New Algorithms Increase Computing Efficiency for IGN Whole-Genome Analysis

Isaac software offers IGN customers an improved time-to-answer solution, while a reduction in Q-score resolution condenses the data footprint without compromising quality or accuracy.

Abstract

The Illumina Genome Network (IGN) Whole-Genome Analysis and Cancer Analysis Services now employ Isaac Genome Alignment and Isaac Variant Caller software to increase computing efficiency of whole-genome sequencing data[1]. A reduction in Q-score resolution and enriched annotation methods—including ENCODE, HGMD, 1000Genomes, and HapMap—enable IGN customers to better manage complex data sets while they identify biological significance. The rapid and accurate resequencing workflow generates data in BAM, VCF, and gVCF file formats in just over seven hours, improving time-to-answer solutions.

This technical note compares sequencing characteristics and small-variant calling results of Isaac Genome Alignment and Variant Caller workflows when deployed as part of the IGN Whole-Genome Sequencing Analysis Pipeline v2.0 and the Cancer Analysis Pipeline v2.0. Compared to previous versions employing the Consensus Assessment of Sequence and Variation (CASAVA) pipeline and Efficient Large-Scale Alignment of Nucleotide Databases (ELAND) aligner, the new informatics pipelines enable significantly increased alignment efficiencies and reduced data footprints, without compromising the quality of the data and variant calls.

Introduction

The volume of whole human genome sequencing has increased dramatically in recent years, leading to an unprecedented rate of variant discovery. Newly identified variants may have the potential to explain disease mechanisms. The growth in whole-genome sequencing is driven by several factors, including improved efficiencies of sample preparation and higher throughput capacity of sequencing instruments. In recent years, sequencer output has outpaced computing, storage, and software evolution. To address the informatics bottleneck created by the massive influx of data, IGN introduced new versions of the Whole-Genome Sequencing Analysis Pipeline v2.0 and the Cancer Analysis Pipeline v2.0. These pipelines leverage the Isaac Aligner, a fast and accurate alignment algorithm to improve time to answer, with reduced Q-score resolution to reduce genome build size and data storage costs for IGN customers. This technical note compares the sequencing characteristics and small-variant calling results from the new versions of the IGN Whole-Genome Sequencing Analysis Pipeline v2.0 and the Cancer Analysis Pipeline v2.0 to the previous CASAVA-based versions of the respective pipelines.

WGS Analysis Pipeline—Highly Accurate SNV and Indel Calling

The Whole-Genome Sequencing Analysis Pipeline uses the Isaac workflow to perform small-variant detection (SNVs and indels up to 50 bp). The Isaac workflow employs Isaac Aligner and Isaac Variant Caller to generate several outputs, including sequencing reads with reduced-resolution Q-scores in the BAM format, and variant data in both the VCF and genome VCF (gVCF) file format. For more information regarding the Isaac workflow and reduced Q-score resolution methods, refer to Illumina's Isaac Genome Alignment and Variant Caller and Reducing Whole-Genome Data Storage Footprint White Paper[2].

To assess variant-calling performance, the Centre d'Etude du Polymorphisme Humain (CEPH) trio of father (NA12891), mother (NA12892), and daughter (NA12878) was processed using both the new Isaac workflow and the previous CASAVA pipeline used in the IGN Whole-Genome Sequencing Service. Various metrics were computed for SNVs and indels. These included the 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 for SNVs (Table 1). Data processed with the Isaac workflow showed a higher average call rate than data processed with the CASAVA pipeline. However, the Isaac workflow calls fewer SNVs due to more stringent filtering criteria applied to the variant calls. For more details on filtering parameters, refer to Illumina's Whole-Genome Sequencing User Guide[3]. For indels, the Isaac workflow limits the indel range from 1 to 50 bp while larger indels (> 50 bp) are detected using the large variant caller components of Whole-Genome Sequencing Analysis Pipeline v2.0 and reported separately. In the CASAVA pipeline, the small indel detection range is 1–300 bp, while indels larger than 300 bp are reported separately.

To examine variant calling performance more closely, specificity, sensitivity, and concordance to the HumanOmni2.5M BeadChip Genotyping microarray were measured for both the Isaac workflow and CASAVA pipeline (Table 2). Measurement of the specificity of variant calling assumes inheritance based on the three analyzed samples forming a trio—two parents and a child. Except for *de novo* mutations in the child, any variant identified in the child should also be called in at least one of the parents. The number of Mendelian conflicts, therefore, was used as a proxy for false positives. This is considered a rough approximation, because failure to call variants in the parent can also create a conflict. Further, not all false positive

Table 1: SNV and Indel Call Quality and Statistics

Variant Class	Quality Metrics	CASAVA	Isaac
SNV	Call Rate	95.90%	96.83%
	Total SNVs	3,434,048	3,386,762
	Ti/Tv	2.03	2.17
	Het/Hom	1.57	1.61
	Novelty Rate	5.5%	4.9%
Indel	Total Indels	400,247	355,773
	Het/Hom	1.62	1.76
	Novelty Rate	19.9%	16.8%

Metrics calculated as an average across the CEPH trio NA12891, NA12892, and NA12878

Call rate: % 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

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

Ti/Tv: Transition to Transversion ratio of SNV calls

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

Novelty Rate: Percent of called SNVs or indels not found in dbSNP 132

Note: In the CASAVA pipeline, the small indel detection range is 1–300 bp, while in the Isaac workflow the range is 1–50 bp.

calls lead to a conflict (and *de novo* mutations in the child are also not accounted for). However, it is a reasonable metric when used to compare multiple workflows, since the flaws in the calculation(s) affect both workflows in the same manner. Sensitivity was measured by the ability to detect a set of well-characterized variants for NA12878, as reported in Kidd et al[4]. It is assumed that those variants were correct, and the ability to detect them within each of the analysis workflows was quantified. 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. The results indicate that variant calling is comparable for the two workflows (Table 2).

Table 2: Comparison of SNV Detection

Quality Metrics	Variant Class	CASAVA	Isaac
Sensitivity	SNV	90.5%	90.8%
	Indel	43.3%	41.5%
Specificity	SNV	99.84%	99.86%
	Indel	97.38%	97.52%
Concordance	Sites	99.45%	99.99%

Sensitivity: Recovery rate of NA12878 variants reported in Kidd et al. (95,005 SNVs and 11,403 indels)

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

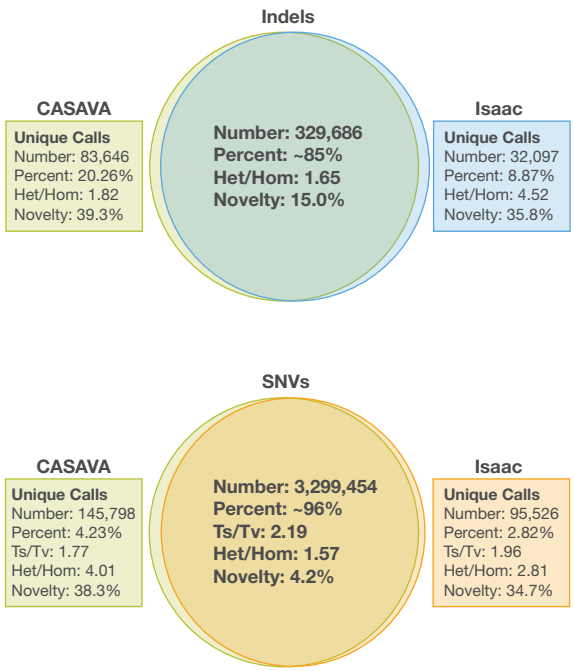
Concordance: Genome concordance with calls from OMNI2.5M array calculated as an average across the CEPH trio

Call Sets Show High Overlap with Other Workflows

In addition to computing summary metrics for variant calls, the overlap of variant calls was measured. 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.

Figure 1 compares the results from the Isaac workflow and the CASAVA pipeline. There is a high level of agreement (96%) for SNVs. The lower agreement (85%) in indel calling is reflective of the relative immaturity of indel calling methods compared to SNV calling methods. These results support an earlier assumption that the Isaac workflow and CASAVA pipeline have comparable small-variant calling accuracy.

Figure 1: Comparison between CASAVA and Isaac Workflows



Cancer Analysis Pipeline—Highly Accurate Somatic SNV and Indel Calling

IGN utilizes the Cancer Analysis Pipeline v2.0 to generate the data package that is delivered as part of the Cancer Analysis Service. The somatic small-variant calling component of the cancer analysis pipeline employs Isaac Aligner and Strelka[5] to generate BAMs with reduced resolution Q score and somatic small-variant data in the VCF format.

