How to generate a custom protein database from RNAseq data

Step 1: Obtain raw-sequencing files (.fastq)

Criteria for choosing a suitable sequencing experiment:

- How many replicates?
- State of the art sequencing machine?
- is the sequencing single-end or paired-end? Paired-end sequencing has 2 fastq files per sample and is preferred
- Sequencing depth?
- Availability of the data? E.g. http://www.ebi.ac.uk/ena/ or https://www.ncbi.nlm.nih.gov/sra/$
- What strand protocol was used (important in Step 2)

Go to: http://www.ebi.ac.uk/ena/data/view/PRJNA297633 and download fastq files of runs SRR2549078 and SRR2549079 (file 1 and file 2) to a data directoy of choice. The data directory should have ~100 GB of free space and be accessible from the euler cluster (e.g. personal euler scratch space). Decompress the fastq files.

Step 2: Configure the RNAseq analysis parameters

There are two shell scripts needed to perform the analysis pipeline on euler.

- rnaseq_pipeline.sh
- rnaseq_pipeline_config.sh

The first contains the calls to the analysis tools and needs to be executed in Step 3.The second contains the parameters for the used tools and is used to configure the pipeline for each run.

Copy the config file for each runyou want to make:

cp rnaseq_pipeline_config.sh rnaseq_pipeline_config_sample1.sh

Open the cpoied file and set the parameters:

- reference: Set path to the reference genome
- annotation: Set path to the annotation file (both ref genome and annotation can be obtained from https://www.gencodegenes.org for mouse and human)
- Set path to the input files (if single end the second file should be set to the same fastq)
- Set library-type parameter

```
# Config file for all input parameters neccessary to run the pipeline
# Change parameters according to your input

# Genome reference FASTA file (full path)
reference=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/GRCh38.primary_assembly.genome.fa

# Genome annotation file (full path)
annotation=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/gencode.v25.primary_assembly.annotation.gtf
#Fastq input files (do not need to be trimmed). If single-end reads second file is ignored (full path)
in_fq_1=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/HEK293/RNA_seq/Raw_sequences/SRR2549078_1.fastq
in_fq_2=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/HEK293/RNA_seq/Raw_sequences/SRR2549078_2.fastq

#-----------------------------------------------------


#Trimmomatic inputs
        # Set number of threads trimmomatic should use (24 seems to work well)
        tr_threads=24

        #Define adapter properties (adapter fasta is distributed with trimmomatic under http://www.usadellab.org/cms/?page=trimmomatic)
        #ILLUMINACLIP:<fastaWithAdapterSeqs>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>:<minAdapterLength>:<keepBothReads>
        #seedMismatches: specifies the maximum mismatch count which will still allow a full match to be performed;palindromeClipThreshold: specifies how accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment.;simpleClipThreshold: specifies how accurate the match between any adapter etc. sequence must be against a read.
        ILLUMINACLIP=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Trimmomatic_adapters/TruSeq3-SE.fa:2:30:10

        #LEADING:<quality>: quality: Specifies the minimum quality required to keep a base
        LEADING=3

        #TRAILING:<quality>: quality: Specifies the minimum quality required to keep a base
        TRAILING=3

        #SLIDINGWINDOW:<windowSize>:<requiredQuality>: windowSize: specifies the number of bases to average across ; requiredQuality: specifies the average quality required.
        SLIDINGWINDOW=4:15

        #MINLEN:<length> ; length: Specifies the minimum length of reads to be kept.
        MINLEN=36


#-----------------------------------------------------


# STAR inputs

        # Set number of cores STAR will use for indexing the genome and read mapping
        st_threads=48

        # specifies path to the directory (henceforth called "genome directory" where thevgenome indices are stored. This directory has to be created (with mkdir) before STAR run and needs to writing permissions
        genomeDir=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/STAR_genome_index

        # specifies one or more FASTA files with the genome reference sequences
        genomeFastaFiles=$reference

        #  specifies the path to the file with annotated transcripts in the standard GTF format
        sjdbGTFfile=$annotation

        # specifies the length of the genomic sequence around the annotated junction to be used in constructing the splice junctions database. Ideally, this length should be equal to the ReadLength-1, where ReadLength is the length of the reads
        sjdbOverhang=99

        # Removes non-canonical spilce junctions (Recommended in STAR manual for cuffinks compatibility)
        outFilterIntronMotifs=RemoveNoncanonicalUnannotated



#-----------------------------------------------


# Samtools inputs

        # Sets number of cores Samtools uses for sorting bam files, recommended: 8
        sa_threads=8


#-----------------------------------------------


# Cufflinks inputs

        # Sets the number of cores used for transcript assembly
        cl_threads=24

        # Specifies the library preparation. Chosse: fr-unstranded (e.g. unstranded Illumina TruSeq), fr_firststrand (dUTP, NSR, NNSR) , etc.
        librarytype=fr-firststrand
```

Other options should be set according to the experiment.

Step 3: Run the analysis script

To run the analysis script log onto euler and copy both the analysis script and the config file to an accessible location.

Make a folder in your data directory for the results:

Cd data/

mkdir sample_1_results

Go to the folder where the run script is stored and make it executable

chmod 755 rnaseq_pipeline.sh

Execute the script without input parameters to get help text

./ rnaseq_pipeline.sh

```
[mfrank@euler04 2017-01-17]$ ./rnaseq_pipeline.sh
------------------------------------------------------------
Trims, aligns and assembles transcript reads from illumina sequencing runs to a transcriptome fasta on the Euler cluster. Par
specified output folder and consist of .sam and .bam alignments from STAR, transcripts in .gtf and .fa format from cufflinks.

USAGE INFORMATION:

rna-seq-pipeline PAIRED[0|1] STAGE[0:4]  RUN_NAME OUT_DIR Path/to/parameter_file

PAIRED  1=Paired end sequencing; 0=Single end sequencing
STAGE   Start with 0:FASTQC, 1: TRIMMOMATIC, 2: STAR, 3: SAMTOOLS, 4:CUFFLINKS
RUN_NAME        Name of the Sequencing run, output files will be saved under that name
OUT_DIR Path to desired output directory
PARAMS Path/to/parameter_file Path to shell script containing parameter variables
```

Execute the script with the right input parameters

./rnaseq_pipeline.sh 1 0 Hek293_R2 /cluster/scratch/mfrank/Hek293/sample_1_results rnaseq_pipeline_config_sample1.sh

```
[mfrank@euler04 2017-01-17]$ ./rnaseq_pipeline.sh 1 0 Hek293_R2 /cluster/scratch/mfrank/Hek293/Hek293_R2 rnaseq_pipeline_config_R2.sh

1st FASTQ Input file '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/HEK293/RNA_seq/Raw_sequences/SRR2549079_1.fastq' checked
2nd FASTQ Input file '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/HEK293/RNA_seq/Raw_sequences/SRR2549079_2.fastq' checked
Reference Fasta file '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/GRCh38.primary_assembly.genome.fa' checked
Genome annotation file '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/gencode.v25.primary_assembly.annotation.gtf' checked

Submitting FastQC-raw read quality control job
Generic job.
Job <35831383> is submitted to queue <normal.4h>.
Submitting Trimmomatic job
Generic job.
Job <35831384> is submitted to queue <normal.4h>.
Submitting FastQC-trimmed read quality control job
Generic job.
Job <35831385> is submitted to queue <normal.4h>.
Found indexed genome in '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/STAR_genome_index'. Starting to map reads with Star...
Generic job.
Job <35831386> is submitted to queue <normal.4h>.
Submitting Samtools SAM/BAM conversion job
Generic job.
Job <35831387> is submitted to queue <normal.4h>.
Submitting Samtools BAM sort job
Generic job.
Job <35831388> is submitted to queue <normal.4h>.
Submitting Cufflinks Aglinment job
Generic job.
Job <35831389> is submitted to queue <normal.4h>.
Submitting Transcriptome-GTF to Fasta conversion job
Generic job.
Job <35831390> is submitted to queue <normal.4h>.
All jobs submitted successfully. You can monitor them with bjobs. An email will be sent to your NETHZ adress when the last job has finished.
```

The script checks the input files and sets up a folder structure within the output folder. Then it submits Several jobs to the euler queue: Quality control with FastQC, Adapter trimming with

Trimmomatic, Alignment with Star, .sam to .bam conversion with samtools and transcriptome assembly with Cufflinks. If the analysis is at an intermediate stage it can be resumed mid-way using the STAGE parameter.

The outputs are saved in each respective folder in the results directory. It may be advisable to delete large intermediary files and keep only the ones used for further analysis

-fastqc contains quality control plots

- star/<Hek293_R2>SJ_out.tab contains the position of all detected splice junctions

- cufflinks/<Hek293_R2>_transcripts.gtf assembled transcripts with FPKM values and more

-cufflinks/<Hek293_R2>_isoforms.fpkm_tracking FPKM of all transcripts

-cufflinks/<Hek293_R2>_genes.fpkm_tracking FPKM of all genes

-samtools/<Hek293_R2>Aligned.out_sorted.bam binary alignment file, needed for variant calling

Step 4: Configure the variant call script

There are two shell scripts needed to perform the analysis pipeline on euler.

- call_variants.sh
- call_variants_config.sh

The first contains the calls to the analysis tools and needs to be executed in Step 5.The second contains the parameters for the used tools and is used to configure the pipeline for each run.

This time all samples can be run with the same config script. Copy both scripts to a place that is accessible from euler and set the parameters in call_variants_config.sh.

- reference: Set path to the reference genome (should be the same as used before)
- DBSNP: Path to vcf file with all dbsnp variants for that species (can be obtained at https://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/#table-1)
- Star parameters should be the same as before

The other parameters can be adjusted to the needs of the analysis but the default represents best practices guidelines from GATK.

```
# Config file for all input parameters neccessary to run the pipeline
# Change parameters according to your input

# Genome reference FASTA file (full path)
reference=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/GRCh38.primary_assembly.genome.fa

# Genome annotation file (full path)
# annotation=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/gencode.v25.primary_assembly.annotation.gtf
#Fastq input files (do not need to be trimmed). If single-end reads second file is ignored (full path)
# in_fq_1=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/HeLa/RNA_seq/Raw_sequences/Vanshee_HeLa-R_AGTTCC_L002_R1_001.fastq
# in_fq_2=/cluster/scratch/mfrank/SRR2549078_2.fastq


#-------------------------------------------------------

# Picard MarkDuplicates

        # Setting stringency to SILENT can improve performance when processing a BAM file in which variable-length data (read, qualities, tags) do not otherwise need to be decoded.
        VALIDATION_STRINGENCY=SILENT

        # Additional Parameters
        MD_additional_params=""
#-------------------------------------------------------

# HaplotypeCaller Parameters

        # The minimum phred-scaled confidence threshold at which variants should be called
        stand_call_conf=20.0
        stand_emit_conf=20.0

        # Additional Parameters
        HC_additional_params=""

#-------------------------------------------------------

# VariantFiltration Parameters

        # The window size (in bases) in which to evaluate clustered SNPs
        window=35

        # The number of SNPs which make up a cluster
        cluster=3

        # Filter FisherStrand values in the info column above a threshold
        FSfilter="FS > 30.0"

        # Filter Qual By Depth values in the info column above a threshold
        QDfilter="QD < 2.0"

        # Additional Parameters
        VF_additional_params=""

#-------------------------------------------------------

# CollectVariantCallingMetrics Parameters

        #Path to dbSNP vcf file for the right species (must be indexed, e.g. with igvtools)
        DBSNP=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/dbSNP/All_20161122.vcf



# STAR inputs

        # Set number of cores STAR will use for indexing the genome and read mapping
        st_threads=48

        # specifies path to the directory (henceforth called "genome directory" where the genome indices are stored. This directory has to be created (with mkdir) before STAR run and needs to writing permissions
        genomeDir=/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/STAR_genome_index
        # specifies one or more FASTA files with the genome reference sequences
        genomeFastaFiles=$reference

        #  specifies the path to the file with annotated transcripts in the standard GTF format
        sjdbGTFfile=$annotation

        # specifies the length of the genomic sequence around the annotated junction to be used in constructing the splice junctions database. Ideally, this length should be equal to the ReadLength-1, where ReadLength is the length of the reads
        sjdbOverhang=99

        # Remove non-canonical splice junctions (Recommended in STAR manual for cufflinks compatibility)
        outFilterIntronMotifs=RemoveNoncanonicalUnannotated


#-------------------------------------------------------


# Samtools inputs

        # Sets number of cores Samtools uses for sorting bam files, recommended: 8
        sa_threads=8


#-------------------------------------------------------
```

Step 5: Call variants

Calls variants, according to the recommended gatk practice (see https://software.broadinstitute.org/gatk/guide/article?id=3891). Requires an aligned input file from star and a reference genome.

To run the analysis script log onto euler and copy both the analysis script and the config file to an accessible location.

Make a folder in your data directory for the results:

Cd data/

mkdir sample_1_variant_calls

Go to the folder where the run script is stored and make it executable

chmod 755 call_variants.sh

Execute the script without input parameters to get help text

./call_variants.sh

```
[mfrank@euler10 2017-01-17]$ ./call_variants.sh
------------------------------------------------------------------
Calls variants, according to the recommended gatk practice (see https://software.broadinstitute.org/gatk/guide/article?id=3891).
le from star and a reference genome.

USAGE INFORMATION:

rna-seq-pipeline INPUT  RUN_NAME OUT_DIR Path/to/parameter_file

INPUT   Input file, Sam file produced by STAR
RUN_NAME  Name of the Sequencing run, output files will be saved under that name
OUT_DIR   Path to desired output directory
Path/to/parameter_file Path to shell script containing parameter variables
```

Execute the script with the right input parameters

./call_variants.sh sample_1Aligned.out_sorted.bam Hek293_R2 /cluster/scratch/mfrank/Hek293/sample_1_variant_calls call_variants_config.sh

```
[mfrank@euler10 Hek293_variant_call_analysis]$ ./call_variants.sh ~/mysonas/Master_Project/data/HEK293/RNA_seq/Alignment/Hek293_R1Aligned.out_sorted.bam Hek2
93_R1 ./Hek293_R1 call_variants_config.sh
Reference Fasta file '/nfs/nas21.ethz.ch/nas/fs2102/biol_ibt_usr_s1/mfrank/Master_Project/data/Human_genome/GRCh38/Gencode_v25/GRCh38.primary_assembly.genome
.fa' checked
STAR Alignment File file '/cluster/home/mfrank/mysonas/Master_Project/data/HEK293/RNA_seq/Alignment/Hek293_R1Aligned.out_sorted.bam' checked

Submitting jobs to sort/index SAM file and mark duplicates (AddOrReplaceReadGroups + MarkDuplicates)
Generic job.
Job <35839482> is submitted to queue <normal.4h>.
Generic job.
Job <35839491> is submitted to queue <normal.4h>.
Submitting Split'N'Trim Job to hardclip splice junction overhangs (SplitNCigarReads)
Generic job.
Job <35839492> is submitted to queue <normal.4h>.
Submitting Variant calling Job (HaplotypeCaller)
Generic job.
Job <35839493> is submitted to queue <normal.4h>.
Submitting Variant filtering Job (VaraintFiltration)
Generic job.
Job <35839494> is submitted to queue <normal.4h>.
Submitting Quality Control comparison to dbSNP Job (CollectVariantCallingMetrics)
Generic job.
Job <35839495> is submitted to queue <normal.4h>.
Generic job.
Job <35839496> is submitted to queue <normal.4h>.
done
```

The pipeline will set up a folder structure in the output directory. One can look through the log files to see if the pipeline behaved correctly. The main output files are

- variant_output/Hek293_R1_filtered_variants.vcf It contains the SNVs and INDELs in tabular format that can be read in Step 6.
- Variant_output/Hek293_R1_filtered_variants_metrics.variant_calling_summary_metrics
  Summary statistics of the output file.

One can quickly assess if the output makes sense by looking at the summary metrics file. There, the output is compared to dbSNP and one would expect to find a high percentage of variants to be annotated, given one works with a well-studied species.

Step 6: Generate custom fasta files

This step is based on the R-package customProDB and takes 3 files (for every replicate as an input:

- .vcf file from step 5
- isoforms.fpkm_tracking file from step 3
- .SJ_out.tab file from step3

The R-package RNASeqToCustomFasta is used to run customProDB functions and also provides additional functionality and generates some plots. A run-script is available that guides the user through the necessary steps.