hgdownload.cse.ucsc.edu/downloads.html).
6. HumanOmni2.5-8 BeadChip Data Sheet (www.illumina.com/products/humanomni25-8_beadchip_kits.ilmn).
7. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 4453: 56-64.
8. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14: R51.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2014-016 Current as of 23 November 2014

illumina®