

# Complete Secondary Analysis Workflow for the Genome Analyzer

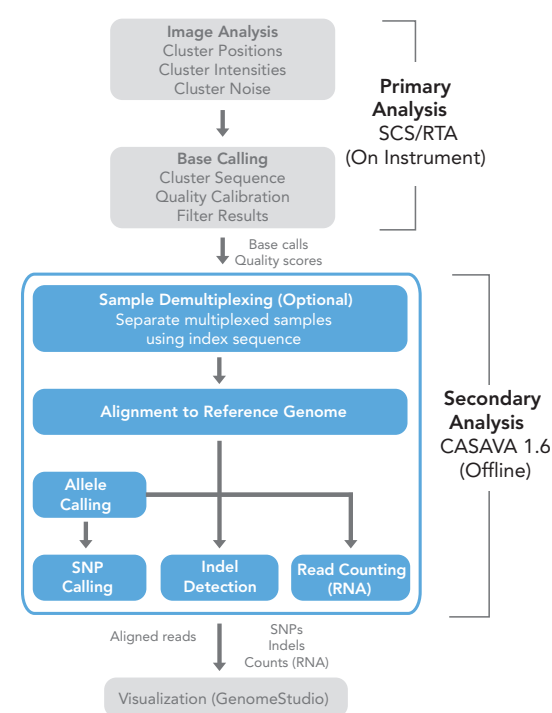
CASAVA 1.6 is a complete secondary analysis package for processing reads from the Genome Analyzer with facilities for alignment, reference-guided assembly, SNP and indel calling, and counting for RNA applications.

## INTRODUCTION

Recent hardware and software improvements have resulted in a significant increase in the data output capacity of the Illumina Genome Analyzer<sub>Hi</sub>. Greater than 33 Gb of sequence output, consisting of approximately 300 million 2 x 100 bp reads, can now be routinely generated within 10 days. The Consensus Assessment of Sequence And Variation (CASAVA) software seamlessly processes this large volume of sequencing data, supporting sequencing of large or small genomes, targeted DNA resequencing, and RNA sequencing. Illumina now offers CASAVA 1.6, with the ability to process multiplexed sample Genome Analyzer runs, a revised alignment method, improved SNP calling, and a new indel detection algorithm.

As shown in Figure 1, sequencing images generated by the Genome Analyzer are analyzed in two steps. In the first step (primary analysis), the Sequencing Control Software Real Time Analysis (SCS/RTA), which runs on the instrument computer, performs real-time image analysis and base calling. In the second step, CASAVA performs complete secondary analysis. CASAVA is a cascade of steps comprising alignment to a reference sequence, allele calling, SNP and insertion/deletion (indel) detection, and counting for RNA. CASAVA automatically processes multiplexed sequencing runs, in which >96 samples can

FIGURE 1: ILLUMINA SEQUENCING DATA ANALYSIS WORKFLOW



Primary analysis is performed by the on-instrument software SCS/RTA. Secondary analysis is initiated by sample demultiplexing (for multiplexed sequencing runs only), alignment to a reference genome, allele/SNP calling, indel detection, and read counting for RNA sequencing.

be separately sequenced by associating unique indexed reads with each sample. CASAVA output can be read into the GenomeStudio® visualization and analysis software for interpretation of results.

## CASAVA 1.6 HIGHLIGHTS

- Complete secondary analysis software package, including facilities for alignment, reference-guided assembly, SNP/indel calling and counting for RNA applications
- New methods for gapped, multiseed alignments that reduce artifactual mismatches
- New indel detection method, and improved SNP calling on longer Genome Analyzer reads

CASAVA WORKFLOW DETAILS

**Demultiplexing**

Using multiplexed sequencing, 12 individual samples can be run in a single lane, for a total of 96 samples per eight-lane flow cell. Each sample is identified as a separate read by an index sequence integrated into the library constructs during sample preparation.

For multiplexed samples, demultiplexing is the first step in the CASAVA workflow. Index sequence information, stored in a “samplesheet” file in .csv format, is used to create a separate BaseCalls directory for each indexed sample. The directories also contain the requisite files for downstream analysis by CASAVA. The demultiplexing step can detect and tolerate a single mismatch or error in the index sequence. This process increases robustness, ensuring that good reads are not discarded simply because of a mismatch or error in the index sequence.

**Alignment**

Efficient Large-Scale Alignment of Nucleotide Databases (ELANDv2) is used to align reads generated by the SCS/RTA software to a reference sequence. In addition, CASAVA generates ELANDv2 error statistics and diagnostic plots to assess data quality. It is important to note that

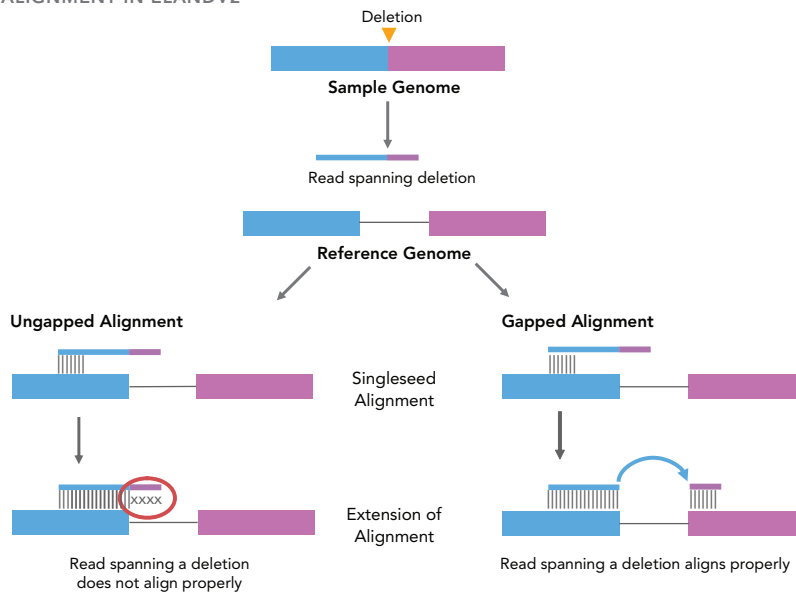
in previous versions of the Illumina secondary analysis workflow, ELANDv1 was available as a part of the Pipeline software. ELANDv1 is now deprecated, and should not be used for alignment.

For each read, ELANDv2 determines positions in the genome to which the read substrings (seeds of 16 to 32 bp) match with a maximum of two errors. Several additional bases before and after the match position are associated with the seed to account for potential gaps during the alignment phase. A global alignment is computed between the read and the reference using the entire read, constraining the maximal gap length to a user-defined size.

Base quality values and the positions of the mismatches in a candidate alignment are used to calculate a probability score (p-value) for each candidate. The p-value is the probability that the candidate position in the reference genome would give rise to the observed read, if its bases were sequenced at error rates that correspond to the read’s quality values. The contribution of each base is weighted according to its quality.

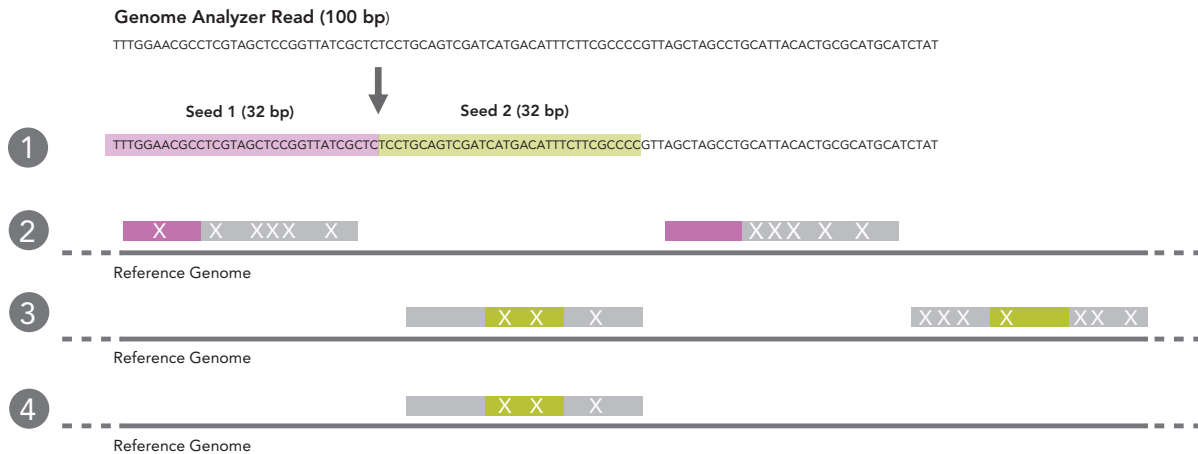
The alignment score of a read is computed by taking the p-value of the candidate and dividing it by the sum of the p-values of all the candidates. The highest p-value in-

FIGURE 2: GAPPED ALIGNMENT IN ELANDV2



An ungapped aligner does not take into account gaps (indels) in the reads to be aligned, resulting in multiple mismatches downstream of the gap (red circle, left). The gapped alignment in ELANDv2 properly accounts for indels such as a deletion in the sample genome, as shown at the top. The gapped alignment results in no mismatches downstream of the gap (right).

FIGURE 3: MULTISEED ALIGNMENT OF 100 BP GENOME ANALYZER READS



The GA read is divided into seeds of 32 bp each (1). Seed 1 is aligned using gapped alignment to multiple candidate regions in the reference genome, allowing for up to two mismatches (X) in the seed, then is extended to the full read, allowing for any number of mismatches (2). Seed 2 is aligned in the same way (3). The best alignment among the multiple alignment positions is chosen based on quality scores (4). A maximum of four seeds are allowed, but this example shows a two-seed alignment.

dicates the best candidate. This is also known as a Bayes' Theorem inversion. The alignment score is expressed on the Phred scale (i.e., the Q20 score corresponding to 1% chance of incorrect alignment, and the Q30 score corresponding to 0.1%). Figure 2 shows that proper accounting of indels during alignment reduces the number of mismatches downstream of the gap position.

### Multiseed Alignment

Multiseed alignment is initiated with the creation of multiple seeds— contiguous, non-overlapping sections of 16 to 32 bp, within the read. Up to four seeds is allowed in ELANDv2, so that up to three 32 bp-seeds are possible for 100 bp Genome Analyzer reads. The seeds are aligned to multiple positions in the reference genome, with a maximum of two mismatches per seed allowed for each candidate seed alignment (Figure 3). The candidate alignments are extended to the full length of the read using gapped alignment, and the best alignment is picked using alignment scores. For paired reads, the first and second reads are aligned separately, and the pairing information is used to choose the best pair of reads among the candidate read pairs. To optimize memory usage and computation time, the following steps are executed:

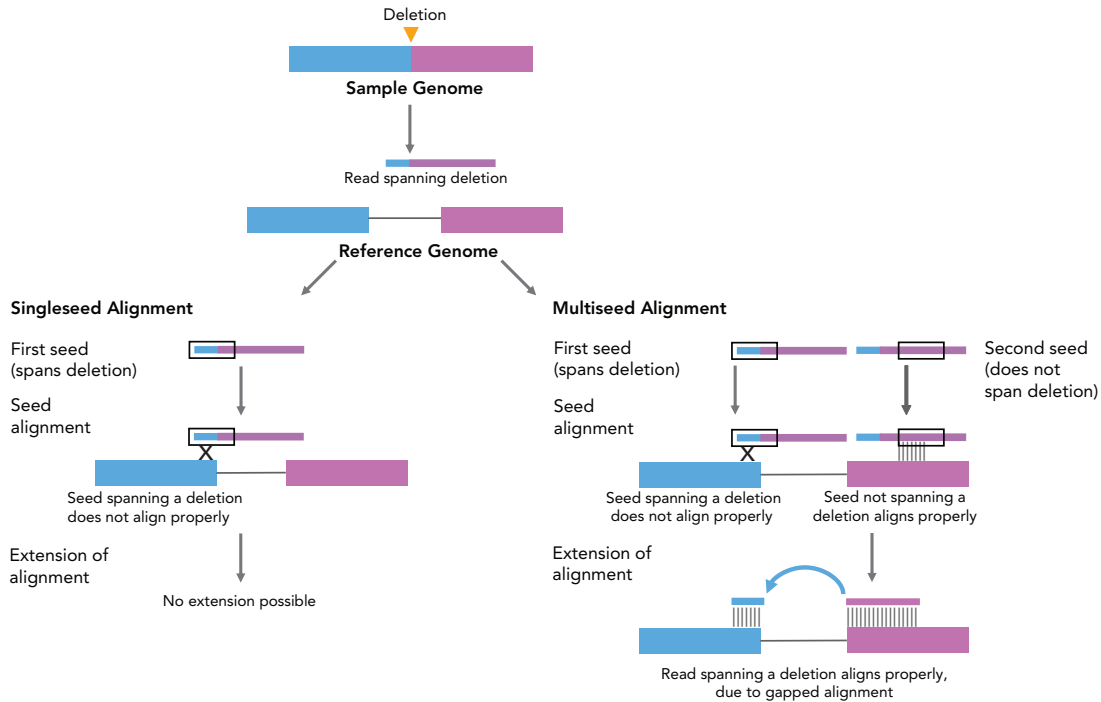
- Singleseed alignment is performed on all reads
  - Reads that do not align are flagged
  - Multiseed alignment is performed on the flagged reads
- Note that gapped alignment and multi-seed alignment are two distinct features that work independently of each other. However, the combination of gapped and multi-seed alignment improves the handling of indels and mismatches, resulting in better alignment, as shown in Figure 4. Better alignments ultimately result in higher yield and genomic coverage per Genome Analyzer run, because good reads that would otherwise be discarded due to indels and mismatches are retained in the analysis.

### Allele Calling

Read pairs that mapped with the expected size (within three standard deviations of the median) and orientation are used for allele calling. The base calls and their associated quality values are sent to a Bayesian allele caller, which produces one or two allele calls and scores for each position in the genome. The allele call score can be approximated as a Phred score divided by 10 (e.g. an allele score of 3 corresponds to a Phred score of 30).

In CASAVA 1.6, a new mismatch density filter reduces noise in longer gapped reads. The new filter has the following features:

FIGURE 4: MULTISEED AND GAPPED ALIGNMENT IMPROVE HANDLING OF INDELS AND MISMATCHES



In this example, the first seed spans the deletion region of the read, while the second seed is completely in the post-deletion region of the read. In singleseed alignment (left), only the first seed is available. The first seed will not align properly with the reference genome (due to the presence of the deletion in the seed). The seed therefore cannot be extended to complete the alignment. In multiseed alignment (right), the first seed leads to a failed alignment as before. However, the second seed aligns well with the post-deletion region of the reference genome. Using gapped alignment, the read can be aligned across the indel region of the reference genome.

- Base calls are ignored where more than two mismatches to the reference sequence occur within 20 bases of the call
- The mismatch limit is applied to the 41 base window at the corresponding end of the read if the call occurs within the first or last 20 bases of a read
- The mismatch limit is applied to the entire read when the read length is 41 or shorter

### SNP Calling

Homozygous SNPs are called at positions where a non-reference allele is observed, the allele call score is  $\geq 10$ , and the depth is no greater than three times the chromosomal mean. For heterozygous SNP calls, there is an additional requirement that the second-highest allele call score be  $\geq 6$ , and the ratio of the highest to the second-highest allele-call scores be  $\leq 3$ . The allele call score cutoff ensures that more than the equivalent of three Q30 bases are needed to make a SNP call. The ratio cutoff distin-

guishes between genuine heterozygous SNPs and any residual background noise, especially for regions with extremely high coverage such as mitochondria in the human genome. The default parameter values mentioned here are recommended, but can be modified.

These changes to allele and SNP-calling in CASAVA 1.6 are designed to improve allele calling and heterozygous SNP sensitivity from longer reads with gapped alignments.

### Indel Detection

Anomalous reads from paired-end alignment are used to detect consensus insertions and deletions. Reads with poor or non-existent alignment due to poor base quality scores are excluded from the analysis. The indel detection is based on the concept of shadow/singleton read pairs, where one “singleton” read in a pair can be confidently aligned, but not the other. Semi-aligned reads with poor ungapped alignment based on the base quality scores are

also used for indel detection. Depending on read length and insert size, indels up to one hundred base pairs can be detected. Indel detection steps are:

- Shadow reads are re-aligned using gapped alignment
- Using the singleton read positions that the shadow reads pair to as a distance metric, clusters of non-aligned “shadow reads” and semi-aligning reads are generated
- Reads in each cluster are assembled into contigs
- Contigs are aligned to the genome, using the positions of associated “singleton” reads to narrow the search to several hundred base pairs
- The diploid genotype for each indel is called using both the reads from the indel contig assembly and any reads from the reference assembly which intersect the indel position

The indel genotype caller provides quality scores (Q-scores) of the indel and the most likely genotype for each call.

**Read Counting for RNA Sequencing**

RNA sequencing analysis in CASAVA supports whole-transcriptome sequencing projects. Exon, splice junction, and gene counts can be used to determine gene expression levels and expressed splice variants. Splice junction sets and non-redundant exon sets for human, mouse, and rat analyses are provided with CASAVA. Additional scripts and documentation are available to prepare these files for other organisms.

Prior to read counting, the RNA alignment mode in ELANDv2 aligns each read against contaminants,

genomes, and splice junctions, in that order. Reads are discarded if they align to the contaminants reference file. A script determines whether the read is a unique alignment against either genomes or splice junctions. If the read is considered neither, it is either discarded or marked as “not matched”.

CASAVA counts the number of bases, not the number of reads, that belong to exons and genes. For splice junctions, the number of reads that cover the junction point is counted. The number of bases that fall into the exonic regions of each gene is summed to obtain gene level counts. Normalized values are calculated as RPKM (reads per kilobase of exon model per million mapped reads)<sup>1</sup>. The RPKM calculation for exons and genes is slightly different than the RPKM calculation for splice junctions, because in the former case, base counts rather than read counts are used.

$$\text{Exons and Genes RPKM} = 10^9 \times \text{Cb/NbL}$$

Cb = the number of bases that fall on the feature

Nb = total number of mapped bases in the experiment

L = the length of the feature in base pairs

The normalized values for splice junctions are counted as described<sup>1</sup>:

