Somatic Variant Discovery in Cancer

Interrogate somatic events and identify driver mutations in tumorigenesis and progression using industry-leading DNA and RNA sequencing tools from Illumina.

**Highlights**

- **Comprehensive Workflow Solutions**
  Sample-to-answer identification of driver mutations by WGS, WES, and RNA sequencing in 1–3 days

- **Highest Read Quality and Alignment**
  Greatest number of high-quality reads for the most accurate somatic variant detection

- **Robust Combined Somatic Variant Calling**
  Best workflow for comprehensive analysis of heterogeneous tumor samples

- **End-to-End Scientific Support**
  Illumina scientists provide installation, application training, and data analysis support at every step

**Introduction**

Next-generation sequencing (NGS)-based discovery and detection of somatic variants in tumor samples is a powerful approach to identify the molecular drivers associated with different cancer types. Discoveries made with this technology can lead to accurate prognoses and effective therapy development, and will be essential to achieving the goals of precision medicine. Using NGS, projects including The Cancer Genome Atlas (TCGA)\(^1\) and International Cancer Genome Consortium (ICGC)\(^2\) have contributed greatly to the catalog of known and suspected mutations that underlie tumorigenesis and progression.\(^3\) These discoveries provide basic research and drug development organizations insight into where to focus their efforts to have the greatest impact in the fight against cancer.\(^4\) In the clinical research and oncology communities, similar sequencing strategies have emerged as powerful weapons to understand the role of somatic variation in the diverse clinical courses of various cancer types. In a handful of examples, predictions in response to therapy would have improved and treatment plans would have changed in a personalized manner based on an individual’s somatic profile.\(^5\) Using a powerful combination of library preparation, sequencing, and informatics methods, Illumina provides researchers with the tools to perform the same comprehensive cancer analysis as groups such as TCGA and ICGC.

DNA sequencing from both the tumor and normal sample along with tumor sample RNA sequencing (RNA-Seq) provides an optimal data set for somatic variant discovery and detection (Figure 1). Sequencing both tumor and normal DNA allows for common germline variants to be efficiently removed from the data set during data analysis. This allows a reduction in the number of candidate driver mutations identified within the tumor sample.

**Library Preparation Tools for NGS Analysis of Somatic Variants**

**A Comprehensive View with Whole-Genome Sequencing**

Through tumor-normal whole-genome sequencing (WGS), researchers can comprehensively compare tumor and normal samples to identify somatic mutations in both coding and noncoding regions of the genome. This hypothesis-free approach highlights the genetic changes specific to the tumor sample relative to a matched normal sample in a genome-wide manner. Illumina TruSeq\(^\text{®}\) DNA PCR-Free Library Prep Kits offer superior coverage of areas traditionally difficult to sequence. Covering these areas such as promoters, and high GC-rich and repetitive regions enable the most complete coverage of the entire genome, detecting the greatest number of possible driver mutations. These kits offer a shortened gel-free workflow with 1 μg DNA input for a streamlined protocol in less than 5 hours (Figure 2). When DNA sample is limited, TruSeq Nano DNA Library Prep Kits provide efficient interrogation of samples down to 100 ng DNA input.
### A Targeted Approach With Whole-Exome Sequencing

Many cancer-causing variants have been found in exonic (protein-coding) regions. Whole-exome sequencing (WES) enables researchers to target cost-effectively 1–2% of the genome that codes for proteins while increasing sensitivity (with high read depth) for somatic variant detection in tumor and normal samples. For a given budget, compared to WGS, the higher read depth of WES enables the detection of low allele fraction variants for accurate identification of somatic mutations within heterogeneous tumor samples. The Nextera® Rapid Capture Exome kits are all-in-one methods for library preparation, enabling rapid identification of coding variants up to 70% faster than other methods (Figure 2). With low DNA input of 50 ng per sample, these kits maintain high coverage uniformity, enrichment rates, and on-target specificity, providing comprehensive coverage of the exome.

### Complete Transcriptome Analysis With RNA Sequencing

RNA sequencing enables researchers to understand the functional effects of DNA mutations through the highly sensitive detection of changes in transcript abundance, discovery of gene fusions, and novel transcript isomers. Further, a broad suite of available library preparation kits addresses the full range of experimental needs, including robust, cost-efficient preparation of low-quality (FFPE) samples and limited available input amounts. Illumina TruSeq RNA Library Prep Kits enable cost-effective, accurate, and complete sequencing of the transcriptome with high-coverage and uniformity. The simple and flexible workflows support quick and robust interrogation of standard (0.1–4 µg), low-input (20 ng), and FFPE samples (Figure 2). These kits provide a comprehensive view of the transcriptome to elicit the functional effects of somatic variants associated with tumorigenesis and progression.

For more details on Illumina library preparation kits for optimal somatic variant detection, see Table 1.

---

**Table 1: Guidelines for Ideal Sequencing Target Depths**

| Tumor-Normal WES subtraction analysis to identify somatic variation | BaseSpace Core Apps.
|---------------------------------------------------------------|------------------|
| Identification of exonic (protein-coding) regions. | BaseSpace Core Apps.
| Whole-exome sequencing (WES) enables researchers to target | BaseSpace Core Apps.
| cost-effectively 1–2% of the genome that codes for | BaseSpace Core Apps.
| proteins while increasing sensitivity (with high read depth) | BaseSpace Core Apps.
| for somatic variant detection in tumor and normal samples. | BaseSpace Core Apps.
| For a given budget, compared to WGS, the higher read depth | BaseSpace Core Apps.
| of WES enables the detection of low allele fraction variants | BaseSpace Core Apps.
| for accurate identification of somatic mutations within | BaseSpace Core Apps.
| heterogeneous tumor samples. | BaseSpace Core Apps.
| The Nextera® Rapid Capture Exome kits are all-in-one methods | BaseSpace Core Apps.
| for library preparation, enabling rapid identification of | BaseSpace Core Apps.
| coding variants up to 70% faster than other methods (Figure 2). | BaseSpace Core Apps.
| With low DNA input of 50 ng per sample, these kits maintain | BaseSpace Core Apps.
| high coverage uniformity, enrichment rates, and on-target | BaseSpace Core Apps.
| specificity, providing comprehensive coverage of the | BaseSpace Core Apps.
| exome. | BaseSpace Core Apps.
| RNA sequencing enables researchers to understand the functional | BaseSpace Core Apps.
| effects of DNA mutations through the highly sensitive detection | BaseSpace Core Apps.
| of changes in transcript abundance, discovery of gene fusions, | BaseSpace Core Apps.
| and novel transcript isomers. Further, a broad suite of available | BaseSpace Core Apps.
| library preparation kits addresses the full range of experimental | BaseSpace Core Apps.
| needs, including robust, cost-efficient preparation of low-quality | BaseSpace Core Apps.
| (FFPE) samples and limited available input amounts. Illumina TruSeq | BaseSpace Core Apps.
| RNA Library Prep Kits enable cost-effective, accurate, and complete | BaseSpace Core Apps.
| sequencing of the transcriptome with high-coverage and uniformity. | BaseSpace Core Apps.
| The simple and flexible workflows support quick and robust | BaseSpace Core Apps.
| interrogation of standard (0.1–4 µg), low-input (20 ng), and FFPE | BaseSpace Core Apps.
| samples (Figure 2). These kits provide a comprehensive view of | BaseSpace Core Apps.
| the transcriptome to elicit the functional effects of somatic variants | BaseSpace Core Apps.
| associated with tumorigenesis and progression. | BaseSpace Core Apps.

---

**Table 2: Workflow Solution for Somatic Variant Detection**

<table>
<thead>
<tr>
<th>Extraction</th>
<th>Library Preparation</th>
<th>Sequencing</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA Extraction Normal Sample</td>
<td>Whole-Genome Sequencing Tumor-Normal Paired Samples</td>
<td>Raw Data Generation Using HiSeq or NextSeq Systems</td>
<td>Alignment, Variant Calling, Tumor-Normal Combined Calling Analysis for WGS, Fusion Gene ID Interpretation</td>
</tr>
<tr>
<td>RNA Extraction Tumor Sample</td>
<td>Whole-Exome Sequencing Tumor-Normal Paired Samples</td>
<td>1-2 days</td>
<td>1-2 days</td>
</tr>
<tr>
<td>RNA-Seq Tumor Samples</td>
<td>or</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2: Workflow Solution for Somatic Variant Detection**

-Tumor-Normal WES subtraction analysis to identify somatic variation requires third-party software (ie, VarScan, www.varscan.sourceforge.net) as this functionality is not currently supported by BaseSpace Core Apps.
Fast and Accurate Bioinformatics Tools

Tumor-Normal for WGS Analysis

Tumor biopsies are often highly heterogeneous and are likely to contain multiple subclones of cancer cells, as well as normal cell contamination. The Illumina Bayesian combined-calling method for somatic variant detection is designed to address this intrinsic heterogeneity.

