Amplicon - DS Somatic Variant Caller

Designed specifically for TruSight Tumor 26 to detect somatic variants below 5% frequency in 26 genes.

Introduction

The highly multiplexed TruSight Tumor 26 panel produces amplicon-based targeted sequencing libraries for the MiSeq® benchtop sequencing system. This panel enables detection of somatic mutations across 14 kb of exons (21 kb of total exons and introns) in genes that are commonly mutated across multiple forms of cancer. By targeting both strands of the template DNA with highly specific amplicon designs, this sensitive assay is ideal for detecting somatic mutations in formalin-fixed, paraffin-embedded (FFPE) samples with the MiSeq system.

Illumina has developed a variant calling algorithm to detect mutations at frequencies below 5% in a mixed cell population with TruSight Tumor 26. The Amplicon - DS (double-stranded) somatic variant caller, compatible with MiSeq Reporter software v2.2, or later, provides a high-confidence method for researchers to evaluate detected mutations. This technical note provides further information about the variant calling algorithm and how to use it for data analysis.

Somatic Variant Calling in Cancer Samples

Deep sequencing of multiplexed amplicon libraries generated from FFPE cancer tissue samples is sensitive, reproducible, and accurate using TruSight Tumor content and the MiSeq system. Generated data can capture heterogeneity that may be present in tumor samples, including normal cells, cells at an early stage of cancer progression, and late-stage cancer cells. Somatic variant calling is inherently difficult because somatic mutations often appear at low frequencies and may be seen in only a small fraction of the reads covering a given base.

Methodology and Usage

The Amplicon - DS somatic variant caller is designed for the analysis of data from multiple cancer samples sequenced in parallel, using results obtained from both strands of the template DNA. The goal of the Amplicon - DS somatic variant caller is to identify variants with high confidence to minimize false positives that can result from less stringent analysis methods. For each call, the variant caller identifies the reference and variant sequences (eg, SNVs, insertions, deletions) and provides a quality (Q) score. This Q-score indicates the degree of confidence that the variant is present in the sample. Variants are reported in variant call format (VCF), a standard tab-delimited format for storing variant calls, which can then be viewed with various tools.

Precision in FFPE Samples

Due to the high level of resolution enabled by deep coverage, next-generation sequencing can detect DNA damage, particularly DNA deamination, caused by formalin fixation. Deamination events effectively result in a C → T single nucleotide change on a single strand of DNA, which appears as a G → A variant when sequenced. If the undamaged complementary strand is sequenced, however, a variant will not be detected. The changes caused by these events can be confirmed as false positive variants by sequencing both strands of the DNA independently and determining whether the variant is present in both strands. This method offers higher precision than alternative methods because true variants are confirmed on each strand independently, which is already common practice in Sanger-based sequencing tests. The Amplicon - DS somatic variant caller analyzes both strands, comparing data from paired, complementary amplicon pools (referred to as FPA and FPB) processed in parallel. Only variants present in both pools are reported as passing, whereas

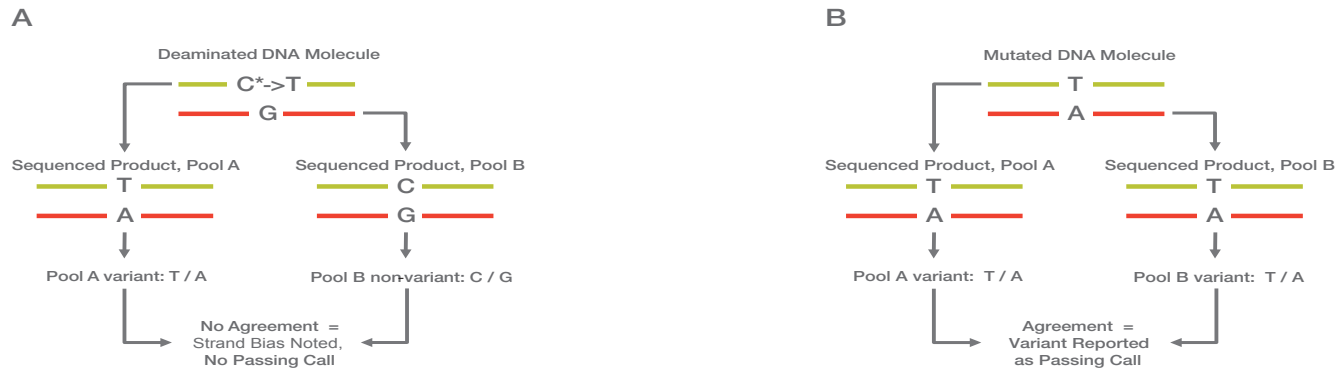

Figure 1: Detection and Differentiation of DNA Damage from Mutation—In panel A (left), cytosine deamination results in a nucleotide change in one strand of a DNA molecule, but does not alter the complementary nucleotide on the opposite strand. Sequencing each strand independently will yield base calls that differ between the 2 strands. In panel B (right), a true DNA mutation results in a nucleotide change in both strands of a DNA molecule. Sequencing each strand independently will yield the same variant call for both strands. The Amplicon - DS somatic variant caller filters false positives and reports confirmed variants that were identified using TruSight Tumor 26.

Table 1: Q-scores for Variant Calls Given Various Coverages

	Reference Bases Called	Variant Base Calls	Expected Miscalls (at 1% error rate[†])	Depth of Coverage	P Value[b]	Q-Score[c]
No Variant Present	100	0	1	100	1	0
5% Frequency SNV	285	15	3	300	6.7×10^{-7}	62
5% Frequency SNV	475	25	5	500	1.6×10^{-10}	98
5% Frequency SNV	950	50	10	1000	0	100[§]
5% Frequency SNV	4750	250	50	5000	0	100[§]
5% Frequency SNV	9500	500	100	10,000	0	100[§]

a. [†]Conservative error rate used by somatic variant caller to provide high confidence calls.
b. The P value indicates the probability that the SNP is a false positive.
c. The Q-score indicates the degree of confidence.
d. [§]Q-scores above 100 are not reported.
The Amplicon - DS somatic variant caller computes the Q-score for an SNV based on a Poisson model. For example, the 5% frequency SNV in the second row identified at a read depth of 300. Assuming a conservative miscall error rate of 1%, this call has a Q-score of 62 and a P value of 6.7×10^{-7}.

variants called in only one pool are flagged and filtered as probe pool biased (Figure 1).

Superior Signal Resolution

The Amplicon - DS somatic variant caller identifies SNPs and short insertions and deletions (indels) present at low frequency in a DNA sample while minimizing false positives. To identify a variant, the Amplicon - DS somatic variant caller considers each position in the reference genome separately, counts the overlapping aligned reads at a given position, and computes a variant score that measures the quality of the call. Variant scores are computed using a Poisson model that excludes calls with scores below Q20 (> 99% predicted accuracy). Also, the algorithm only calls variants for bases that are covered at a depth of 300× or greater for a single amplicon (Table 1). Table 1 shows the likelihood of a false variant being called from a single strand given various depths of coverage. Variants are first called separately for each pool and are then compared and combined into a single output file. If a variant is present in both pools, yields at least 1,000× cumulative coverage between the 2 strands, and has a frequency of 3% or greater, it is marked as passing.

For simplicity, and to be conservative, the Amplicon - DS somatic variant caller only considers 2 alleles: 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 nonreference 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 nonreference 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.

The Amplicon - DS somatic variant caller addresses indels similarly. It analyzes aligned reads covering a given position that includes a particular indel to determine the variant count versus the overall coverage at that position. Due to the amplicon-based generation of the sequencing libraries, the variant caller does not perform some of the indel realignment steps included in other variant callers, such as Genome Analysis Toolkit (GATK). Therefore, small indels (up to 25 bases in length) are reliably detected with this aligner. Illumina recommends that users review and confirm indels using visualization software such as the Integrative Genomics Viewer (IGV)[1,2]. All identified variants are provided in a VCF file, along with the frequencies at which they are detected in either the single or combined pool data.

High Reproducibility

The Amplicon - DS somatic variant caller delivers a high level of reproducibility in detecting low-frequency mutations associated with cancer. In a survey of reference standards with quantified mutations in *BRAF*, *EGFR*, and *KRAS* genes, the Amplicon - DS somatic variant caller results were within ~1% of those reported by the manufacturer (Table 2). These results highlight the reproducibility and reliability of TruSight Tumor 26 and the MiSeq system when detecting somatic mutations.

Data Analysis

The Amplicon - DS somatic variant caller is compatible with MiSeq Reporter software v2.2, or later. When used with the Amplicon - DS workflow, the algorithm exports a VCF file that contains the identified variants and their associated Q-scores and a genome VCF (gVCF) file. It is important to note that variant calls will only be made for those regions that meet the minimum coverage requirements mentioned previously. Illumina recommends that users review the Amplicon Coverage report generated by MiSeq Reporter software to identify any regions where coverage may be below the minimum requirement. In addition to assay variation, large deletions (greater than 25 bases) or complex substitutions can lead to a loss of coverage in the regions where they occur. This loss of coverage, unlike that caused by assay variation, will be evident in the data from both the FPA and FPB assay pools for such variants. These variants will not be captured in the VCF file and will therefore require manual annotation.

Q-scores calculated with the Amplicon - DS somatic variant caller are not equivalent to those generated using the standard TruSeq® Custom Amplicon analysis workflow. As the statistical model and details of

Table 2: Results from Known Reference Standards

Sample	HD103	HD111	HD132	HD123	HD118	HD159	HD129	HD126
Gene	BRAF	EGFR	EGFR	BRAF	KRAS	KRAS	EGFR	EGFR
Mutation	V600E	ΔE746-A750	L861Q	V600K	G12R	G12C	L858R	T790M
Reported Frequency	3.5%	4.0%	4.5%	5.0%	5.0%	6.0%	6.0%	20.0%
Average Frequency	3.3%	4.6%	4.7%	4.3%	5.2%	6.2%	6.6%	18.8%
Average Coverage	16,470	7469	6132	9731	8747	9381	5347	6403

All samples listed here are commercially available through Horizon Discovery Ltd.[3]

variant calling are not compatible, the Q-scores generated by the 2 analysis tools are not directly comparable.

Summary

The Amplicon - DS somatic variant caller is a powerful tool for analyzing cancer samples. This tool can detect variants at frequencies below 5% in high-quality sequencing data generated by the MiSeq system and TruSight Tumor 26. For more information, review the MiSeq Reporter Amplicon - DS Workflow Reference Guide[4] or visit www.illumina.com/trusight-tumor.

References

1. www.broadinstitute.org/igv
2. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nat Biotechnol 29: 24–26.
3. www.horizondiscovery.com
4. MiSeq Reporter Amplicon - DS Workflow Reference Guide, support.illumina.com/documents/documentation/software_documentation/miseqreporter/miseq-reporter-amplicon-ds-workflow-guide-15042903-b.pdf

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

illumina®