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.

DNA sequencing is a powerful method for the identification of single nucleotide variants (SNVs), small insertion and deletion events (indels), and large-scale copy number variants (CNVs). RNA sequencing is ideal for identifying genes that are differentially expressed in tumor tissues and for detecting the presence of gene fusions. As such, sequencing RNA from tumor samples enables full transcriptome analysis. In combination with WGS or WES, RNA-Seq allows researchers to prioritize candidate somatic driver mutations identified through DNA sequencing.

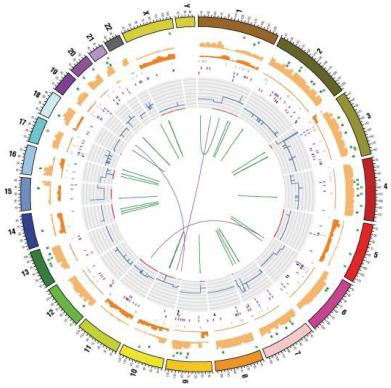

Figure 1: Somatic Variant Detection Using WGS or WES With RNA-Seq—An optimal data set for somatic variant detection using a combined approach of WGS or WES with RNA-Seq can lead to insights into the mechanism of tumorigenesis. This example Circos plot shows point mutations, copy number aberrations, and chromosomal rearrangements that can be found by using a combination of DNA and RNA sequencing approaches.[6]

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® 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.

For Research Use Only. Not for use in diagnostic procedures.

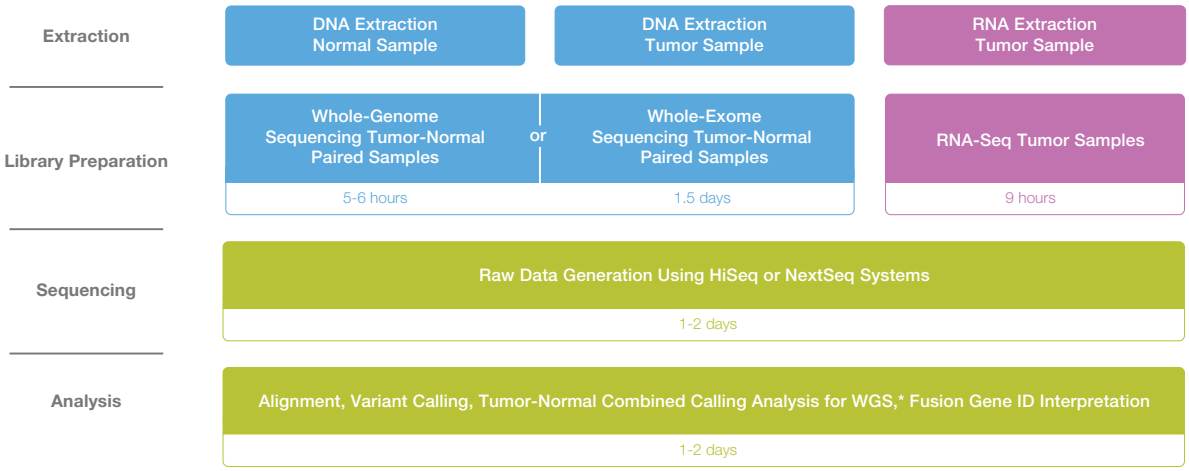

Figure 2: Workflow Solution for Somatic Variant Detection—*Tumor-Normal WES subtraction analysis to identify somatic variation requires third-party software (ie, VarScan, www.varscan.sourceforge.net) as this functionality is not currently supported by BaseSpace Core Apps.

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.

High-Accuracy Somatic Variant Sequencing

Because of tumor heterogeneity and normal cell contamination, sequencing of a tumor sample often identifies a mix of DNA signatures representing the constituent cell types of that particular tumor sample. Thus, detection of true somatic mutations requires deep sequencing to detect driver mutations at low allelic fractions within the sample. Deep sequencing refers to sequencing a genome region multiple times with overlapping short reads to obtain high confidence that identified mutations within that region are accurately identified. Figure 3 shows the impact of read depth on the ability to call somatic variants of various allelic fractions. The required read depth depends on several factors including the purity (ie, normal cell contamination) and heterogeneity (ie, clonal architecture) of the tumor sample and desired experimental sensitivity. Thus, any increase in raw sequencing error will limit the ability to detect true variants within these heterogeneous samples. This is important for somatic variant detection as some somatic driver events are present at low allelic level within a given sample. Illumina NGS technology not only delivers the highest-quality raw data on a genome-wide level, but allows access to more of the genome, successfully sequencing difficult regions such as GC-rich and homopolymers. As a result, significantly more variants can be robustly and accurately identified.