Sequencing reads from both normal and tumor samples are mapped to the reference sequence using Illumina Isaac Aligner software. The somatic variant caller combines tumor and normal sample data sets together (combined calling) to model varying levels of tumor purities present with superior accuracy and sensitivity compared to other subtractive analysis methods. The combined analysis takes into account the somatic variation and noise that can occur at any allele frequency ratio, and reports somatic SNVs and small indels. Similarly, structural variants (SVs) are obtained by comparing the SVs identified in the tumor and matched normal samples. The combined-calling method from Illumina identifies somatic and true variants with superior sensitivity > 97.0% and specificity > 98.5% for tumor purity as low as 40% (Figure 5).

RNA-Seq Analysis

New Illumina RNA-Seq Apps in BaseSpace (Figure 4) provide expert-preferred data analysis tools (TopHat and Cufflinks) packaged in an intuitive, click-and-go user interface designed for informatics novices. These apps deliver preconfigured workflows that support a range of common transcriptome data analysis needs, including high-confidence alignment for abundance measurement and detection of splice junctions, gene fusions, and cSNPs. Simple-to-follow prompts guide users through the entire process, starting with file selection generated by the sequencer to filtering and visualizing analyzed data and results. RNA-Seq Apps software generates output files that can be directly imported into a broad range of available downstream analysis tools.
Somatic Variant Discovery and Detection in the Clinic

Much progress has been made in the discovery of driver mutations in cancer, but recent reports suggest that many thousands of variants remain undiscovered. Therefore, some groups are working to combine their basic research discovery efforts with clinical research by associating detected somatic variants with clinical phenotype data. Still others are collecting their NGS data sets in a CLIA-certified setting for both research purposes and clinical decision making. While whole-genome and whole-exome approaches provide the most opportunity for discovery and detection, many groups have opted to focus only on a handful of confirmed driver mutations with established clinical significance. They have chosen to bypass discovery efforts in favor of a focus on clinical implementation. While this highly focused approach has some benefits, the emerging genetic complexity of cancer suggests significant advantages in investing in a translational data set that is complete. Not only will the new data set be available for research, but also for reanalysis as discoveries are made and the spectrum of available targeted therapies increases.

Summary

Cancer is a diverse and deadly disease of the genome. Due to the molecular complexity, cancer has historically been a challenge to identify driver mutations and understand the biology of tumorigenesis and progression. However, by combining DNA sequencing from both the tumor sample and normal tissue with tumor RNA sequencing, researchers can obtain an optimal data set for somatic variant discovery. Furthermore, rapid and comprehensive library preparation and enrichment kits, industry-leading sequencing accuracy, and streamlined data analysis workflows from Illumina provide efficient and precise interrogation of the molecular complexity of cancer. These workflows now enable all researchers to identify a complete picture of somatic variation in tumor samples, which until now, have been accessible only to international consortiums such as TCGA and ICGC.

Table 1: Library Preparation Kits for Somatic Variant Data Sets

<table>
<thead>
<tr>
<th>Application</th>
<th>Whole-Genome Sequencing</th>
<th>Whole-Exome Sequencing</th>
<th>RNA Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TruSeq DNA PCR-Free</td>
<td>TruSeq Nano DNA</td>
<td></td>
</tr>
<tr>
<td>Sample Input DNA/RNA</td>
<td>1–2</td>
<td>0.1–0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>(μg/sample)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Prep Time</td>
<td>5 hours</td>
<td>6 hours</td>
<td></td>
</tr>
<tr>
<td>Hands-On Prep Time</td>
<td>4 hours</td>
<td>5 hours</td>
<td></td>
</tr>
<tr>
<td>Kit Configurations</td>
<td>24, 96</td>
<td>24, 96</td>
<td></td>
</tr>
<tr>
<td>Pre-Enrichment Sample</td>
<td>N/A</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Pooling</td>
<td>Up to 12</td>
<td>Up to 12</td>
<td>N/A</td>
</tr>
<tr>
<td>Recommended Read Length</td>
<td>2 × 100 bp</td>
<td>2 × 100 bp</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Recommended Coverage Levels or Number of Reads for Accurate Somatic Variant Detection

<table>
<thead>
<tr>
<th>Application</th>
<th>Normal Sample</th>
<th>Tumor Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole-Genome Sequencing</td>
<td>40×</td>
<td>80×</td>
</tr>
<tr>
<td>Whole-Exome Sequencing</td>
<td>100×</td>
<td>130×</td>
</tr>
<tr>
<td>Total RNA Sequencing</td>
<td>N/A</td>
<td>&gt; 50M reads</td>
</tr>
<tr>
<td>mRNA Sequencing</td>
<td>N/A</td>
<td>25M reads</td>
</tr>
</tbody>
</table>

References