Table 3: Sequencing Summary and Statistics for Paired Tumor-Normal Analysis*

Quality Metrics	COLO 829				HCC 1187				HCC 2218				NA12878**			
	C+S		I+S		C+S		I+S		C+S		I+S		C+S		I+S	
	T	N	T	N	T	N	T	N	T	N	T	N	N	N	N	N
Mapped Sequence (%)	90.1	89.0	95.08	94.94	89.92	88.71	95.17	95.36	89.9	89.4	93.99	95.36	90.3	89.5	95.32	95.39
Total Somatic SNVs	45454		44890		15649		15437		29192		28500		1612		1169	
Specificity (%)	98.6		98.5		97.6		97.2		97.2		97.1		90.1		90.1	
Sensitivity	98.04		98.05		98.7		98.7		97.5		97.6		NA		NA	
Total Somatic Indels (Indel length < 50 bp)	874		724		1139		804		1231		647		87		9	
Specificity	82.1		83.6		85.2		86.4		76.9		78.8		93.9		88.9	
Sensitivity	80.1		81.54		87.5		100		89.2		100		NA		NA	

C=CASAVA; I=Isaac; S=Strelka; T=Tumor; N=Normal

Mapped Sequence: Percent of all passing filter reads which map to a unique position in the reference genome

Sensitivity:

HCC1187 and HCC2218: recovery rate of the confirmed somatic variants reported in COSMIC (2011)[6]

COLO-829: recovery rate of the confirmed somatic variants reported in Pleasance et al. (2010)[7]

Specificity:

Percent of called SNVs not found in dbSNP 132

Percent of indels (< 50 bp) not found in the 1000G dataset

* Tumor/Normal sequencing coverage > 40×/80×, respectively

** To establish a baseline false-positive rate, two NA12878 technical replicates were sequenced and analyzed through the Cancer Analysis Pipeline v2.0 and the previous CASAVA-based Cancer Analysis Pipeline v1.0. Sequencing coverage was similar to that of Tumor/Normal pairs, with 40×/80× coverage respectively.

To assess the performance of the somatic small-variant component of the Cancer Analysis Pipeline v2.0, three tumor-normal pairs (COLO 829, HCC*1187, and HCC 2218) were sequenced at a sequencing coverage > 40×/80×, respectively. Analysis results from this pipeline were compared to the previously used analysis pipeline comprised of CASAVA (with ELAND aligner) and Strelka. COLO 829 is a fibroblast cell line derived from a patient with metastatic melanoma. The epithelial cell lines HCC 1187 and 2218 are poorly differentiated cells derived from invasive ductal carcinoma. Normal samples for each of the tumor cell lines, generated from peripheral blood cells, were also analyzed. As a means of establishing a baseline false-positive rate, somatic variant-calling analysis was performed on replicates of NA12878. The summary alignment and variant-calling statistics are presented in Table 3. These results indicate that the accuracy of the somatic small-variant calling in the Cancer Analysis Pipeline v2.0 is comparable to the previous version used in the Cancer Analysis Service.

High Overlap in Somatic Variant Call Sets

To compare the two somatic small-variant calling components, the overlap of variant calls was measured. The same tumor/normal experimental conditions were used for consideration of overlap count. As shown in Figures 2 and 3, there is a high level of agreement (91%) for somatic SNVs and lower agreement (83%) for somatic indel calling.

Figure 2: Overlap Between Old Versus New Somatic SNV Call Sets

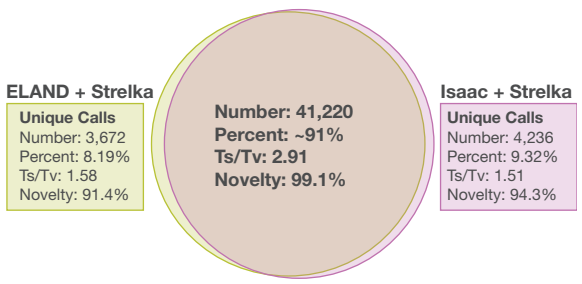
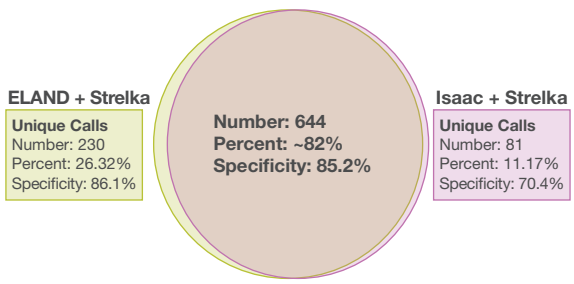


