illumina®

Isaac Genome Alignment and Isaac Variant Caller

Ultrafast, accurate, and user-friendly human whole-genome sequencing workflow to align and call variants.

Abstract

Illumina introduces a very fast and accurate resequencing workflow that begins with BCL or FASTQ files and produces BAM and VCF files in just over 7 hours for a 30× genome on commodity computing hardware. This workflow provides a four- to seven-fold speed increase over currently available methods and enables a sample to variant workflow of less than 40 hours, using PCR-free library preparation and sequencing with the HiSeq® 2500 System.

The core algorithms are the Isaac Genome Alignment and the Isaac Variant Caller software[1]. The workflow is available in BaseSpace®, through RapidTrack Services, and on any computer meeting the recommended specifications, including the IlluminaCompute Systems, under the Illumina Open-Source Software License.

Introduction

Illumina technology is now used to sequence a 30× human genome with high base-calling quality in a day, enabling applications that require fast time to analyzed data. For example, speed is paramount in the diagnosis and treatment of certain neonatal conditions and in the selection of treatments for late-stage cancers[2,3]. In recent years, sequencer output has outpaced the evolution of computing power, storage capacity, and software. As a consequence, performing an alignment and variant calling analysis has required a large investment in IT infrastructure (hardware and personnel) to set up analysis and bioinformatics expertise to execute the workflow. The daunting task of managing and analyzing large data sets has been a bottleneck in the utilization of next-generation sequencing technology.

To address this bottleneck, Illumina introduces a user-friendly workflow that enables scientists with no bioinformatics experience to efficiently align and variant call whole human-genome data generated by the HiSeq instrument. The Isaac workflow employs the Isaac Genome Alignment software (referred to as "Isaac Aligner"), Genome Analysis Toolkit (GATK)[4] indel realignment, and the Isaac Variant Caller. In the future, the GATK realignment will be replaced by a realignment step within Isaac. The Isaac workflow can either start with BCL or FASTQ files. The output consists of the realigned and duplicate marked reads in a BAM file, the variants in a VCF file, an additional Genome VCF (gVCF[5]) file that has an entry for every base in the reference (thus differentiating a reference call and a no call), and a summary of the run quality. HiSeq users can access the workflow in BaseSpace at no cost. The intuitive interface available in BaseSpace enables users without bioinformatics expertise or Linux experience to run the workflow. An installer for the software is available upon request for customers who prefer to use the workflow locally. The component algorithms are released under an open-source license[6,7].

Isaac Aligner and Variant Caller

The Isaac Aligner is targeted at DNA sequencing data with low error rates and read lengths (single or paired ends), which would be typical for data produced by the HiSeq platform. The aligner uses a novel indexing scheme to very quickly perform seed searches. It runs on a single node and very efficiently uses CPU power while minimizing input/output.

High-Level Summary of the Isaac Aligner Algorithm

- The complete set of relevant candidate mapping positions is identified. This is a usual seed-based search, using 32-mer seeds. The main increase in speed at this stage comes from sorting the reference index by 32-mers.
- The best mapping among all candidates is selected. After optional trimming of low-quality 3′ ends and adapter sequences, the possible mapping positions of each fragment are compared, taking into account paired-end information when available.
- Alignment scores for the selected candidates are determined. The alignment scores are based on a Bayesian model, where the probability of each mapping is inferred from the base qualities and the positions of the mismatches.
- The final output consists of a sorted, duplicate-marked BAM file and summary statistics.

The Isaac Variant Caller identifies and genotypes single-nucleotide variants (SNVs) and small indels in the diploid genome case. The output in the VCF file captures the genotype at each position, the probability that the consensus call differs from reference, and the probability of the called genotype.

High-Level Summary of the Isaac Variant Caller Algorithm

- A set of possible indel candidates is identified. For example, any gap opened by the aligner is an indel candidate. All reads overlapping the candidates are realigned using a multiple sequence aligner. (Due to this step, Isaac Variant Caller can be used whether or not BAM files have gone through the GATK indel-realignment step.)
- The SNV caller uses a Bayesian model to compute the probability of each possible genotype, given the aligned read data and a prior distribution of variation in the genome.
- A heuristic is used to reflect the dependence structure of reads, i.e., base calls at a particular position are not assumed to be independent.
- A series of filters is applied (e.g., low-quality filter or high-depth filter) and the results annotated in the VCF.

- Indel genotypes are called and probabilities are assigned using a Bayesian model. An approximate model is used to handle overlapping indels.
- The final output is a block-compressed gVCF. This file follows the specifications of the variant call format but provides information for every position in the reference genome.

A more complete description of the Isaac Aligner and Variant Caller algorithms is available upon request (manuscript in preparation).

Isaac Workflow Significantly Reduces Run Time

To assess the performance of the new workflow, the Centre d'Etude du Polymorphisme Humain (CEPH) trio of father (NA12891), mother (NA12892), and daughter (NA12878) was sequenced. Mapping and alignment was performed using three algorithms: Efficient Large-Scale Alignment of Nucleotide Databases (ELAND)—the aligner in Consensus Assessment of Sequence and Variation (CASAVA), Isaac Aligner, and Burrows-Wheeler Alignment (BWA)[8]. After performing indel realignment,variants were called using either Isaac Variant Caller or GATK. Table 1 shows the end-to-end wall clock time for each of the three workflows tested. The Isaac workflow is almost six times faster than BWA and GATK on a standard IlluminaCompute node.

Table 2 illustrates that this gain in speed is achieved without compromising mapping quality, as the Isaac Aligner produces comparable values for various quality metrics such as percent reads mapped, percent mismatch to the reference, and average coverage by uniquely mapping reads.

Isaac Workflow Provides High-Accuracy SNV and Indel Calling

To evaluate the quality of variant calling, 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.

The results are provided in Table 3. For most metrics, the Isaac workflow provides values in line with the results from BWA and GATK. One exception is the higher heterozygote to homozygote ratio for both variant classes. Manual review of a subset of the SNV calls using a genome browser suggests that many of the heterozygous calls are correct, so there is likely to be some heterozygous undercall in the BWA and GATK workflow and some overcall in Isaac. The ratio of 1.62 for SNVs is also consistent with findings using an orthogonal platform, which determined a range of 1.25–1.7 for European populations.[9]

Table 1: End-to-End Time for Alignment and Variant Calling on a 30× Human Data Set

From BCL to VCF (NA12878, 30x)	CASAVA	Isaac	BWA+GATK
IlluminaCompute standard system	18h 38m	7h 12m	41h 18m

Duplicate removal and indel realignment were included for each pipeline. An additional 2 hours was added in BWA+GATK for converting BCLs to FASTQ before alignment.

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

On an optimized server (128G/2 CPU/local SSDs, Intel® Xeon® CPU E5-2687 @3.1 GHz), the total Isaac workflow took 3h 40m.

Table 2: Comparison of Mapping and Alignment Accuracy

	ELAND	Isaac	BWA
% Mapped reads	89.11	94.98	95.17
% Mismatch bases	0.56	0.36	0.37
Average coverage	36.2	35.8	36.1

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

% Mismatch bases: Percent of aligned bases that 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. Note that ELAND does not trim reads, while Isaac Aligner and BWA do. This may explain the higher average coverage of ELAND, given the lower fraction of mapping reads.

Table 4 provides additional quality metrics. Measurement of the specificity of variant calling is based on the assumption that the three analyzed samples form a trio of two parents and a child. With the exception of 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 was used as a proxy for false positives. This is a rough approximation, because failure to call variants in the parent can also create a conflict, and not all false positive calls lead to a conflict. However, it is a reasonable metric when used to compare multiple workflows. To measure sensitivity, a set of well-characterized variants was used as reported in Kidd et al[10], who also analyzed sample NA12878. It was assumed that those variants are correct and the ability to detect them was quantified with each of the analysis workflows. The last column in the table is concordance. This measures the agreement of SNV calls in the sequencing data with calls made using a high-density microarray. For all metrics, the results are comparable for the three workflows, though the BWA and GATK workflow has slightly higher sensitivity in these data sets. For indels, the sensitivity gain comes at the expense of specificity.

Table 3: SNV and Indel Call Quality and Statistics

	Quality Metrics	CASAVA	Isaac Workflow	BWA+GATK
Variant Class	Call Rate	95.90%	96.69%	97.39%
SNV	Total SNVs	3,434,048	3,390,119	3,279,583
	Ti/Tv	2.03	2.01	2.10
	Het/Hom	1.57	1.62	1.38
	Novelty Rate	5.5%	4.9%	3.5%
Indel	Total Indels	400,247	345,681[a]	364,389
	Het/Hom	1.62	1.83	1.32
	Novelty Rate	19.9%	16.4%	15.8%

Metrics calculated as the average across a CEPH trio (NA12878, NA12891, and NA12892)

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 FILTER key of VCF file

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

Ts/Tv: Transition to Transversion ratio of SNV calls

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

Novelty Rate: percent of called SNVs or Indels not found in dbSNP 132

a: There are two contributing factors to the drop in indels relative to CASAVA. First, there is a filter which fails indels called in long homopolymer stretches. Second, we have yet to integrate a large variant caller used in CASAVA, so the indel calls are limited to roughly 10 bp or fewer.

Table 4: Comparison of SNV Detection

	Specificity		Sensitivity		Concordance
	SNV	Indel	SNV	Indel	Sites
CASAVA	99.84%	97.38%	90.5%	43.3%	99.45%
Isaac	99.87%	97.86%	90.4%	41.0%	99.99%
BWA+GATK	99.88%	95.88%	91.3%	42.7%	99.45%

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

Isaac Workflow Generates Call Sets with High Overlap to Other Workflows

In addition to computing summary metrics for variant calls, the overlap of variant calls was also 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 has significant overlap. 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 Isaac workflow with the CASAVA workflow. There is a high level of agreement for SNVs. The lower agreement in indel calling reflects the fact that indel-calling methods are not yet as mature as SNV-calling methods.