$$\text{Splice Junctions RPKM} = 10^9 \times \text{Cr/NrL}$$

Cr = the number of reads that cover the junction point

Nr = total number of mapped reads in the experiment

L = the length of the feature in base pairs

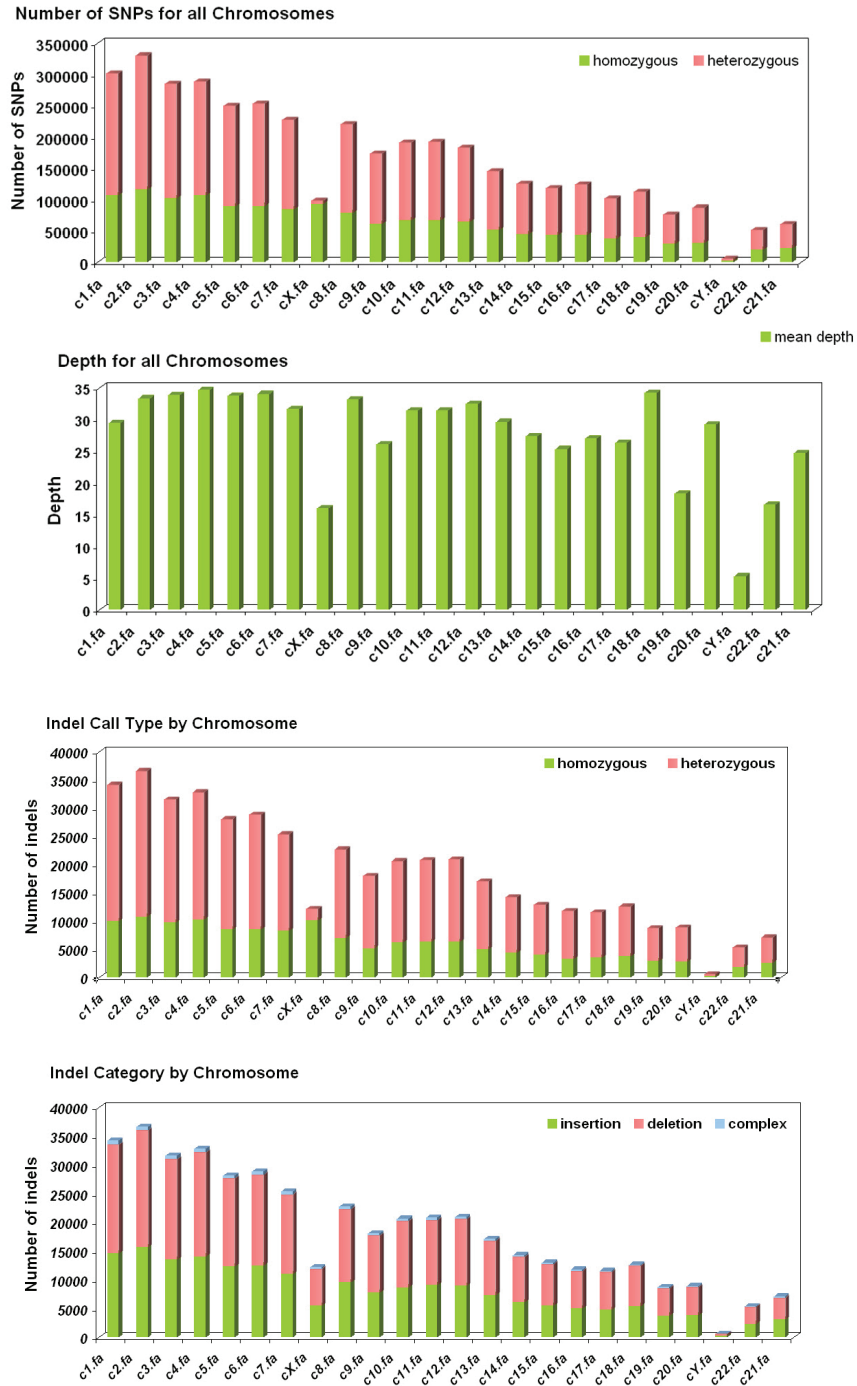
Exons that include overlapping exons from other genes on the forward or reverse strands are excluded from counting and are not included in calculating the total gene length for normalization.

FIGURE 5: SNP CALLS FOR CHR 15 GENERATED BY CASAVA

	Position of the SNP on the indicated chromosome	A	B	C	D	E	Base called	Total bases called at that position	Bases used for making the call	Score of the call(s)	reference	Call type
	#position	A	C	G	T	modified_call	total	used	score			type
1	18260646	0	34	0	0	C	36	34	112.79		T	SNP_diff
2	18261869	0	0	13	0	G	13	13	41.19		T	SNP_diff
3	18262422	0	0	9	26	TG	40	35	82.78:28.20		T	SNP_het1
4	18262476	12	0	0	37	TA	54	49	104.98:36.99		T	SNP_het1
5	18262564	0	31	0	16	CT	50	47	94.33:53.20		C	SNP_het1
6	18263563	0	0	36	0	G	42	36	128.39		C	SNP_diff
7	18264404	0	0	28	0	G	30	28	83.48		T	SNP_diff

CASAVA generates a formatted text file with information about SNP positions, the bases called, the total number of bases used to make the call, the call score, and the type of call.

FIGURE 6: CASAVA STATISTICS OUTPUT



CASAVA generates graphs for statistics such as heterozygous/homozygous SNP ratios, mean depth of coverage per chromosome, indel call types, and indel categories.

TABLE 1: SNP STATISTICS STUDY FROM TWO ANALYSESE OF SAMPLE FROM HAPMAP INDIVIDUAL NA18507

STUDY*	ALL	HET	HOM	AUTOSOMAL HET/HOM SNP RATIO	IN DBSNP % ALL**	IN DBSNP % HET	IN DBSNP % HOM
Bentley et al., 2008	3,828,342	2,411,022	1,417,320	1.811	82.39	75.17	94.66
CASAVA 1.6	4,008,522	2,518,681	1,489,841	1.800	82.70	75.40	95.04

\*The read depth is 32 for both datasets.

\*\*"In dbSNP%" refers to the percentage of identified SNPs that match previous entries in dbSNP. The SNP analysis in the Bentley et al. paper was carried out using dbSNP release 128. The analysis reported here was carried out using dbSNP release 129, which is the latest available version before the SNPs from the Bentley study were included.

### CASAVA OUTPUT

CASAVA output is placed into a folder structure (called a CASAVA build), with sub-folders for each chromosome. For each chromosome folder, the following files are generated, and sorted by respective chromosome positions:

- *Sequence files*: divided into 50 MB bins containing experimental details including machine name, run number, lane and tile information, reads and quality scores, and chromosome positions
- *SNP file*: containing alleles called at each SNP position, call scores and other information (Figure 5)
- *Indel file*: containing indel locations, sizes, call types, and contexts
- *RNA count files*: containing three files per chromosome: for the exon, gene, and splice counts, each containing start and end positions, gene symbols, and absolute and normalized counts

The CASAVA build containing the reads and allele calls can be imported directly into GenomeStudio for the visualization and interpretation of results.

CASAVA automatically generates per-chromosome statistics in individual files, and graphical summary statistics in .html reports. These statistics include:

- Duplicates statistics
- SNP statistics such as number of heterozygous (het) and homozygous (hom) SNPs (Figure 6)
- Mean depth of coverage (Figure 6) and percentage chromosome coverage
- Indel call type (het/hom, Figure 6), indel categories (insertion, deletion or complex, Figure 6)
- Exon, gene and splice counts per chromosome

### RESULTS

The SNPs reported in the Bentley et al., 2008 paper for a sample from HapMap individual NA18507, Yoruban male, are shown in Table 1 (top row). The same sample was sequenced using 37 lanes of 100 bp paired-end reads generated from five flow cells, yielding 142.96 Gb of sequence. Gapped, singleseed alignment using ELANDv2, followed by allele and SNP calling was carried out using the new reads in CASAVA 1.6. SNP statistics from this secondary analysis are shown in Table 1 (bottom row). Notably, the new analysis using CASAVA 1.6 calls an additional ~180,000 SNPs while maintaining a slightly higher concordance to entries in the dbSNP public database.

### SUMMARY

CASAVA 1.6 is a complete secondary analysis package for processing reads from the Genome Analyzer, including facilities for alignment, reference-guided assembly, SNP/indel calling, and counting for RNA applications. Recent alignment enhancements including gapped and multi-seed alignment reduce artifactual mismatches between called bases and the reference sequence. Improved SNP calling from longer reads (100 bp) and a new indel-detection method make CASAVA 1.6 a strong addition to the large collection of open source and commercial software products currently available for secondary analysis of Genome Analyzer reads.

#### REFERENCES

- (1) Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):621–628.
- (2) Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456 (7218):53-59.

#### ADDITIONAL INFORMATION

For more information about Illumina systems and software for sequencing, visit [www.illumina.com](http://www.illumina.com) or contact us at the address below

**Illumina, Inc.**  
**Customer Solutions**  
9885 Towne Centre Drive  
San Diego, CA 92121 USA  
1.800.809.4566 toll-free  
1.858.202.4566 tel  
[techsupport@illumina.com](mailto:techsupport@illumina.com)  
[www.illumina.com](http://www.illumina.com)

---

#### FOR RESEARCH USE ONLY