Figure 3: Overlap Between Old Versus New Somatic Indel Call Sets



Significantly Reduced Run-Time

To demostrate improvements in compute efficiencies provided by the new Whole-Genome Sequencing analysis pipeline, the CEPH trio was sequenced. Both alignment and variant calling speed and accuracy were tested by comparing analysis results from the new analysis pipeline (Isaac) versus the previous pipeline (CASAVA). Table 4 shows the end-to-end wall clock time for each of the two workflows tested. The Isaac workflow is more than twice as fast as CASAVA on a standard IlluminaCompute node. This gain in speed is achieved without compromising mapping and alignment accuracy (Table 5) or the average percent coverage of the genome at variable depths (Table 6). The Isaac aligner produces comparable values for various quality standards, such as percent of reads mapped, percent of mismatches to the reference, and average coverage by unique mapping reads.

Summary

The Isaac software-based Whole-Genome Sequencing Analysis Pipeline v2.0 and Cancer Analysis Pipeline v2.0 provide high-quality sequencing data and variant-calling accuracy. IGN's deployment of these enhanced analysis pipelines increases computing efficiencies through a reduction of Q-score resolution and improved compute timing, providing IGN customers with a reduced data footprint for lower data storage costs. In addition, enriched variant annotations and full genome summary files arm IGN customers with more analysis tools to identify biological context from their complex data sets, delivering a faster time to answer.

References

1. 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 10.1093/bioinformatics/btt314
2. http://res.illumina.com/documents/products/whitepapers/whitepaper_data-compression.pdf
3. http://support.illumina.com/downloads/whole_genome_sequencing_service_user_guide.ilmn
4. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.
5. Saunders CT, Wong WSW, Swami S, Becq J, Murray L, et al. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28: 1811–1817.
6. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 39: D945–950.
7. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196.

*HCC cell lines were invented by Drs. Adi F. Gazdar and John D. Minna at the University of Texas Southwestern Medical Center. Rights in and to the HCC cell lines, progeny, and unmodified derivates thereof belong to the Board of Regents of The University of Texas System. Illumina, Inc. has obtained permission from the Board of Regents of The University of Texas System through the University of Texas Southwestern Medical Center to use the HCC cell lines and publish the data and results herein displayed.

Table 4: End-to-End Time for Alignment and Variant Calling on a 30× Human Data Set

From BCL to VCF (NA12878, 30x)	CASAVA	Isaac
IlluminaCompute standard system	18h 38m	7h 12m

Duplicate removal, indel realignment, and statistics generation were included for each pipeline.

IlluminaCompute standard system: 128G/32 CPU/local raid6, AMD Opteron™ Processor 6212 (Numa)

Using an optimized server (128G/2 CPU/local SSDs, Intel® Xeon® CPU E5-2687@ 3.1GHz) the total Isaac workflow took 4h 29m, Isaac aligner,1h 07m, and Isaac Variant Caller, 0h 59m.

Table 5: Comparison of Mapping and Alignment Accuracy

Quality Metrics	CASAVA	Isaac
% Mapped reads	89.11	93.35
% Mismatch bases	0.56	0.47
Average coverage	38.02	39.97

% Mapped reads: Percent of all passing filter reads, which map to a unique position in the reference genome.

% Mismatch bases: Percent of aligned bases, which do not match the reference. Includes variation and sequencing error.

Average coverage: The average number of uniquely mapped reads covering a position in the reference. All numbers represent an average over the three CEPH trio datasets described earlier.

Table 6: Comparison of Isaac and CASAVA Percent Coverage at Various Sequence Depths*

Quality Metrics	Coverage	CASAVA	Isaac
% ≥ 1× Coverage	Full Genome	98.84	98.87
	Exon	98.77	99.10
% ≥ 10x Coverage	Full Genome	97.27	98.21
	Exon	98.11	98.69
% ≥ 30x Coverage	Full Genome	82.06	85.77
	Exon	83.87	84.73

Percent coverage of exons as determined by RefSeq. RefSeq database is a non-redundant set of reference standards derived from the INSDC databases that includes chromosomes, complete genomic molecules (organelle genomes, viruses, plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.

*All numbers represent an average over the three CEPH trio datasets.

Illumina, Inc. • 5200 Illumina Way, San Diego, CA 92122 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

Pub. No. 770-2013-016 Current as of 18 July 2013

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA