illumina®

Evaluating Somatic Variant Calling in Tumor/Normal Studies

The Illumina tumor/normal data analysis workflow enables identification of true somatic variants even in low purity samples.

Introduction

Identifying somatic variants within a tumor sample can be accomplished by performing whole-genome sequencing (WGS) of DNA extracted from a cancer sample and DNA from a matching normal sample. Cancer-specific variants are those observed in the tumor sample but absent from the normal sample. The ability of this approach to detect somatic variants depends, in part, upon the variant frequency within the tumor sample as well as the total sequence depth. The purity and heterogeneity of the tumor sample strongly impact the allele frequencies of the somatic mutations present and thus also impact the ability to detect these variants.

This white paper quantitates the ability to detect somatic variants in samples of variable purity at various sequencing depths using a HiSeq® System and the Illumina tumor/normal data analysis workflow.

Assessment Methodology

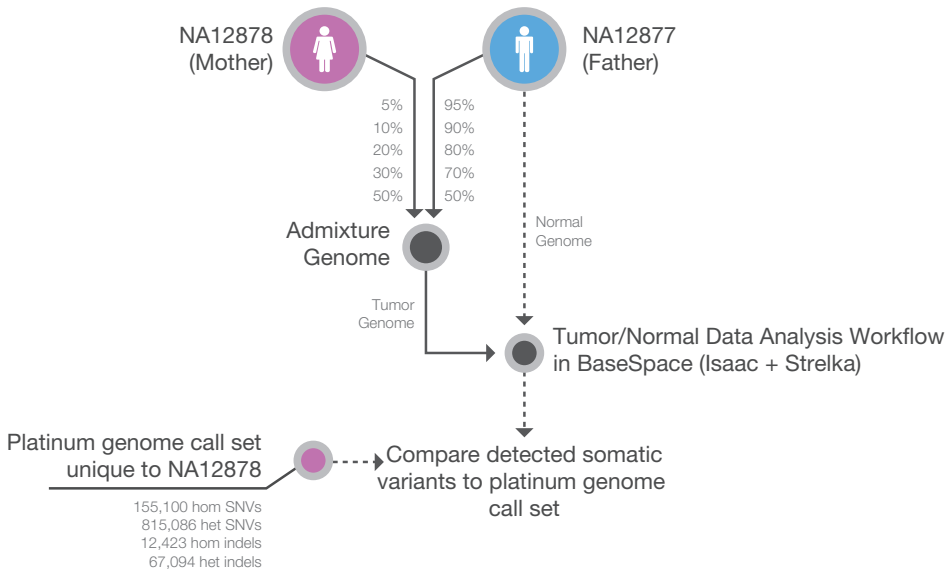
The assessment of the Illumina tumor/normal data analysis workflow is based upon analysis of a highly accurate catalog of variants produced in the "platinum genomes" project[1]. Illumina scientists generated platinum genome data sets by sequencing Coriell Institute for Medical Research samples of a 13-member CEPH pedigree at high depth on a HiSeq System. The high-confidence germ line variants present in the parental genomes were cataloged using hereditary patterns.

Platinum regions representing ~95% of the human genome were characterized as either true homozygous reference or true germ line variants.

A catalog of simulated somatic variants was created by merging the sequencing data of both the maternal and paternal genomes from the platinum genome data set in different proportions to approximate different levels of tumor purity. The true variants within these samples were known through the pedigree anlaysis, enabling *in silico* estimation of sensitivity and specificity. By varying the mixing ratios, the frequency of the variants within the artificial "tumor" sample was adjusted, thus simulating different levels of tumor purity. The mixed genome data sets were used as simulated tumor genomes in the sample and analyzed using the Illumina tumor/normal data analysis workflow, with the paternal genome acting as the surrogate normal (Figure 1).

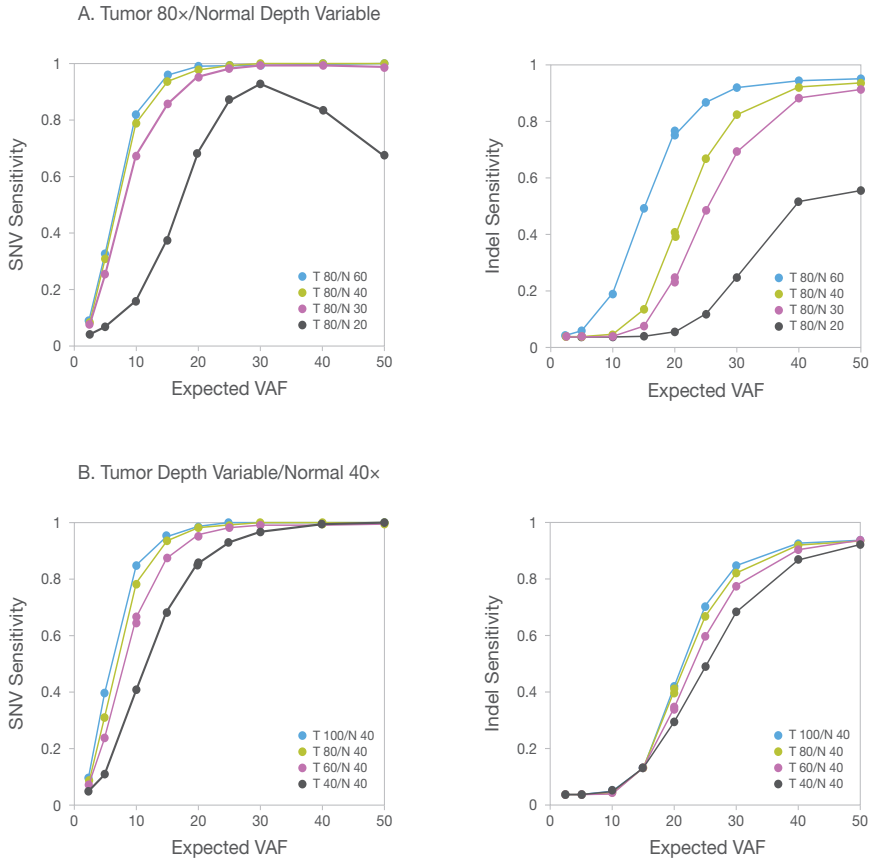
WGS data from tumor and normal samples sequenced on a HiSeq System enter the tumor/normal data analysis workflow in BaseSpace® as FASTQ files. Each FASTQ file contains base calls and quality scores from the tumor and normal samples. Through seamless integration in BaseSpace, the Isaac Alignment Software[2] aligns the files to a reference, followed by somatic small-variant calling using Strelka[3]. To derive sensitivity values, the identified somatic variants were then compared to the high-confidence platinum genome call set unique to

Figure 1: Generation of Simulated Cancer Samples from Platinum Genome Data Sets



*BaseSpace is the Illumina genomics computing environment for next-generation sequencing data analysis

Figure 2: Somatic Calling Sensitivity Using the Illumina Tumor/Normal Data Analysis Workflow



the maternal genome (NA12878) (Figure 1). Specificity was assessed by counting the number of variants detected when subtracting independent replicates of the paternal genome.

Sensitivity Evaluation

Sensitivity of recalling somatic variants was assessed over a range of sample purity values and sequencing depths. The expected variant allele frequency (VAF) acts as a surrogate for purity (Figure 2). For example, if a heterozygous somatic variant is present in all tumor cells and the tumor cells represent 40% of the sample, then the observed VAF would be 20%. Sensitivity is defined as the ratio of known variants called by the tumor/normal data analysis workflow out of a total of 970,186 single nucleotide variants (SNVs) and 79,517 indels. Figure 2A illustrates the sensitivity of somatic variant calling for a fixed tumor depth (80×) with a variable normal depth of sequencing (20–60×). While a depth of 20× in the normal genome reduces the ability to detect SNVs even at high purity, a depth of 30× in the normal genome may still be slightly suboptimal, especially for low-purity samples (VAF < 20%, ~ purity < 40%).

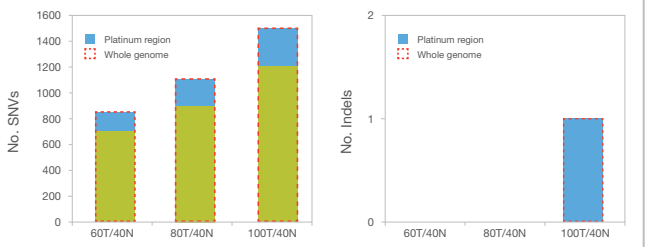
Figure 2B illustrates the sensitivity for a fixed normal depth (40×) and a variable tumor depth (40–100×). A depth of 60× in the tumor genome will perform well for higher purity samples (VAF > 20%, equivalent purity = 40%). Higher sequencing depth may be advisable to maintain high sensitivity in samples with lower purity or higher heterogeneity levels.

Indel calling performance increases significantly with higher normal depth and tumor depth. A minimum depth of 40× in the normal sample is required in order to detect up to 80% of the true somatic variants with VAF 30% or more (equivalent purity = 60%). Increasing the tumor depth will improve sensitivity for purity levels of 40–80% (VAF ≈ 20–40%).

Specificity Evaluation

To measure specificity in this study, two replicates of the same genome were subtracted from each other, and the total number of somatic variants detected over the whole genome and over the platinum regions was determined (Figure 3). As expected, false positives increased with increasing depth of tumor sequencing, but remained low (~1,500 variants/2.6 billion at 100× depth). Most false positives are located in genomic regions known to be difficult to align.

Figure 3: Low Number of False Positives Using Illumina Sequencing



Summary

The Illumina tumor/normal data analysis workflow, consisting of Isaac alignment followed by Strelka small somatic variant calling, demonstrates excellent performance for the detection of somatic variants at high allele frequencies. Sensitivity values exceeded 99.9% for SNVs and exceeded 90% for indels at a sequencing depth of 60x (tumor) and 30–40× (normal) in high-purity (> 80%) samples. Lower frequency variants are expected in tumor samples of lower purity, greater heterogeneity, and within regions containing copy number changes. Higher depth of sequencing will increase sensitivity to detect these variants.

References

1. Illumina Platinum Genomes (2013) www.illumina.com/platinumgenomes
2. 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 29: 2041–2043.
3. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, et.al. (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics 28: 1811–1817.

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

FOR RESEARCH USE ONLY

Pub. No. 1170-2014-001 Current as of 21 March 2014

illumina®