illumina®

Somatic Variant Caller

Detect mutations below 5% frequency with TruSeq® Amplicon - Cancer Panel.

Introduction

The highly multiplexed TruSeq Amplicon - Cancer Panel (TSACP) targeted resequencing assay for the MiSeq® system can detect somatic mutations across hundreds of mutational hotspots in cancer genomes. With a streamlined workflow and highly specific amplicon targeting, this sensitive assay is uniquely suited for detection of somatic mutations in formalin-fixed, paraffin-embedded (FFPE) samples[1].

Illumina developed a new variant calling algorithm specifically for TSACP to detect low-frequency mutations (even below 5%) in a mixed cell population. The somatic variant caller is built into the MiSeq Reporter v1.3+ software and is also available in BaseSpace™. It reports all mutations, providing researchers with the flexibility to evaluate both confident and near-confident scoring mutations. This technical note provides information about the somatic variant caller algorithm and how to employ it for data analysis.

Somatic Variant Calling in Cancer Samples

Deep sequencing of multiplex amplicon libraries generated from FFPE cancer tissue samples is sensitive, reproducible, and accurate using TSACP and the MiSeq system. The data generated reflects the heterogeneity of the tumor, which includes some normal cells, some cells at an early stage of cancer progression (with fewer mutations), and some late-stage cells (with more mutations)[2]. Somatic variant calling is extremely important for characterizing cancer samples, yet inherently difficult because somatic mutations often appear at very low frequencies. For example, a SNP might have a very low variant allele frequency (VAF) and be seen in < 10% of the reads covering a given base.

Methodology and Usage

The goal of the somatic variant caller is to identify somatic variants with high confidence to minimize spurious false positives that less stringent analyses can report. For each variant, the somatic variant caller notes the reference and variant sequences (e.g., A to C mutation, or AG to A deletion), as well as the frequency, and provides a quality (Q) score indicating confidence that the variant is indeed present in the sample. Variants are reported in .VCF format (a standard tab-delimited format for storing variant calls), which can then be displayed in a variety of tools, including MiSeq Reporter. The somatic variant caller was designed for instances where a single cancer sample has been sequenced; it is not intended for analyzing tumor/normal pairs.

For SNP calling, the somatic variant caller considers each position in the reference genome separately, starting with the bases of the aligned reads. Alignments come from the TruSeq Amplicon aligner (banded Smith-Waterman). Only base calls above Q20 (> 99% estimated accuracy) are used and variants are only called for a base position that has depth of coverage 10 or greater (Table 1). Finally, based on the reference and mutant counts, a variant score (accuracy of the call) is computed based on a Poisson model that excludes the variant if its Q score is below 20 (1/100 chance of being false positive).

The somatic variant caller handles indels similarly, analyzing the number of alignments covering a given position that include a particular indel (the variant count) versus the overall coverage at that position. Note that the somatic variant caller does not perform some of the indel re-alignment or “indel cleaning” steps included in other variant callers, such as GATK.

Table 1: Q Scores for SNPs Given Various Coverages

	Reference Bases Called	Variant Base Calls	Expected Miscalls (at 1% error rate*)	Depth of Coverage	P Value	Q Score
No Variant Present	100	0	1	100	1	0
5% Frequency SNP	95	5	1	100	0.004	24
5% Frequency SNP	190	10	2	200	4.6×10^{-5}	43
5% Frequency SNP	475	25	5	500	1.6×10^{-10}	98
5% Frequency SNP	4750	250	50	5,000	0**	100**

P value = chance SNP is false positive

Q score = higher the score, higher the confidence

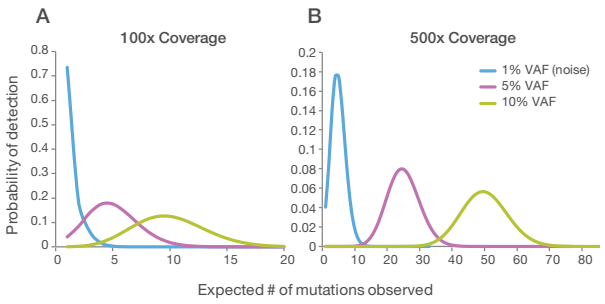
The somatic variant caller computes the Q score for a SNP based on a Poisson model. For example, the 5% frequency SNP in Row 2 (identified at a read depth of 100 and assuming a conservative miscall error rate of 1%) has a Q score of 24 and a P value of 0.004 (i.e., a 4/1,000 chance of being a false positive).

*Conservative error rate used by somatic variant caller to provide high confidence calls.

**Q scores above 100 are not reported.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 1: Impact of Coverage Depth on VAF Overlap



With only 100× coverage (panel A), there is considerable overlap between 5% and 1% VAF, inhibiting the ability to confidently call low-frequency variants below 5%. In contrast, variants below 5% frequency can be reliably called when coverage depth is increased to > 500× coverage (panel B).

Superior Signal Resolution

For simplicity, and to be conservative, the somatic variant caller only considers two alleles or variant states: reference and variant. Any other calls are considered reference calls. For example, A/C/G/T counts of 100/10/1/1 for reference A are considered to be K = 10 variants out of N = 112 coverage. Under the null hypothesis, it is assumed that no variant is present and that any non-reference calls are due to noise. Given a Q20 base filter, the acceptable noise level is 1%. For simplicity, it is assumed that the expected number of non-reference calls due to noise should follow a Poisson distribution with a mean of $\lambda = 0.01*N$. The equation **P = 1 - CDF(K - 1, λ)** represents the probability (P) of having K or more variant calls, where CDF is the cumulative distribution function of the Poisson distribution. P is the probability that no variant is present, given K or more observations. In this way, P is the theoretical false-positive rate, and this probability is converted to a Q score with the maximum Q score set to 100. As seen in Figure 1, greater depth of coverage yields greater discrimination of signal from noise.

High Reproducibility

The somatic variant caller delivers a high level of reproducibility in detecting variants within cancer samples. In a study comparing MiSeq versus Sanger sequencing, the somatic variant caller results for KRAS and BRAF mutations were within 5% of that reported by Sanger sequencing (Table 2). These results highlight the reproducibility and reliability of the TSACP and MiSeq system for detection of somatic mutations.

Somatic Variant Caller Data Analysis

The new somatic variant reporter tool (Figure 2) is built into the MiSeq Reporter v1.3+ software, and is available in BaseSpace. Its output is a .VCF file that contains the identified SNPs and associated Q scores.

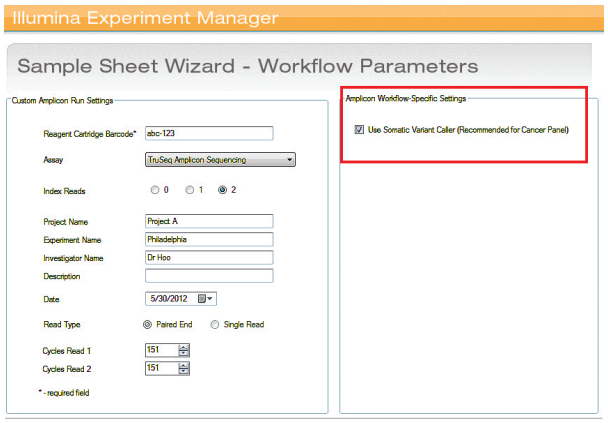
Researchers who have legacy cancer samples analyzed with a prior version of MiSeq Reporter may re-queue their legacy data for analysis using the new somatic variant caller by simply re-editing the sample sheet to specify somatic variant calling. To download the latest version of MiSeq Reporter, please visit www.illumina.com.

Table 2: Reproducibility of % Mutation

Sample	1	2	3
Mutated Gene	KRAS	KRAS	BRAF
Expected % Mutation*	60%	40%	60%
% Mutation, Run 1 (Coverage)	57% (1767×)	37% (1948×)	60% (3902×)
% Mutation, Run 2 (Coverage)	55% (664×)	37% (656×)	59% (2616×)

*Expected mutation rate estimated from Sanger sequencing chromatogram.

Figure 2: Choosing Somatic Variant Caller Data Analysis



To perform somatic variant caller data analysis, go to the Sample Sheet Wizard in the MiSeq Reporter's Illumina Experiment Manager and select it under Amplicon Workflow-Specific Settings (red box).

Summary

The somatic variant caller is a powerful new tool for the analysis of cancer samples and can detect mutations below 5% frequency with high-quality sequencing from the MiSeq system and the TruSeq Amplicon - Cancer Panel.

References

1. www.illumina.com/Documents/products/datasheets/datasheet_truseq_amplicon_cancer_panel.pdf
2. www.illumina.com/Documents/products/other/review_cancer_research.pdf
3. www.illumina.com/documents/products/technotes/technote_eland_variantcalling_improvements.pdf

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

FOR RESEARCH USE ONLY

© 2012-2014 Illumina, Inc. All rights reserved.
Illumina, BaseSpace, MiSeq, TruSeq, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 970-2012-014 Current as of 08 May 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA