geneXplain® platform 6.0 release

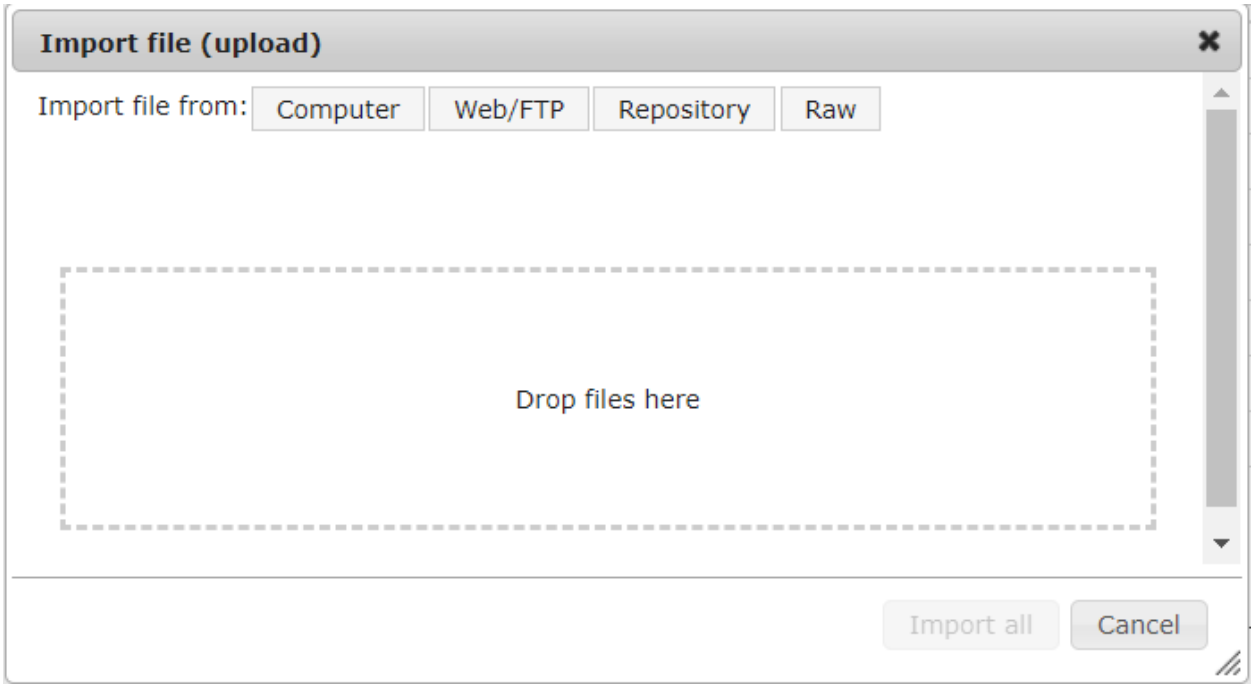
Database updates

- Ensembl is updated to release 99 (January 2020).
- TRANSFAC® is updated to version 2020.2 (May 2020).
- TRANSPATH® is updated to version 2020.2 (May 2020).
- HumanPSD™ is updated to version 2020.2 (May 2020).
- Gene Ontology database is updated to version 2020-03-25.

New features

- **Import with Drag and Drop function**

Using standard import function to upload files to the platform now supports easy Drag and Drop functionality.



EdgeR for two tables

The tool *Empirical Analysis of Digital Gene Expression Data in R* (edgeR) can be applied to any technology that produces read counts for genomic features and estimates differential expression at the gene, exon, transcript, or tag level. In fact, read counts can be summarized by any genomic feature.

This tool uses the edgeR quasi-likelihood pipeline (edgeR-quasi) for differential expression analysis. This statistical methodology uses negative binomial generalized linear models, but with F-tests instead of likelihood ratio tests. This method provides stricter error rate control than other negative binomial based pipelines, including the traditional edgeR pipelines or DESeq2. While the limma pipelines are recommended for large-scale datasets, because of their speed and flexibility, the edgeR-quasi pipeline gives better performance in low-count situations.

This method supports now two separate read counts tables for samples and control samples while identifying differentially expressed genes.

MTB report: From somatic variants to treatment options

Disclaimer: This report is intended for research use only and should not be used for medical or professional advice. geneXplain GmbH makes no guarantee of the comprehensiveness, reliability, or accuracy of the information on this report. You accept full responsibility for all risks associated with using this report.

GENE-DRUG PREDICTIVE ASSOCIATIONS Method: Somatic variants of one patient (mutations, amplifications, deletions, rearrangements) are searched in curated databases of predictive biomarkers (GKDB, CIViC) and are reported according to their clinical evidence (see Level of Evidence). In the following table, basic information of the somatic variants with relevant clinical implications can be found:

Gene	Patient's Variant	Level of Evidence
APC	V1822D	B3
BRCA1	Q614PX, Q1756PX, Q652PX, Q1777PX, Q1709PX, Q66PX, Q247PX	B1 , B2
BRCA2	V2466A	B1 , B2 , A2 B2 , A2
ERCC6	R850K	B3
FAT1	K4059N, K4061N, V862L, A43V	B3
SETD2	P1962L, P1596L	B3

Level of Evidence: Findings are classified into six levels of evidence combining the axis A-B and the axis 1-2-3. Level A means evidence was found in the same cancer type. Level B means evidence was found in any other cancer type. On the 1-2-3 axis, level 1 means the evidence is supported by drug approval organizations or clinical guidelines, level 2 contains a clinical evidence (clinical trials, case reports) and level 3 consists of a preclinical evidence.

The report summarizes all predictive associations in a detailed table. The results are sorted by 1) drug frequency and 2) level of evidence (A1-B1-A2-B2-A3-B3). To allow a quick interpretation, the type of the association (response, resistance) is colored (green, red) and new variants are gray and underlined.

Patient				Gene-Drug	Associations			
Gene	Variant	Disease	Known Variant	Association	Drugs	Evidence	PMID	Level
BRCA1	Q614PX Q1756PX Q652PX Q1777PX Q1709PX Q66PX Q247PX	unspecified	any mut.	sensitivity/respons	Olaparib	approved	19553641	B1
BRCA2	V2466A V2466A	unspecified	any mut.	sensitivity/respons	Olaparib	approved	19553641	B1
			any variant (LoF)	response	PD1 blockade	case report	26997480	B2

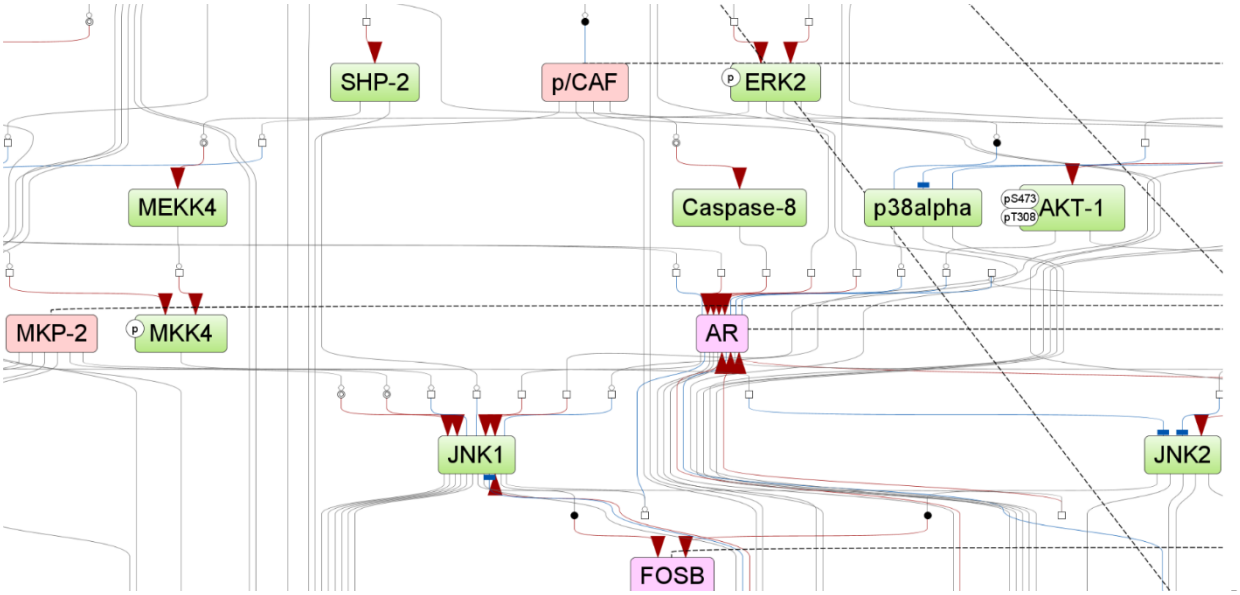
This tool and the corresponding report structure were developed by Julia Perera-Bel in the research group of Prof. Dr. Tim Beißbarth at the University Medical Center Göttingen (UMG).

Perera-Bel J, Hutter B, Heining C, et al. From somatic variants towards precision oncology: Evidence-driven reporting of treatment options in molecular tumor boards. *Genome Med*. 2018;10(1):18. Published 2018 Mar 15. doi:10.1186/s13073-018-0529-2).

Calculate CMA regulation

Composite modules (CMA) are combinations of transcription factor binding sites (TFBSs), which are common in promoters of functionally related genes. A postulated CMA is responsible and the major component of transcriptional regulation for a given gene expression pattern of a set of genes.

The method calculates regulatory scores for transcription factors (TFs) from a CMA result and results from Master regulator (keynode) search. The result of Calculate CMA regulation is a visualization (see below) of the postulated Master regulators together with the Composite modules, underlying genes, and predicted feed forward loops (red triangle) as transcriptional regulation.



Create profile from CMA model

A profile is a matrix collection of given transcription factors (TFs). The method takes a cutoff for each matrix from a user's CMA model result (should be >0). A new profile is created, which uses the described cutoffs and other profile parameters from the original TRANSFAC® profile, that was used for the CMA calculation. This new profile can be used to determine and identify CMA's based on the user specific CMA.

Find regulatory regions with mutations

The method scans genes with flanking regions (= window of an input size of 1100 bp usually) and defines a window position with the highest sum of mutation weights inside the whole window. Weight and gene properties are optional and are calculated with 'Mutation weight calculator' analysis. If no weight property is given, each mutation gains a weight of 1.

If zero mutations are found in the window around the gene, Fantom5 transcriptional start site (TSS) position knowledge is used for a selected tissue or Ensembl transcriptional start site (TSS) position knowledge is used if no tissue was selected.

The output is a track file, which can be visualized with geneXplain's platform genome browser.

PSD pharmaceutical compounds analysis

This method is based on HumanPSD™ database, where a valid license is needed to perform this feature. As input a table with genes is used. The output shows a ranking of potential drug targets. The gene table can have an optional weight column that might be used to rank the postulated drug targets. The calculated rank will be used if no weight column is selected or present in the original input gene list.

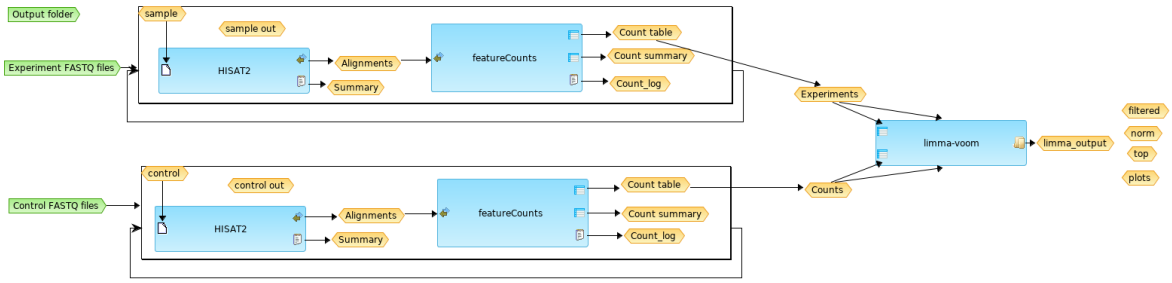
There are two output tables: One output table with drug name, drug chemical structures names and various drug annotation. The second output table with potential targets is a subset of the input genes matched to given and known drugs in the HumanPSD™ 2020.2 database.

New workflows

There are three new workflows, which uses state-of-the-art tools to align raw FASTQ files from RNA sequencing data to human hg38 genome, calculate feature counts for genes and perform a statistical analysis to identify differentially expressed genes in a sample collection compared to a control sample collection. The user only defines folders with samples and control samples. Results are manifold and include alignment files, alignment summary, quality report, normalized count tables, filtered count tables, DEGs and plots.

Variants of these workflows with other genome versions and for paired reads will soon be available (July 2020).

Full RNAseq analysis with HISAT2, featureCounts and limma



The components of this workflow are:

HISAT2 is a fast and sensitive alignment tool for mapping next-generation sequencing reads (RNA) to a human genome. HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies,

enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

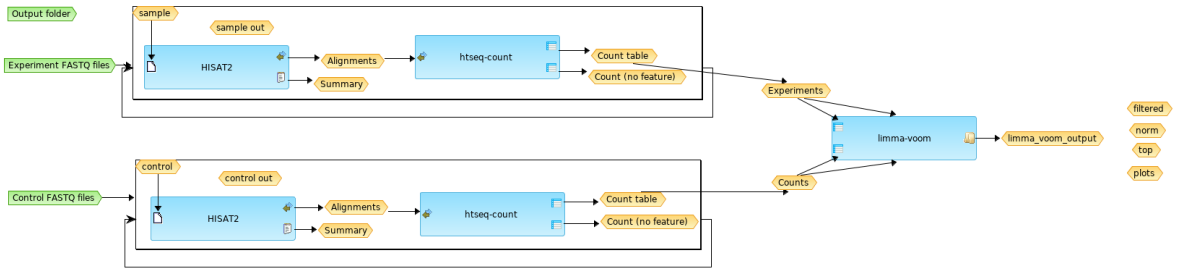
The featureCounts function counts the number of reads or read-pairs that overlap any specified set of genomic features. It can assign reads to any type of genomic region. Regions may be specified as simple genomic intervals (promoter regions) or can be collections of genomic intervals (genes comprising multiple exons). Any set of genomic features can be specified in GTF, GFF or SAF file format. SAF is a Simplified Annotation Format with columns GeneID, Chr, Start, End and Strand.

featureCounts produces a matrix of gene wise counts and can be used as input for gene expression analysis with limma, edgeR or DESeq2. Alternatively, a matrix of exon-level counts can be produced suitable for differential exon usage analyses using limma, edgeR or DESeq2.

featureCounts outputs the genomic length and position of each feature as well as the read count, making it straightforward to calculate summary measures such as RPKM (reads per kilobase per million reads).

limma-voom is a differential expression analysis for pre-processed RNA-seq data (single channel experiments) with sample-specific quality weights when the library sizes are quite variable between samples or the presence of outlier samples is given. The output reports the top100 differentially expressed genes and a pdf document containing density plots from raw and filtered counts, plot about the Mean–variance trend and gives visual information about sample clustering.

Full RNAseq analysis with HISAT2, htseq-counts and limma



The components of this workflow are:

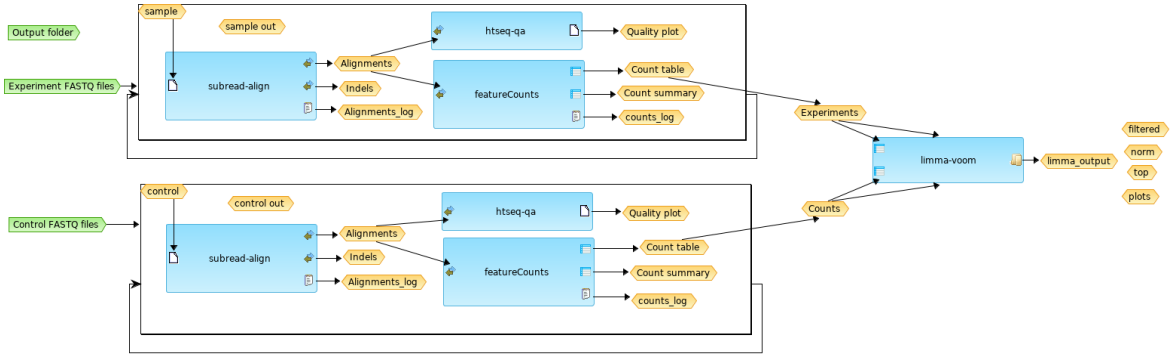
HISAT2 is a fast and sensitive alignment tool for mapping next-generation sequencing reads (RNA) to a human genome. HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome (each index representing a genomic region of 56 Kbp, with 55,000 indexes needed to cover the human population). These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

HTseq-count takes an alignment file in SAM or BAM format and a feature file in GFF format and calculates the number of reads mapping to each feature. It uses the *htseq-count* script that is part of the HTSeq python module.

A feature is an interval (i.e. a range of positions) on a chromosome or a union of such intervals. In the case of RNA-Seq, the features are typically genes, where each gene is considered here as the union of all its exons.

limma-voom is a differential expression analysis for pre-processed RNA-seq data (single channel experiments) with sample-specific quality weights when the library sizes are quite variable between samples or the presence of outlier samples is given. The output reports the top100 differentially expressed genes and a pdf document containing density plots from raw and filtered counts, plot about the Mean–variance trend and gives visual information about sample clustering.

Full RNAseq analysis with subread, featureCounts and limma



The components of this workflow are:

Subread-align is a general-purpose read aligner which can align RNA-seq reads, based on its unique seed-and-vote design, by which a large number of 16mer subreads from each read are mapped to the reference genome. The subread function accept raw reads, in the form of Fastq, SAM or BAM files, and output read alignments in either SAM or BAM format. The output contains the total number of reads, the number of uniquely mapped reads, the number of multi-mapping reads and other mapping statistics. The align function is exceptionally flexible. It performs local read alignment and reports the largest mappable region for each read.

The featureCounts function counts the number of reads or read-pairs that overlap any specified set of genomic features. It can assign reads to any type of genomic region. Regions may be specified as simple genomic intervals (promoter regions) or can be collections of genomic intervals (genes comprising multiple exons). Any set of genomic features can be specified in GTF, GFF or SAF file format. SAF is a Simplified Annotation Format with columns GeneID, Chr, Start, End and Strand.

featureCounts produces a matrix of gene wise counts and can be used as input for gene expression analysis with limma, edgeR or DESeq2. Alternatively, a matrix of exon-level counts can be produced suitable for differential exon usage analyses using limma, edgeR or DESeq2.

featureCounts outputs the genomic length and position of each feature as well as the read count, making it straightforward to calculate summary measures such as RPKM (reads per kilobase per million reads).

limma-voom is a differential expression analysis for pre-processed RNA-seq data (single channel experiments) with sample-specific quality weights when the library sizes are quite variable between samples or the presence of outlier samples is given. The output reports the top100 differentially expressed genes and a pdf document containing density plots from raw and filtered counts, plot about the Mean–variance trend and gives visual information about sample clustering.

Enhancements

Workflows RNAseq analysis with HISAT2 or Subread

Transparent genome version control is now available for the two RNAseq alignment workflows: RNAseq analysis with HISAT2 and RNAseq analysis with Subread.

Integration of the following genome versions for the alignment of reads and selection within the input mask of the workflows:

Ensembl GRCh38, Ensembl GRCh37, Ensembl NCBIM37, Ensembl NCBIM38 and Ensembl RGSC3.4.