Table 2 provides guidelines for ideal sequencing target depths for optimal somatic variant detection, providing sufficient confidence to identify mutations within tumor samples while limiting false positive and negative rates. Leveraging Illumina technology, the HiSeq® and NextSeq® Systems deliver industry-leading accuracy (> 80% sequenced bases over Q30) and high coverage capability, addressing the challenges and increasing sensitivity in detecting somatic variants.

For Research Use Only. Not for use in diagnostic procedures.

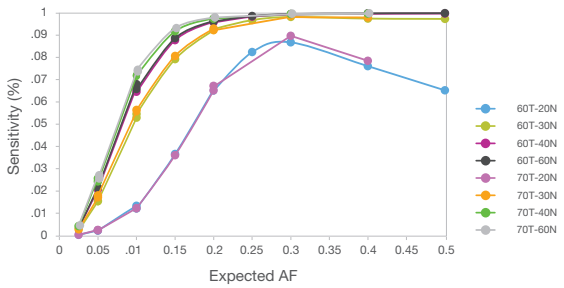

Figure 3: Impact of Read Depth on Sensitivity for Somatic Variant Detection—With increased read depth (coverage), the sensitivity (ability to call variants) for detecting somatic variants at low allelic levels increases for tumor-normal data sets.

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 Alignment[7] 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.[8] 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).

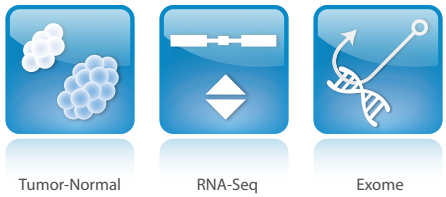

Figure 4: BaseSpace Apps Empower Somatic Variant Discovery—A suite of BaseSpace Apps is available to detect somatic variants and perform transcriptome analysis in tumors. Somatic variant detection from whole-genome tumor-normal sequence data can be achieved using the Tumor-Normal App. The RNA-Seq apps can be used to analyze gene expression, transcripts, and fusion genes, while the exome apps detect mutations. **Note:** Currently, tumor-normal subtraction is not enabled in BaseSpace for exome data. For somatic variant detection using subtractive analysis of tumor-normal samples after whole-exome sequencing, Illumina recommends use of a third-party software such as VarScan from Washington University (www.varscan.sourceforge.net).

The tumor-normal WGS workflow described here is available as part of a suite of cancer data analysis apps in the Illumina BaseSpace® computing environment (Figure 4). BaseSpace is available to users with access to an Illumina sequencing system or a third-party service provider. Raw data from Illumina sequencing instruments can be uploaded to and analyzed directly in BaseSpace. Analysis results for all apps, including annotated variants and performance statistics, are presented in intuitive, easy-to-interpret reports.

RNA-Seq Analysis

New Illumina RNA-Seq Apps in BaseSpace (Figure 4) provide expert-preferred data analysis tools (TopHat[9] and Cufflinks[10]) 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.

