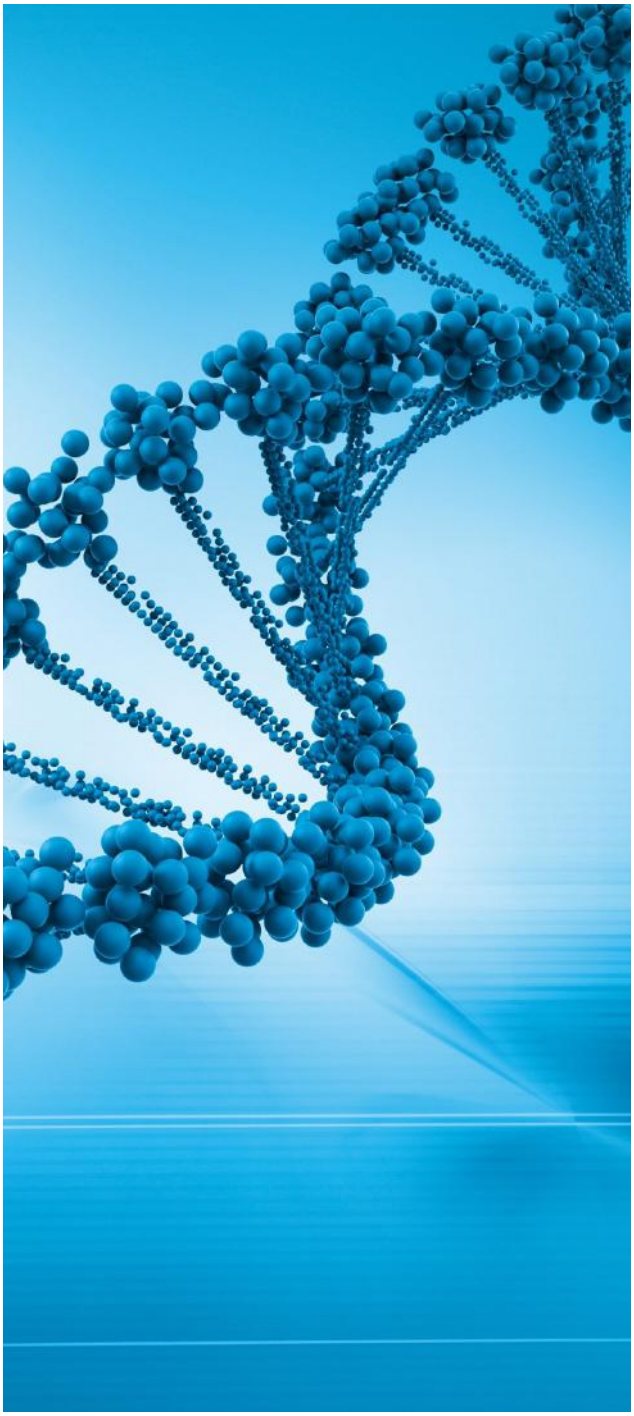
GENETEK

GENOMICS, TRANSCRIPTOMICS ANALYTICAL PIPELINE AND PATHWAY ANALYSIS SOFTWARE PACKAGE

ACADEMICA SOLUTIONS
31, STAPLERS ROAD, NEWPORT, ISLE OF WIGHT
PO30 DB, UK

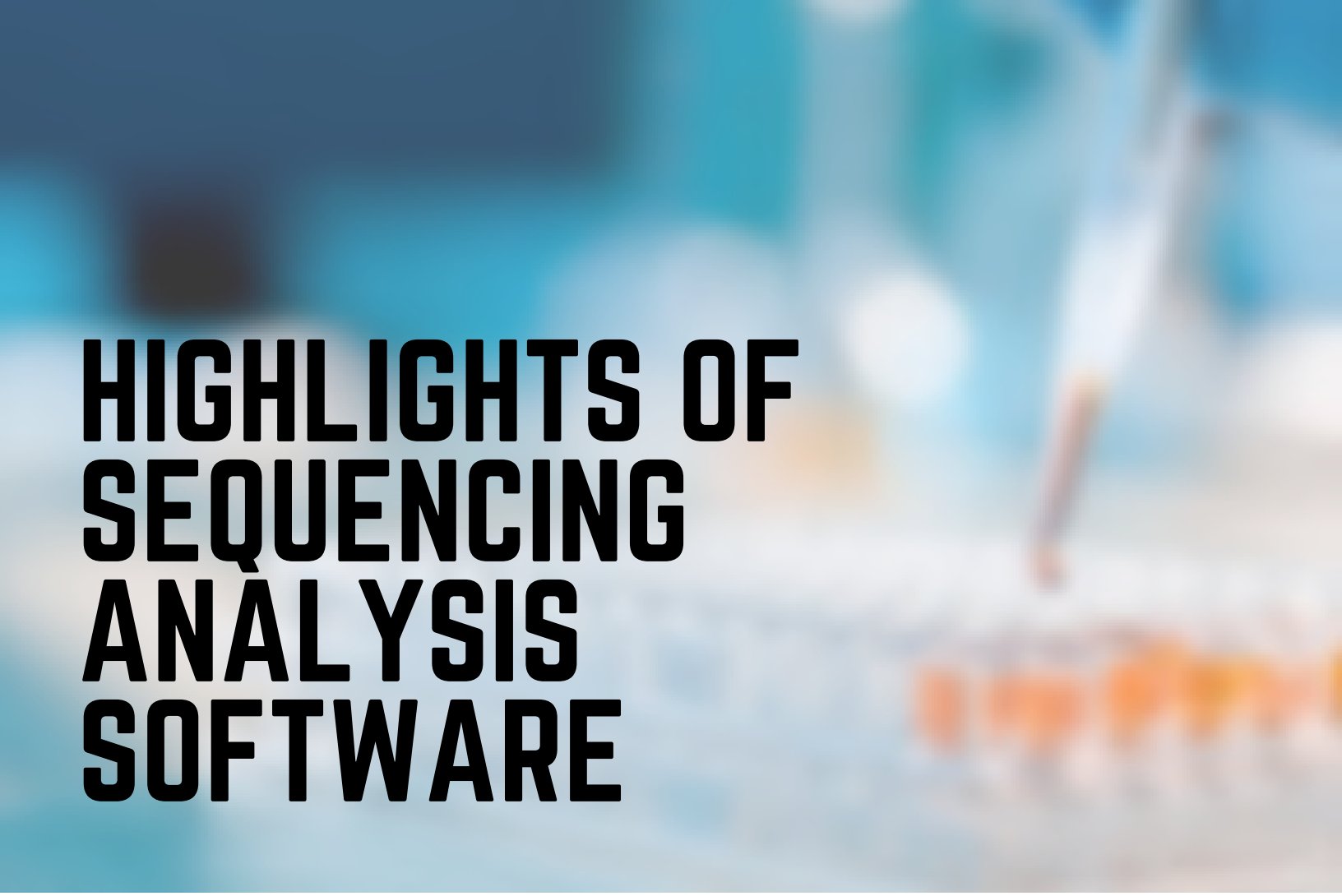
“

A genome research lab remains incomplete without specialized computing infrastructure and appropriate personnel to take advantage of the information generated from it. The vast amount of data generated by next generation sequencing platforms requires careful data reduction and management system as well as bioinformatics tools for downstream analyses. Academica has developed ‘GeneTek’ a robust set of software package to analyze the massive output and provide end-to-end solution of biological relevant data.



INTRODUCTION

Comparative analysis of whole genome sequence data is becoming an increasingly important and accessible approach for addressing both fundamental and applied biological questions. Genetek integrates bioinformatics tools and databases for comparative analysis of a large number of genomes. The pipeline offers tools and algorithms for annotation and analysis of completed and draft genome sequences. pipeline is developed using Perl, C++ and MySQL on Ubuntu Linux version. Currently, the software package accompanies 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

GENOMICS ANALYTICAL PIPELINE

Next-generation sequencing technologies are very high throughput and produce huge datasets. After sequencing this data needs to process in an efficient, unbiased and accurate manner. In this package, we provide tools for the analysis of such data from raw sequencing data to complex downstream analysis.

Short Read Aligner

Short read datasets are the most common data produced by NGS technologies. The alignment of these reads back to a reference genome is an important first step in the analysis pipeline. This module executes the process of figuring out where in the genome a short sequence is from in an efficient manner.

Long Read Aligner

Some next-generation sequencing technologies produce very long fragments of up to 100s of kilobases. Exiting short read technology is unable to correctly process this information. We provide tools to efficiently and accurately map long reads to a reference genome.

TYPES

Genomics analytical Software

- *Pre-processing*
- *Variant identification*
- *Copy number alteration region identification*

Analyses data at DNA level to identify changes in target (e.g., disease) DNA sequence from reference (e.g., healthy) DNA sequence (or between DNA from different groups.

