# BIOTEK TOOLS IN A BOX

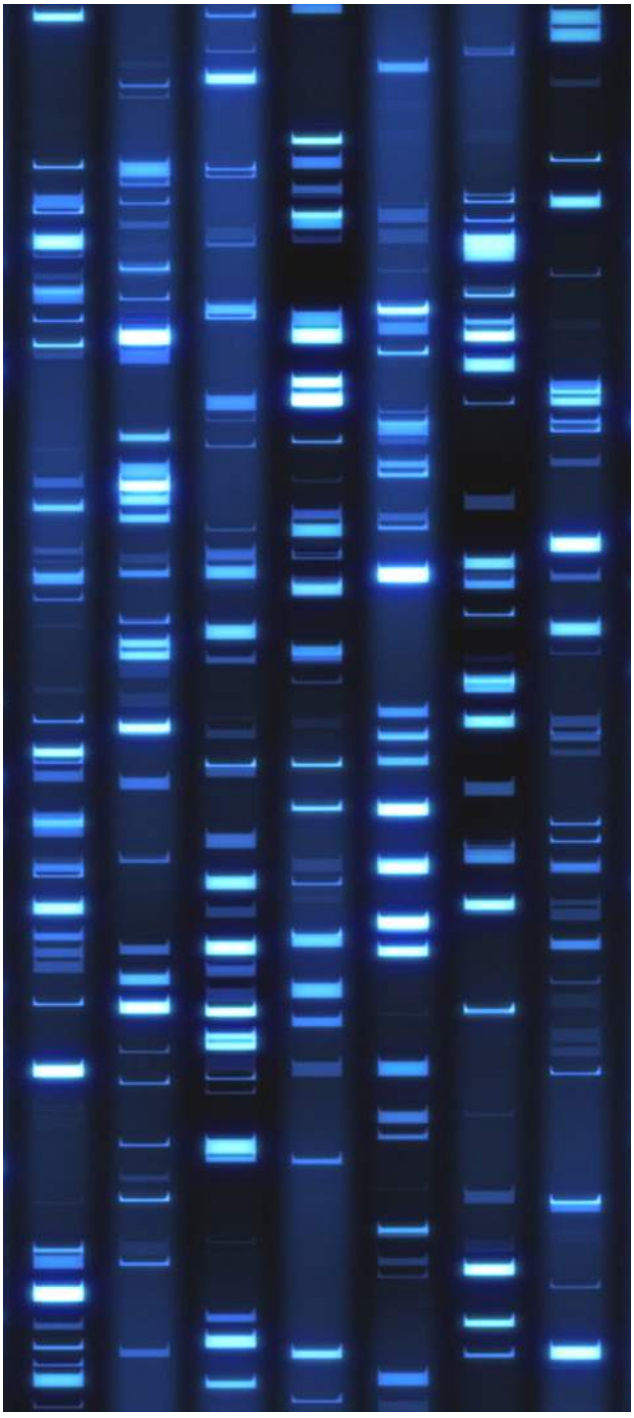
FOR BIOINFORMATICS RESEARCHERS

**BUILDING BIOLOGICAL DATA REPOSITORY, VARIANT ANALYSIS AND PROTEIN STRUCTURE PREDICTION SOFTWARE PACKAGE**

SciTech
consulting and solutions

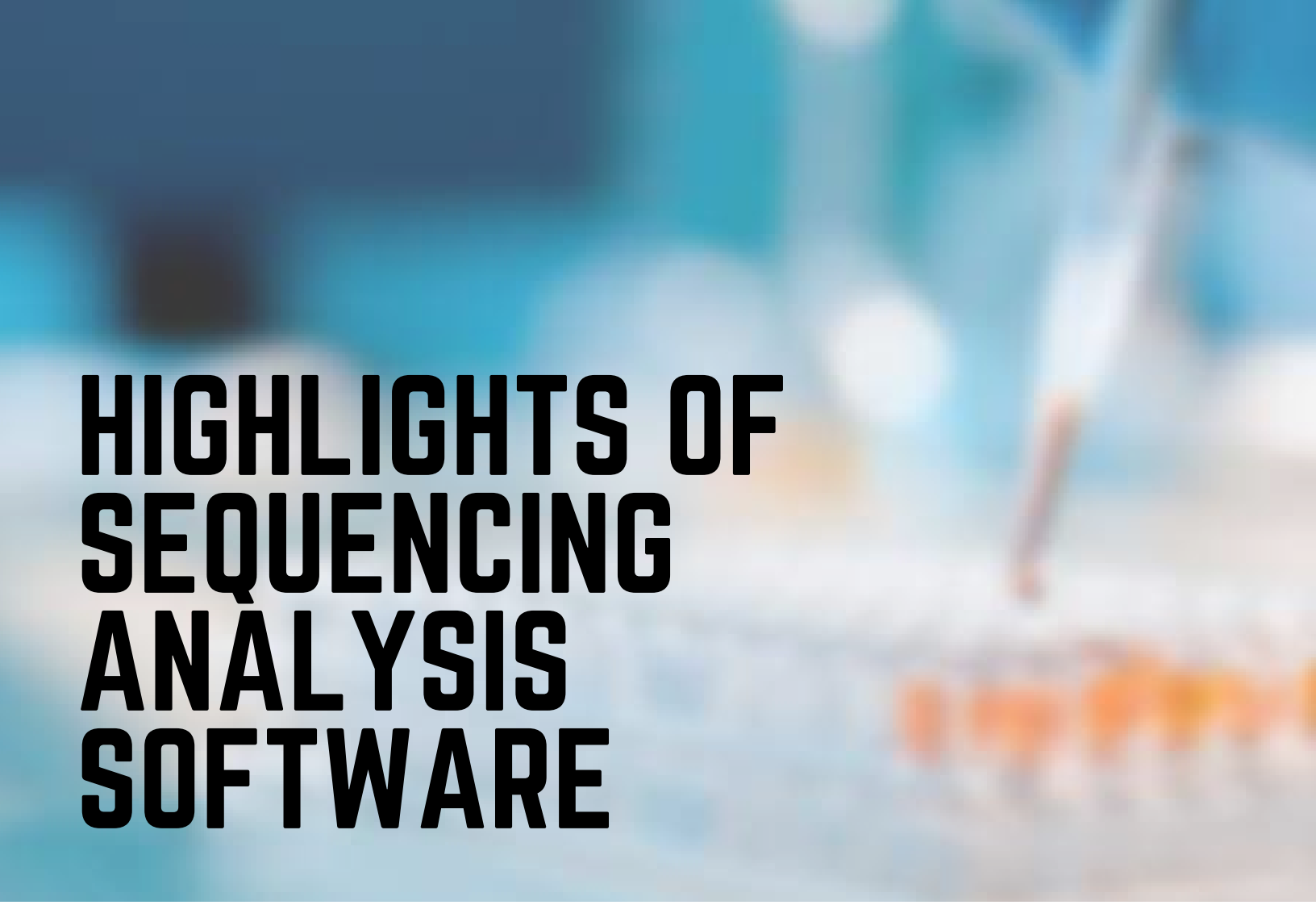
**VIEW FULL STORY**

Each experiment on NGS (Next Generation Sequencing) Platform generates huge amount of data. Further analysis in turn also generates similar multitudes of processed and interpreted data at different steps. Terse types of data are useful for the immediate research in question and are also extremely useful for future researchers in the context of conducting validation, reanalysis and meta-analysis studies. ScITech has developed a software portfolio that consists variant analysis, protein structure prediction and building of biological data repository and provides visually representation of relevant data. The graphical interface is intuitive for use by non-informaticians.



# INTRODUCTION

Comparative analysis of whole genome sequence data is becoming an increasingly important and accessible approach for addressing both fundamental and applied biological questions. BioTek integrates bioinformatics tools and databases for variant analysis and protein structure predictions. The portfolio offers tools and algorithms for annotating the effects of genome variation  and to identify the 3D structures of novel proteins and predicting their functions.  The tools are developed using Perl, C++ and MySQL on Ubuntu Linux version. Currently, the software package contains an accompanying script for automated installation of necessary external programs on Red Hat Linux; however, the pipeline is compatible with other Linux and Unix systems after necessary external libraries are installed.



# HIGHLIGHTS OF SEQUENCING ANALYSIS SOFTWARE

## SPECIFICATIONS

### VARIANT ANALYSIS

Annotating the effects of genome variations is often a key step in the identification of causes of biological phenomena.

Phylogenetic Tree Building

This method is often used when the common evolutionary events are considered to have a different probability of occurrence. Under the probability distribution which is given as input, these methods try to construct a tree with the minimum number of evolutionary events to explain the observed variation.

Hi C Visualization Tools

Hi-C, 4C, and other chromatin capture technologies will allow for the detection of long-range interactions in the genome. After processing the data, the user will be able to produce plots to visualize and interpret the data. This module will also provide the following type of visualization functionality.

- Virtual 4C plot
- 2D heat map
- Intra-chromosomal heat map

## TYPES

Variant analysis Software
Annotating the effects of genome variations.

Building biological data repository Software
Extremely useful for future researchers in the context of conducting validation, reanalysis and meta-analysis studies as well as for immediate research.

Protein structure prediction Software
Identifying the 3D structure of novel proteins from hypothetical protein sequences and predicting their functions.

SciTech
Consulting and Solutions

- Inter-chromosomal heat map
- Circular plot
- Rotated local heat map
- Local arc track
- Multi-dataset visualizations
- Hi-C signal transformations
- Locus-specific circular plot
- Continuous-valued tracks

The module provides the ability to zoom in and out of inter and intrachromosomal. The tool also allows for Hi-C heatmaps to be explored interactively. The user can zoom down to a high level of detail, letting the user explore interesting areas at a suitable resolution. The virtual 4C plots are supported alongside the heat map visualization.

### Hi C Processing Tools

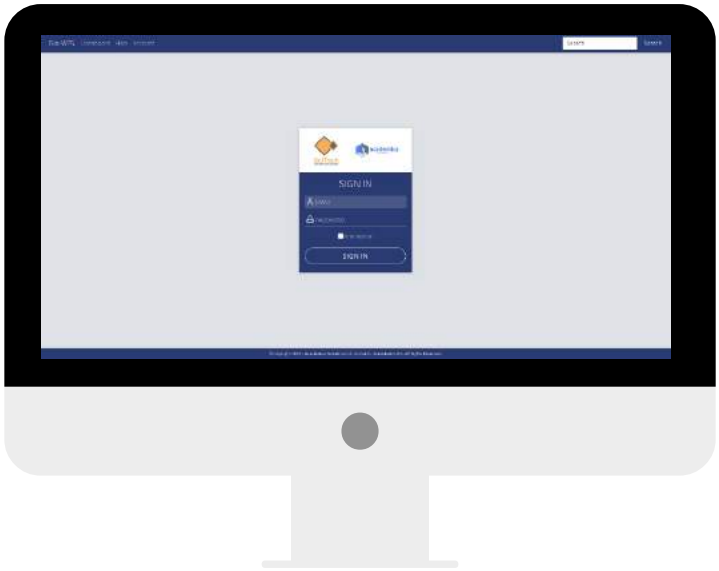
This module provides an end to end HI-C processing toolkit able of processing raw reads without any quality control step, perform Hi-C specific quality control, align the reads to a reference and produce normalized contact maps.

### Structural Variant Caller

Structural variants are an important class of genomic rearrangements which are generally defined as having a size greater than 50 bp. This tool will detect common classes of structural variations including deletions, insertions, tandem repeats, transpositions, translocations, inversion, copy number variation and mobile-element transpositions. In addition to the primary functionality, there are several additional helper scripts included in this module to help in further downstream analysis.

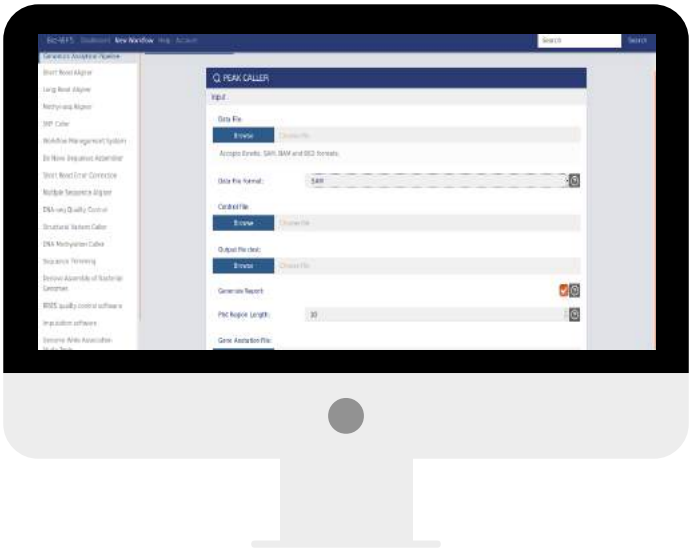
## HARDWARE REQUIREMMENT

- Linux (Red Hat recommended)
- 16 Core processors
- 32 GB RAM
- > 50 GB Free Disk Space




## SOFTWARE HIGHLIGTS

- Comprehensive: End-to-end solution for sequencing data analysis
- Multithreaded: Threads will run on separate processors and provide faster performance.
- Compatible: The modules support all major input types from all major sequencing machines
- Customizable: All major output formats are supported and can be specified by user
- Intuitive: Graphical Interface



# BUILDING BIOLOGICAL DATA REPOSITORY SOFTWARE

Public biological data repositories have become a key to genomic analysis. As such the ability to build, modify and use such repositories is a key skill within bioinformatics.

## Sequence Database Searching

Tools for efficiently searching sequence databases search within a database of DNA sequences using alignment techniques and scoring matrices to determine how similar two sequences are. This module also supports a wide range of options such as searching for nucleotide sequences, protein sequence, whole genomes and more.

There is also a simple but intuitive web-based GUI for numerous sequence searching and sequence processing tools.

## Sequence Data Manipulation Tools

This module supports a variety of data manipulation tasks merging sorted files, splitting files, conversion between formats, printing files to standard output (including compressed BAM files), generate a new file in pileup format, convert CIGAR notation from binary to human readable form. It will also support the most common data formats produced during next generation sequencing and its related downstream analysis are SAM, BAM, CRAM and files.

## Bed File Manipulation Tools

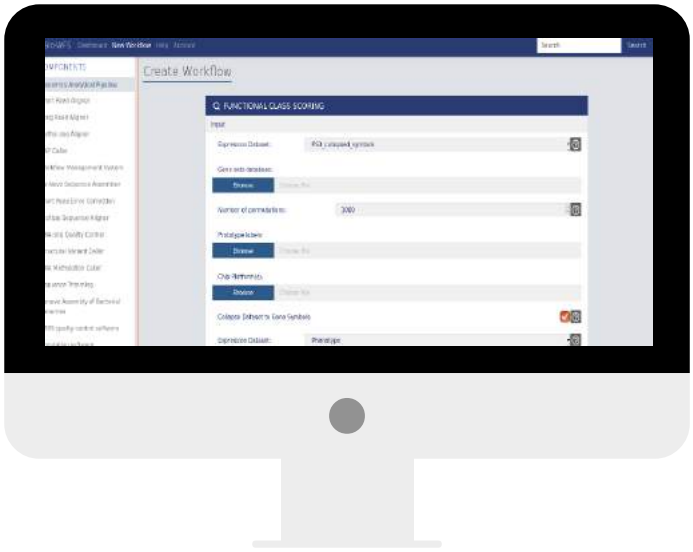
This module provides a range of basic tools for the manipulation of bed files. Many additional common bed file manipulation tasks have also been identified and are provided by this module. Any searching and filtering techniques in this method are implemented using sophisticated string matching technologies to ensure efficient search times.

## VCF File Manipulation Tools

This module allows for the efficient and flexible processing of VCF files. A script is provided that computes a summary of the file information. This information includes allele frequency statistics, transition statistics as well as many other statistical tests and functions.

## Data Normalization Tools

Various different next-generation sequencing tasks require data normalization before comparisons take place. This module allows for various different normalization techniques to be used including quantile normalization and specific methods for RNA-seq and Chip-seq experiments.

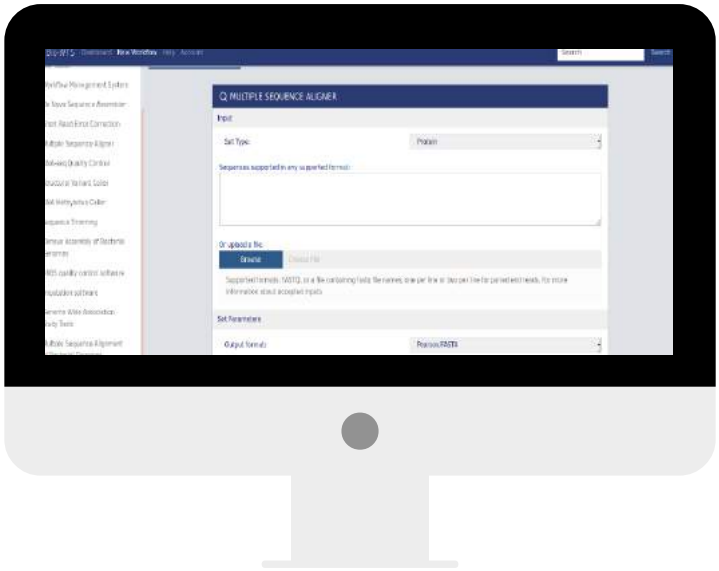


# PROTEIN STRUCTURE PREDICTION SOFTWARE

Identifying the 3D structure of novel proteins from hypothetical protein sequences and predicting their functions.

## Homology Modeling

Homology modeling is the comparative modeling of proteins. In homology modeling, we aim to construct a model of the protein from an amino acid sequence. The module allows for the construction of a three-dimensional structure using a related protein which is used as a template which the new model is based on.




In summary BioTek is a wide-ranging set of tools necessary for a Bioinformatics researcher in a box.

SCITECH CONSULTING AND SOLUTIONS LIMITED

S-3, PARK PLAZA, HOUSE #31, ROAD #17
BANANI, DHAKA 1213, BANGLADESH
INFO@SCITECHBD.COM

SciTech
consulting and solutions