Figure 2 compares the Isaac workflow with the BWA and GATK workflow. As expected, the overlap between SNV calls is again higher than that between indel calls. In this comparison, the quality metrics of the SNV calls unique to each workflow are significantly different. Specifically, the transition to tranversion ratio, the heterozygote to homozygote ratio, and the fraction of calls not found in dbSNP (novelty rate) are all much higher with Isaac. One hypothesis is that the variant quality recalibration step applied in GATK leads to a bias against novel calls because the method is trained on known variants available in dbSNP. An alternative hypothesis is that the unique SNVs identified by Isaac are false positives. However, the transition to transversion ratio of 2.08 is in line with expectations under the assumption that the variants are real. The high heterozygote to homozygote ratio is also expected, because novel SNVs are expected to be rare in the population and therefore to be observed mostly as heterozygotes.

Figure 1: Variant overlap: CASAVA versus Isaac

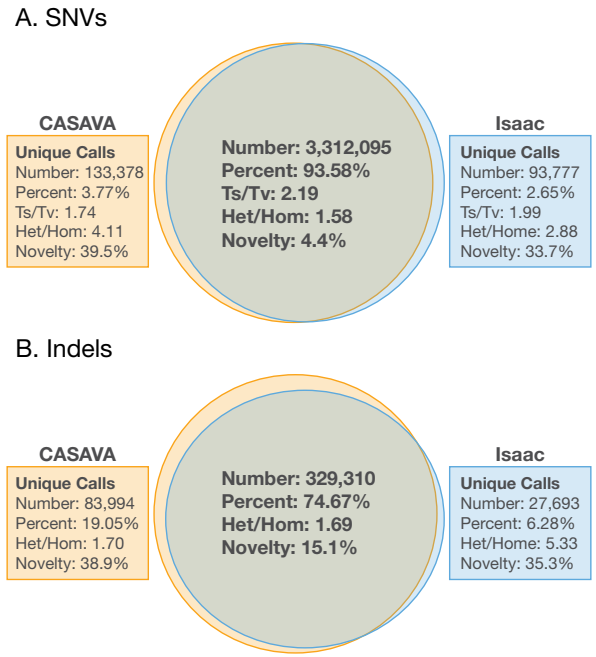
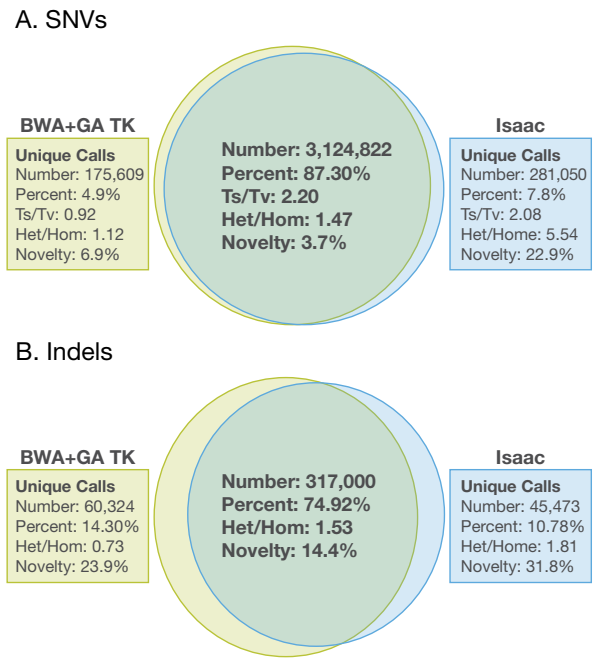


Figure 2: Variant overlap: BWA and GATK versus Isaac



Conclusions

The Isaac workflow efficiently and accurately aligns and calls variants for whole human genome resequencing data. The workflow is almost six times faster than the most popular workflow currently in use. The quality of results is similar across workflows, both for high-level metrics and direct measures of variant overlap. The workflow is freely available in BaseSpace, through RapidTrack services, and will be made locally available in the near future. Illumina will continue to improve the workflow in both performance and quality of results. Future releases will focus on improvements in the calling of indels and the integration of methods that call structural variants.

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. Saunders CJ, Miller AM, Soden SE, Dinwiddie DL, Noll A, et al. (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units Sci Transl Med 4: 154ra135
3. Jones SJM, Laskin J, Li YY, Griffith OL, An J, et al. (2010) Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol 11: R82.
4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis C, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
5. gVCF: https://sites.google.com/site/gvcftools/home/about-gvcf
6. The supported Isaac workflow installer will be posted on MyIllumina.com and the open source versions of Isaac Genome Alignment Software and Isaac Variant Caller are available at https://github.com/sequencing
7. The most recent Illumina Open Source Software License can be found at https://github.com/sequencing/licenses/
8. Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754-1760.
9. http://media.completegenomics.com/documents/GSG-FirstLookatData_June2011.pdf
10. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Nature 453: 56-64.

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. 770-2013-009 Current as of 24 November 2014

illumina®