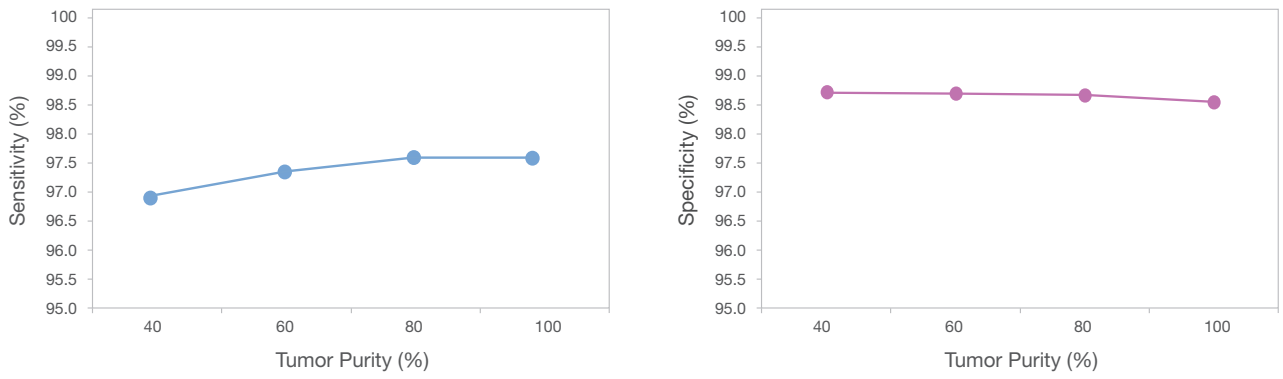

Figure 5: Combined-Calling Method Enhances the Ability to Call True Somatic Variants—A melanoma cell line was analyzed using a combined-calling method. Tumor sample data was mixed with varying fractions of a normal sample data ranging from 40%–100% to simulate different tumor purities. Results indicate high sensitivity (ability to call variants) and specificity (ability to call true variants) even at low tumor purity.

For Research Use Only. Not for use in diagnostic procedures.

AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AATCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACC
AACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTA
TTAACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGAT
AAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAA
AAGATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGA
AACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAG

Table 1: Library Preparation Kits for Somatic Variant Data Sets

Application	Whole-Genome Sequencing		Whole-Exome Sequencing		RNA Sequencing	
Recommended Kit	**TruSeq DNA PCR-Free**	**TruSeq Nano DNA**	**Nextera Rapid Capture Exome**	**Nextera Rapid Capture Expanded Exome**	**TruSeq Stranded Total RNA**	**TruSeq Stranded mRNA**
Sample Input DNA/RNA (µg/sample)	1–2	0.1–0.2	0.5	0.5	0.1	0.1
Total Prep Time	5 hours	6 hours	30 hours for up to 96 samples	30 hours for up to 96 samples	8 hours	9 hours
Hands-On Prep Time	4 hours	5 hours	< 5 hours	< 5 hours	5.5 hours	5.5 hours
Kit Configurations (samples/kit)	24, 96	24, 96	24, 48, 96	24, 48, 96	48, 96	48, 96
Pre-Enrichment Sample Pooling	N/A	N/A	Up to 12	Up to 12	N/A	N/A
Recommended Read Length	2 × 100 bp	2 × 100 bp	2 × 75 bp – 2 × 100 bp	2 × 75 bp – 2 × 100 bp	2 × 75 bp	2 × 75 bp

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.[11] 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.[12] 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.[13]

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

Application	Normal Sample	Tumor Sample
Whole-Genome Sequencing	40×	80×
Whole-Exome Sequencing	100×	130×
Total RNA Sequencing	N/A	> 50M reads
mRNA Sequencing	N/A	25M reads

References

1. Home - The Cancer Genome Atlas - Cancer Genome - TCGA (cancergenome.nih.gov) Accessed 27 May 2015.
2. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. *Nature.* 2010;464:993-998.
3. COSMIC: Catalogue of Somatic Mutations in Cancer (cancer.sanger.ac.uk/cosmic) Accessed 27 May 2015.
4. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. *Science.* 2013;339:1546-1558.
5. Schuh A, Becq J, Humphray S, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood.* 2012;120:4191-4196.
6. COSMIC: Catalogue of Somatic Mutations in Cancer (www.sanger.ac.uk/genetics/CGP/cosmic/help/for_sample.shtml) Accessed 27 May 2015.
7. Illumina (2013) Isaac Genome Alignment and Isaac Variant Caller (www.illumina.com/documents/products/whitepapers/whitepaper_iassc_workflow.pdf)
8. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics.* 2012;28:1811-1817.
9. TopHat (ccb.jhu.edu/software/tophat/index.shtml) Accessed 27 May 2015.
10. Cufflinks (cole-trapnell-lab.github.io/cufflinks) Accessed 27 May 2015.
11. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499:214-218.
12. Jardim DL, Conley A, Subbiah V. Comprehensive characterization of malignant phyllodes tumor by whole genomic and proteomic analysis: biological implications for targeted therapy opportunities. *Orphanet J Rare Dis.* 2013;8:112.
13. Targeted Cancer Therapies Fact Sheet - National Cancer Institute (www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet) Accessed 27 May 2015.

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

For Research Use Only. Not for use in diagnostic procedures.

© 2015 Illumina, Inc. All rights reserved. Illumina, BaseSpace, HiSeq, Nextera, NextSeq, TruSeq, and the pumpkin orange color are trademarks of Illumina, Inc. and/or its affiliate(s) in the U.S. and/or other countries. Pub. No. 1170-2013-005 Current as of 04 June 2015

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA