Calling Sequencing SNPs

Illumina provides a SNP caller in the CASAVA software that identifies SNPs in RNA or DNA sequencing experiments.

Introduction

Next-generation sequencing provides a powerful way to identify novel single nucleotide polymorphisms (SNPs) and call known SNPs in genome or transcriptome samples. Accurately calling these SNPs requires high-quality sequencing data, high coverage, and a thorough bioinformatics approach to identify the SNPs in a statistically relevant manner.

Illumina's Genome Analyzer provides the high throughput and high-quality data necessary for efficient DNA and RNA sequencing experiments. Illumina's bioinformatics solutions for DNA and RNA sequencing consist of the Genome Analyzer Pipeline software that aligns the sequencing data, the CASAVA software that assembles the reads and calls the SNPs, and the GenomeStudioTM DNA and RNA Sequencing Modules that enable visualization and analysis of the SNPs.

This technical note explains Illumina's SNP caller algorithm in the CASAVA software (Figure 1).

Figure 1: SNP Caller in Illumina's DNA and RNA Sequencing Workflow

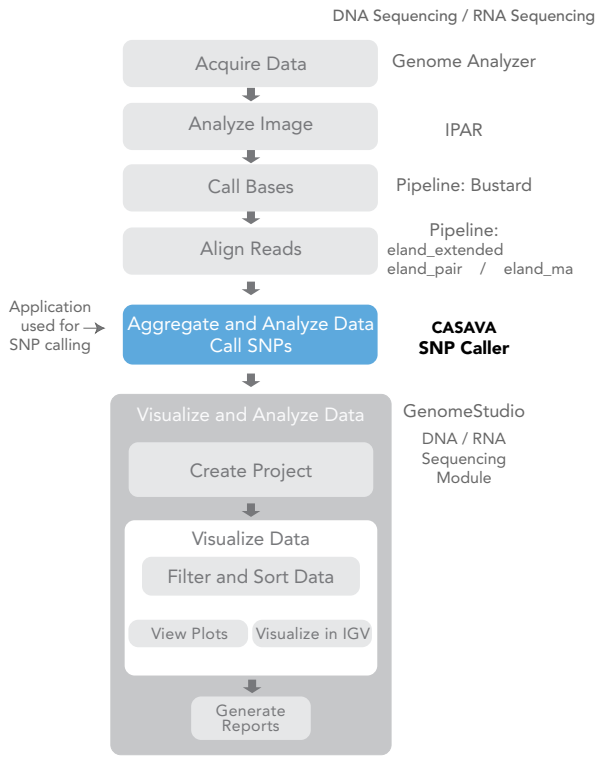
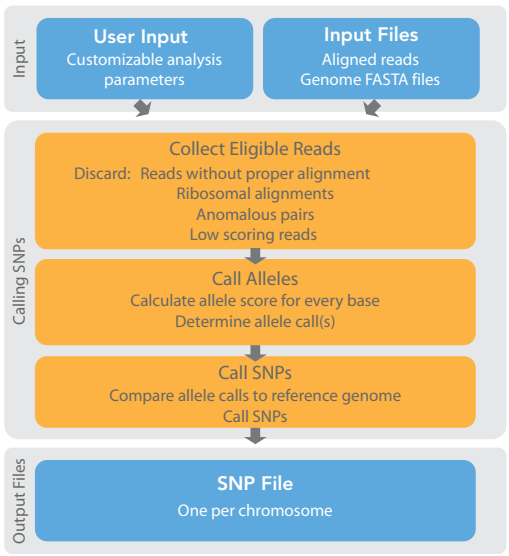


Figure 2: SNP Calling Workflow



The SNP Caller Algorithm

The SNP caller algorithm is described below; an overview is given in Figure 2.

Collecting and Selecting Eligible Reads

The SNP caller uses reads that CASAVA collects from the different sequencing runs as primary input. CASAVA selects the reads that match any of the reference sequence (chromosome.fa). Reads with no match or matching ribosomal DNA (for RNA sequencing) are not used by the SNP caller.

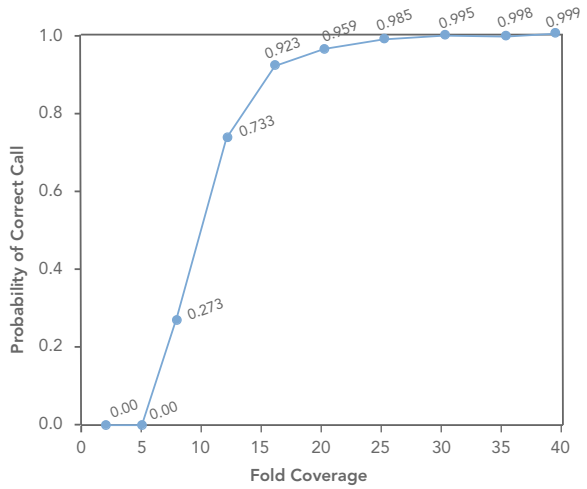
For paired-end runs, files contain reads from normal pairs, anomalous pairs, and orphan reads. Only reads from normal pairs are used for SNP calling. Read pairs need to map with the expected size (within three standard deviations of the median for the default setting) and orientation to be used for allele calling. Paired reads also are required to have a minimum paired read alignment score of 6 and single reads are required to have a minimum alignment score of 10.

Calling Alleles

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 caller computes log10 P(observed base | no "A"s are present) for A and similarly for C, G, T, at each position. The scores are then normalized by subtracting the lowest score from each of the other three and converted to a log-odds score. This score is the allele call score and a score of 3 is equivalent to a Phred score of Q30.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 3: Probability of Correct SNP Call
Calculation of the probability of a correct SNP call at different coverage levels for a theoretical heterozygote position. The quality of the base calls was assumed at Q30.



Calling SNPs

In the default setting, a SNP is called if the following conditions are met:

- **A non-reference base allele is observed**
- **The allele call score is ≥ 10**
- **For DNA sequencing, the depth at this position is no greater than three times the chromosomal mean (there is no coverage cutoff for RNA SNP calling because the reads have much greater depth)**
- **For heterozygous calls, both alleles should have an allele-call score ≥ 10, and the ratio of their scores should be ≤ 3**

The allele call score cutoff ensures that more than the equivalent of three Q30 bases are used to make a SNP call. The ratio cutoff ensures that genuine heterozygous SNPs and any residual background noise can be distinguished, especially for extremely high coverage (e.g. mitochondria in the human genome).

Customizing The SNP Caller

The SNP caller can be customized by changing the following parameters.

SNP Threshold

The SNP threshold is the minimum allele call score required to call a SNP. For a heterozygous SNP to be called, the score for both alleles must exceed this value. The default threshold is 10.

SNP Max Ratio

The SNP max ratio is used to evaluate possible heterozygous SNPs. This sets the maximum ratio between the first and second allele call scores. Situations where the major allele is much stronger than the minor allele should be called as homozygous SNPs, as the minor allele may simply be noise. The default value is 3.

Read Mode

The SNP caller can be used for single-read or paired-end data. Note that the Pipeline alignment module for RNA sequencing, eland_rna, currently does not support paired-end analysis.

Quality Value Cutoff

This cutoff sets the minimum quality value for the read to be considered. The default cutoff is 6.

SNP Coverage Cutoff

SNPs are called only at the positions where the depth is no greater than the SNP coverage cutoff times the chromosomal mean depth. This prevents SNP calling in regions with extreme depth, such as near the centromere of a human chromosome. The default cutoff is 3. SNP coverage cutoff is turned off for RNA SNP calling.

Characteristics of the SNP Caller

Depth of Coverage

The SNP caller needs a minimal level of coverage to be reliable. A mean coverage of 30× is recommended for DNA sequencing. This will lead to confident SNP scores and tolerates areas with somewhat lower coverage.

The required minimum coverage is illustrated in Figure 3, which displays a graph of calculated theoretical probabilities of correct SNP calls at different coverages at Q30. The probability of a correct call is > 0.99 for 30× coverage, while 20× and 25× coverage still deliver reasonably confident SNP scores (probabilities > 0.95 and > 0.98).

The calculated required depth of coverage was confirmed by analysis of sequencing data from a Yoruba male. When SNP caller was run at different depths of coverage for chromosome 2, SNP calls for heterozygote positions accumulated up to a depth of coverage of 33×1.

Validation of SNP Caller

SNPs called after Illumina sequenced a Caucasian X chromosome and the entire genome of a Yoruba male validated the SNP caller approach. There was excellent agreement of the SNP calls with microarray-based SNP genotyping results of the same samples1.

For the X chromosome, 99.77% of the 13,604 X chromosome loci of HumanHap550 BeadChip were covered and excellent agreement between sequence-based SNP calls and genotyping data was found (99.52%). There was complete concordance of all homozygous calls and a low level of 'under-calling' from the sequence data (denoted as 'GT > Seq' in Table 1) at a small number of the heterozygous sites. This was likely caused by inadequate sampling of one of the two alleles.

For the Yoruba male, a total of ~4 million SNPs were called, with 74% matching previous entries in dbSNP. Sequence-based SNP calls covered almost all of the 552,710 loci of HM550, with 99.57% concordance between sequencing and genotyping calls (Table 1). The few disagreements were mostly under-calls of heterozygous positions (GT > Seq) in areas of low sequence depth, providing a false-negative rate of 0.35%.

Table 1: Comparison of SNP Calls From Sequencing Versus Genotyping Data

Sample	X Chromosome	Yoruba Male
SNPs	13,604	552,710
Covered by sequence	99.77	99.60
Concordant calls	99.52	99.57
All disagreements	0.48	0.43
GT > Seq	0.48	0.35
Seq < GT	0	0.05
Other discordances	0	0.03

Limitations of the SNP Caller

The SNP caller has the following limitations:

- **It is tuned to human genome analysis (i.e. diploid genome).**
- **It needs approximately 30× coverage (i.e., 15× per allele).**
- **The SNP caller is not strand aware.**
- **The SNP caller may identify some indels as SNPs. An indel that has erroneously been called a SNP can be recognized by a concomitant drop of coverage, since ELAND does not align reads that span an indel well.**
- **Casava v1.0 only uses reads mapped to the genome for SNP calling, not those spliced over two exons. Allele frequencies for RNA SNPs close to splice sites may be inaccurate.**
- **In default mode, the SNP caller does not call SNPs in high-copy regions. These regions can be prime research targets in areas such as cancer research.**

Conclusion

The Illumina SNP caller accurately calls statistically relevant SNPs from the high throughput, high-quality sequencing data generated by the Genome Analyzer.

Reference

1. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 970-2008-028 Current as of 10 March 2009

illumina®