Transcriptomics analytical Software

- Quality Control
- Differential Expression

Analyses data at gene/transcript level to identify changes in gene/transcript expression (i.e., quantity) from one group to another. The change is often represented in terms of over/under expression or up/down regulation of genes

Pathway analysis

Based on the differentially expressed genes between target and reference samples, identifying statistically significant biological pathways (a group of proteins associated with performing a particular task).

academica SOLUTIONS

Methyl-seq Aligner

Methyl-seq technologies use NGS technologies to produce millions of short reads which have methylation specific properties. Here we provide a tool to accurately map methyl-seq data with care taken to provide unbaised estimates of methylations levels across the genome.

SNP Caller

This tool will provide a powerful way to identify novel single nucleotide polymorphisms (SNPs) and call known SNPs in genome or transcriptome samples.

Workflow Management System

This will provide a comprehensive workflow management system to allow the easy construction of end to end analysis workflows. This module will support the following functionalities.

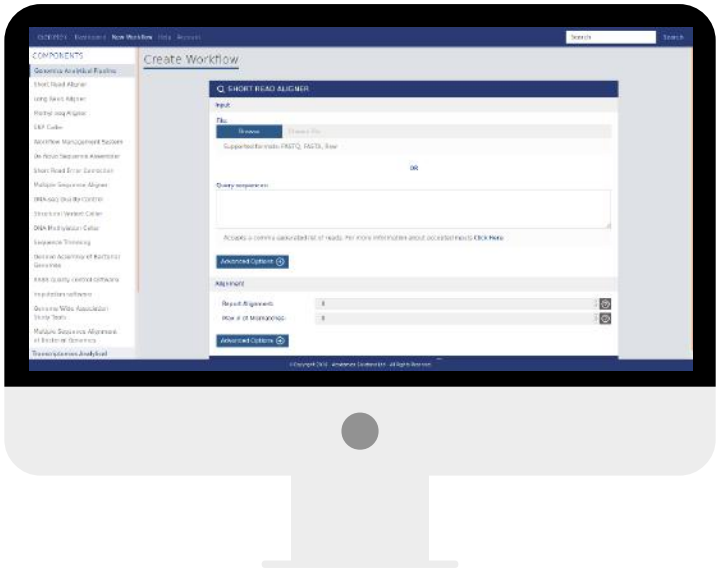
- New Pipeline
- Load Pipeline
- Save Pipeline
- Generate Pipeline

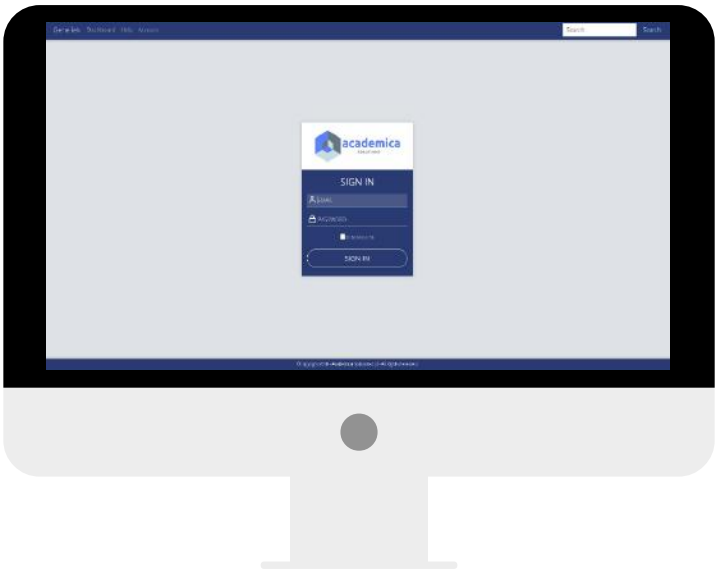
DE Novo Sequence Assembler

This will construct a genomic sequence without a reference. These methods use the reads generated from next generation sequencing technologies and will have support for both long and short reads.

Short read error correction

This module will provide the functionality to correct errors in reads where possible. It will correct errors in reads rather than simply removing them as this provides more data for analysis.





Multiple Sequence Aligner

Multiple sequence alignment considers the alignment of 3 or more sequences and is a significantly harder problem than pairwise sequence alignment. Several efficient heuristic approaches are known that optimize for different cost functions. This module will provide functionality for at least 3 of the most common approaches.

DNA-seq Quality Control

This software package conforms to all the standard QC statistics and quality metrics for NGS data that is computed and displayed sensibly in a QC report.

Structural Variant Caller

This module is able to detect common classes of structural variations including deletions, insertions, tandem repeats, transpositions, trans-locations, inversion, copy number variation and mobile-element transpositions.

DNA Methylation Caller

Once the bi-sulphite converted DNA has been sequenced and aligned this module will be used to call the methylation value for CpGs in BED format to ensure compatibility with other downstream analysis tools and ensure re-usability once the data is released to the public.

Sequence Trimming

This module provides a method that uses fast exact string matching methods to find all of the common adapter sequences and trims reads for all of the common error types that can arise as sequencing artefacts.

Denovo assembly of bacterial genomes

The module is heavily based on the de novo assembler and provides basis de novo functionality in an efficient and tested way.

RRBS Quality Control

This module is based on the implementation of the general sequence quality control software. Simulated genome digestion is provided by a script in this module which uses sophisticated string matching techniques to quickly identify digestion sites and cleave them. This process outputs the genome fragments for later analysis

Imputation Software

This module provides statistical inference of unobserved genotypes in a dataset, a very useful step in GWAS studies allowing for more statistical power when testing for traits of interest. This also provides a number of fine tuning options.

Genome wide association study tools

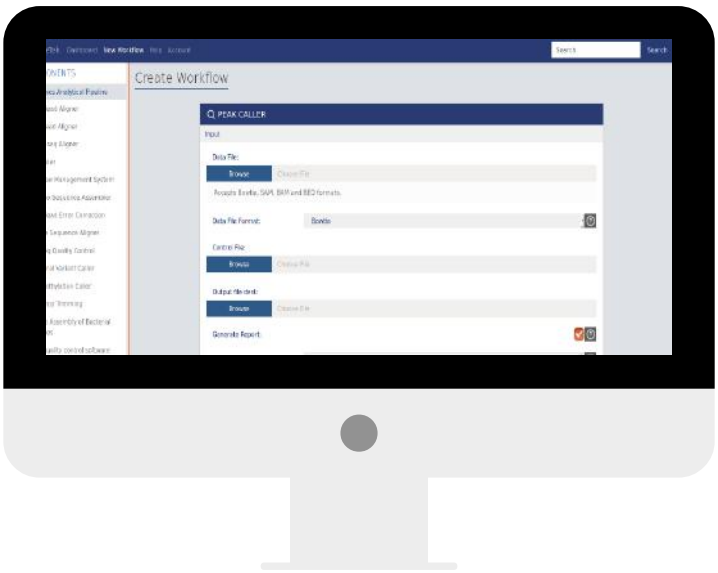
This is an observational study which tests for a significant association between a trait and a genome wide set of variants (often SNPs). The module uses both univariate and multivariate linear mixed model approaches in the analysis of genome wide significance.

Multiple Sequence Alignment of Bacterial Genome

This module provides all of the basic multiple sequence alignment functionalities in an efficient and tested way. This is based on the more general multiple sequence alignment module with extra optimizations for bacterial genomes.

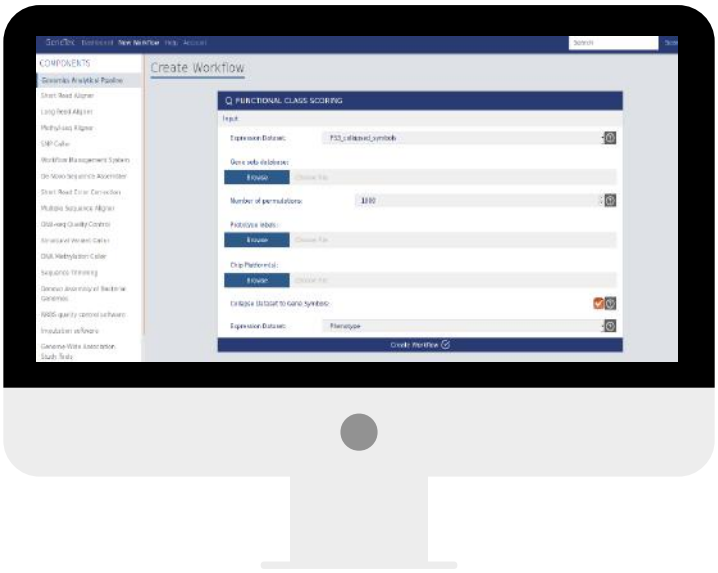
HARDWARE REQUIREMMENT

- 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



TRANSCRIPTOMICS ANALYTICAL PIPELINE

Transcriptome sequencing provides insights into the quantity, structure and composition of RNA molecules in a biological sample.

RNA Sequence Aligner

This module provides Alignment of RNA-seq data. The goal of the alignment process is finding the right position of the input reads in the reference genome. This module provides such functionality in an efficient way and takes into account properties of RNA-seq that would not be present in other data types.

Differential Peak Calling

This module identifies areas in a genome that
have been enriched with aligned reads as a consequence of performing an experiment such as Chip-seq or MeDIP-seq. These methods attempt to identify peaks which are statistically significantly when compared to the background.

Transcriptome Assembler

This module provides a sophisticated algorithm to de novo reconstruct transcriptomes from RNA-seq data. This method provides the fast and robust transcriptomes to be produced. This module partitions RNA-seq data into many smaller datasets and uses a de Bruijn graph method to model transcriptional complexity for some locus.

Peak Caller

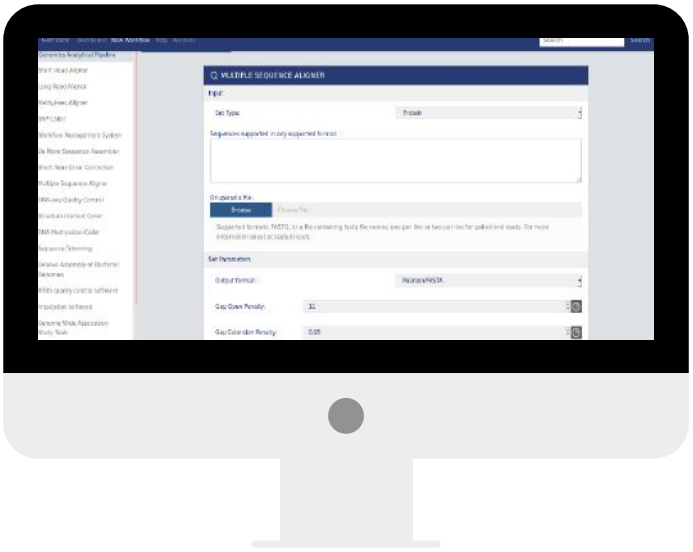
This module allows the user to apply the appropriate statistical tests to be able to determine the peaks which are statistically significantly when compared to the background. This module provides a flexible set of options and different types of statistical tests to allow the user to tailor the test to the experiment.

RNA-seq Quality Control

This module will compute the most common quality control metrics/statistics including

- Sequence quality
- Sequencing depth
- Nucleotide composition bias
- rRNA contamination
- Mitochondria contamination

There are a variety of output options that the user can specify including outputting the data in an easy to read the report for data inspection analysis of gene sets. This tool is able to handle large databases of gene sets in a robust and sensitive way.



academica SOLUTIONS

PATHWAY ANALYSIS

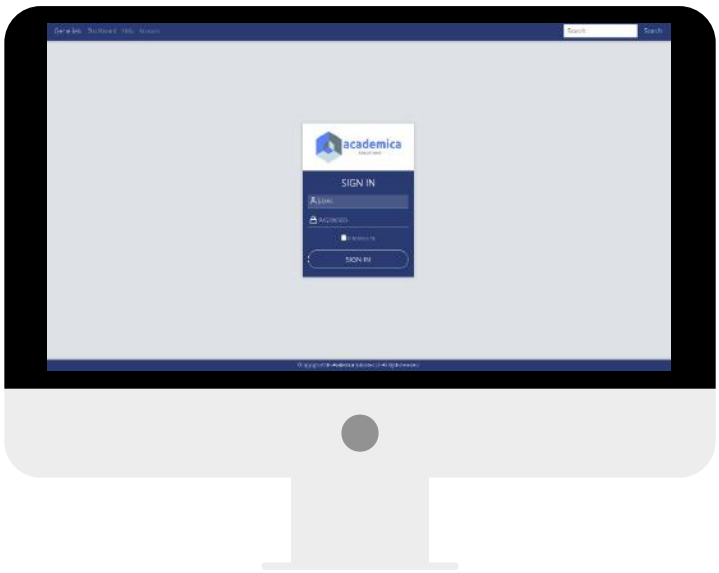
Based on the differentially expressed genes between target and reference samples, identifying statistically significant biological pathways.

Functional Class Scoring

This module provides a robust tool for the analysis of gene sets. This tool is able to handle large databases of gene sets in a robust and sensitive way.

Pathway Topology

This module uses the pathway regulation score technique for the analysis of pathway enrichment for RNA-seq or other gene expression data.



In summary Genetek is a set of necessary tools for a researcher in a box.

ACADEMICA SOLUTIONS
31, STAPLERS ROAD, NEWPORT
ISLE OF WIGHT, PO30 DB, UK
CONTACT@ACADEMICASOLUTIONS.COM