Qlucore Omics Explorer 3.8 NGS Module feature overview

INTRODUCTION

Qlucore Omics Explorer (QOE) supports the user with fast, simple and visual analysis of measured data. The NGS Module is an add-on module that will enable additional functionality related to data generated with NGS technologies and will make it possible to interactively and dynamically analyze and explore NGS data both from DNA and RNA.

All functionality will be provided as integrated parts of Qlucore Omics Explorer and work with the well-known functionality.

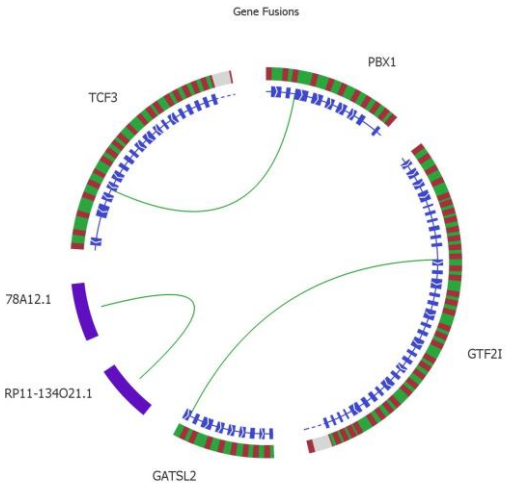
MODULE OVERVIEW

The main components of the NGS module are:

- An interactive and fast genome browser for analysis and visualization. Support for many samples and many tracks per sample
- A flexible genome filter control for focus on relevant parts of data
- A project manager for project set-up and definition of which files to include
- A built-in variant caller for short indels and variants
- A Gene Fusion workbench with circle plot visualizations and possibility to view and navigate the gene fusion sequence in the genome browser.

EXAMPLE OF SUPPORTED ANALYSIS

- RNA-seq
- Peak analysis (ChIP-seq and ATAC-seq)
- Variant analysis
- Fusion gene analysis



INTERACTIVITY

The NGS module is unique in providing true interactive and dynamic analysis of NGS data. The Genome Browser content is dynamically updated when filters and filter cut-off are changed, for instance using sliders and checkboxes. The circle plot is updated as the filter settings change in the Gene Fusion workbench.

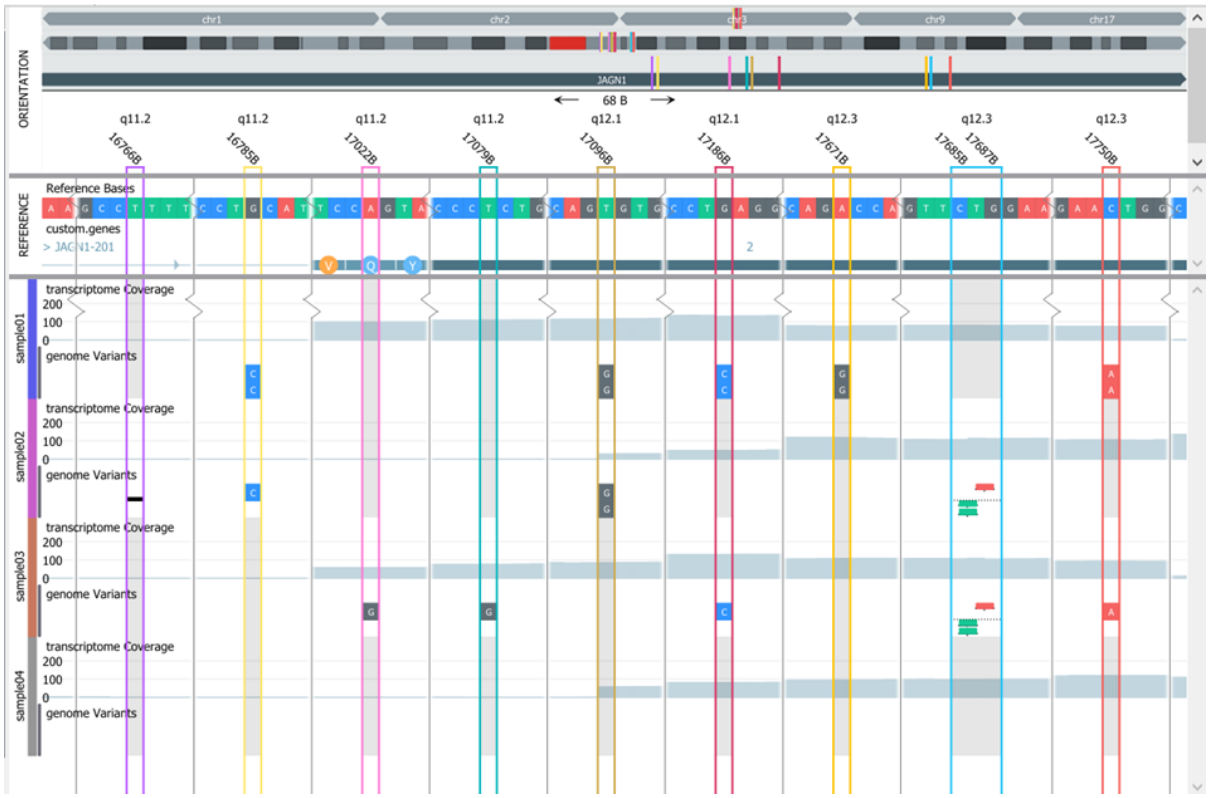
COPYRIGHT 2021 QLUCORE AB

WORKFLOW AND DATA STORAGE

One unique feature of the NGS Module is the two-stage workflow. The first stage is a pre-processing step that prepares data for the fast and interactive analysis in stage two.

The project files and associated data for indexing samples are stored locally on the user's computer to secure fast access and interactivity while files such as large BAM files can be stored anywhere on a network (it is though preferred to have the BAM files locally stored since it makes the Genome Browser faster when showing individual reads).

GENOME BROWSER



The Genome Browser enables visualization in a variety of flexible ways. Results after filtering, zooming and scrolling are fast and interactive.

Visualization options and selection possibilities are extensive with the fast and interactive genome browser. Examples of options include:

- Display one or several samples
- Add and display any number of annotation tracks
- Detailed track with summary of reads
- Labeling, coloring and sorting to aid visualization in the sample track
- Navigation tracks for improved orientation on three levels (whole genome, chromosome and gene)
- Tracks for all accepted data types showing, for example, deletions and insertions, as well as gene fusions
- Tracks with coding regions, genes, introns, exons etc. extracted from GTF file
- Zoom

COPYRIGHT 2021 QLUCORE AB

- Padding to control how much of filtered bases are shown to make sure that edge effects are not affecting the analysis
- Detailed info on selected features in annotations

RNA-SEQ ANALYSIS

The options available for RNA-seq analysis are impressive. Utilizing the existing functionality in QOE base module for expression data and combining it with the NGS functionality enables significantly increased analysis options.

When a RNA-seq project is initiated, both quantitative normalized expression levels (select from FPKM, TPM and TMM) will be calculated at the same time as data is prepared for the genome browser. The program is built to seamlessly handle information both from a quantitative and genomic view. As an example, the user can define discriminating genes with a t-test, visualize them in a heatmap and study the genes in the genome browser. If a sample group is selected or de-selected all plots will be updated and the analysis can directly continue.

PEAK ANALYSIS

The peak analysis support allows for comprehensive analysis of peak data such as ChIP-seq and ATAC-seq. The main components of the peak analysis processing are peak detection, consensus peaks and count matrix generation.

QOE supports de novo peak analysis by detecting peaks using MACS2 peak calling algorithm. The peaks detected in each peak BAM-file can then be merged into a consensus peaks file. This merged file can then be used as a feature file to define the variables of a count matrix. It is also possible to supply another bed file with known peaks of interest for a non-de novo analysis.

A gtf file is used as the feature file to generate a gene centric peak count matrix. Since peaks associated with a gene might be located before or after the actual gene, it is possible to set a padding region before and/or after each gene. Any reads within the padding region will be added to the counts. If the experiment also includes RNA-seq data, it is possible to generate a count matrix for RNA-seq and swap between the RNA-seq and ChIP-seq count matrices during the analysis.

The genome browser includes powerful filtering options for peak analysis, making it easy to find interesting features. The browser allows the user to annotate peaks and export the results as a bed file.

DATA AND FILE TYPES

A reference genome is mandatory. All organisms are supported, provided a reference genome expressed with the .fasta file format is available.The following file formats can be imported and used:

COPYRIGHT 2021 QLUCORE AB

- .bam files of version 1.3 or higher. Version 1.5 or higher is[1] recommended
- .vcf files according to VCF version 4.2
- .bed files[2]
- .gtf files according to specification at http://www.ensembl.org/info/website/upload/gff.html [3]
- .txt file for cytobands. Format according to UCSC
- .txt or .tsv files for fusion genes. Gene fusion output files from the folling callers can currently be analysed:
    - Star Fusion: star-fusion.fusion_predictions.abridged.tsv
    - Fusion Catcher: final-list_candidate-fusion-genes.txt
    - Arriba
    - Manta (TS500, Illumina)

Vcf is a flexible format. A subset of the full VCF file format is supported and the following features are available:

- Handling of small substitutions, insertions and deletions
- Full header parsing
- Support for vcf:s with multiple samples
- Support for all mandatory SNV fields as well as INFO and FORMAT
- Partial support for large mutations and structural variants

VARIANT CALLING

Variant calling is done using Samtools (http://www.htslib.org/) and small variants (insertions, deletions and substitutions) are called by default.

FILTERS

The filter component is flexible and lets the user combine multiple filters to select the relevant parts of data to analyze. The filter can be applied to calculated values such as read coverage as well as presence and matches with certain names. An example is:

---

[1] There is no well-defined standard for a FASTA file. In QOE a FASTA file should have the following structure. A FASTA file contains a list of genomic sequences. Each sequence starts with a header row in this format:
>SequenceID or >SequenceID [Optional Sequence Description]
The header row must start with '>' and be followed immediately (no whitespace in between) by the Sequence ID. The Sequence ID is terminated by whitespace or end-of-line. An optional sequence description can be supplied on the header row – after Sequence ID and the separating white space. The header row is followed by one or more rows of sequence data. Nucleotide sequences are described by A,C,G,T and N characters. Both upper and lower case characters are allowed. The sequence data is terminated by end-of-file, or by a new header line starting with '>'.

[2] The four first mandatory fields: chr, start, end and name are supported. All further fields are ignored. Track and browser lines are ignored.

[3] QOE requires some non-mandatory fields to import files.

COPYRIGHT 2021 QLUCORE AB

Create a combination of filters on parameters such as SNV:s being present in the vcf file and named “A” and not present in the vcf file and named “B” and combine this with a requirement that all displayed bases shall have a read depth of at least 50.

Filtering on any names in GTF files such as gene names, gene ids or transcript ids is supported.

Filtering SNVs on many criteria including INFO or Format values as well and QUAL and FILTER with relevant operators depending on the type of the values is supported.

THE GENE FUSION WORKBENCH AND CIRCLE PLOT

The Workbench has two main tabs, Fusions as detected in the sample(s) and Databases. The Database tab enables inspection of Database content. The program is shipped with the Mitelman and TumorFusions databases. The Fusion tab provides extensive filtering, viewing and sorting options. Examples are:

- Filter on Breakpoint reads
- Filter on Spanning fragments
- Filter on presence in database

Selected Gene fusions are visualized in the Circle plot and in the Genome Browser. In the gene fusion track is the breakpoint, the genetic code and corresponding amino acids shown. The breakpoint sequence is displayed in the NGS tab.

REPORT, EXPORTS AND INFORMATION WINDOW

- High quality 2-D graphics exports of the genome browser content in png, jpg, bmp, TIFF or pdf (vector graphics)
- Report list of SNV’s in the browser and the corresponding information such as BED and Cytoband. HGVS (http://varnomen.hgvs.org) strings are provided as well as reference sequence specifiers missing and protein levels
- Gene fusion break point sequence

SAMPLE MANAGEMENT

- Samples can be organized into groups using sample annotations
- Active/visible samples can be managed using sample annotations
- Import of sample annotation files (*.txt, *.csv)

SYSTEM REQUIREMENTS

The list below highlights the most important requirements. For full information see the Qlucore Omics Explorer System Requirement.

- 64bit Operating System (Windows or Max OS X)
- 12GB of RAM memory
- A graphical card with support of at least Open GL 3.3.
- At least 500GB free hard disk space, preferable a SSD disk
- A fast processor, preferable Core i7 or similar with at least 4 cores

COPYRIGHT 2021 QLUCORE AB

RESEARCH PURPOSE ONLY
Qlucore Omics Explorer is intended for research purposes.

DISCLAIMER
The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore has no liability for any error or damages of any kind resulting from the use of this document.

TRADEMARK LIST
Windows is a trademark of Microsoft.

Core i7 is a trademark of Intel.

Mac and OS X are trademarks of Apple.

COPYRIGHT 2021 QLUCORE AB