3’ RNA sequencing for robust and low-cost gene expression profiling

Eric Charpentier (✉ eric.charpentier@univ-nantes.fr)
CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS016, CNRS UMS 3556, Université de Nantes, 44000, Nantes, France https://orcid.org/0000-0002-8571-7603

Marine Cornec (✉ marine.cornec@univ-nantes.fr)
CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS016, CNRS UMS 3556, Université de Nantes, 44000, Nantes, France https://orcid.org/0000-0001-9102-753X

Solenne Dumont
Université de Nantes, CHU Nantes*, CNRS, INSERM, l’institut du thorax, F-44000 Nantes, France.
https://orcid.org/0000-0003-3237-7382

Dimitri Meistermann
Université de Nantes, CHU Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, 44000 Nantes, France https://orcid.org/0000-0001-6614-2325

Philippe Bordron
Université de Nantes, Inserm, TENS, The Enteric Nervous System in Gut and Brain Diseases, IMAD, Nantes, France https://orcid.org/0000-0003-1975-0920

Laurent David
Université de Nantes, CHU Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, 44000 Nantes, France https://orcid.org/0000-0003-3594-0353

Richard Redon
Université de Nantes, CHU Nantes*, CNRS, INSERM, l’institut du thorax, F-44000 Nantes, France.
https://orcid.org/0000-0001-7751-2280

Stéphanie Bonnaud (✉ stephanie.bonnaud@univ-nantes.fr)
Université de Nantes, CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS 016, CNRS UMS 3556, F-44000 Nantes, France

Audrey Bihouée (✉ audrey.bihouee@univ-nantes.fr)
Université de Nantes, CHU Nantes, Inserm, CNRS, SFR Santé, Inserm UMS 016, CNRS UMS 3556, F-44000 Nantes, France https://orcid.org/0000-0002-8689-2083

Method Article

Keywords: transcriptome,expression,sequencing,pipeline,snakemake

DOI: https://doi.org/10.21203/rs.3.pex-1336/v1

Abstract

3’seq-RNA Profiling (3’SRP) approach is based on multiplexing samples and molecular indexing mRNA in order to drive genome-wide transcriptional profiling at reasonable cost in comparison to standard RNA-sequencing. The protocol is performed according to the 3'-digital gene expression (3'-DGE) approach developed by the Broad institute. The libraries are prepared from small amounts of total RNA where the mRNA poly(A) tails are tagged with universal adapters, well-specific barcodes and unique molecular identifiers (UMIs). We have improved the fragmentation step by implementing tagmentation based on the activity of a bead-linked transposome. This technique allows sample multiplexing on 96-well plates. Libraries are then sequenced using standard procedures, e.g. on Hiseq2500 or NovaSeq 6000 SP Flow Cells. We have developed a snakemake pipeline including every analysis step from raw fastq de-multiplexing to functional annotation of the differentially expressed genes, producing a complete HTML report for end-user.

Introduction

For the past decade, RNA sequencing has progressively become the gold-standard approach in genome-wide transcriptional profiling. Experimental design has to balance between sufficient numbers of sample measurements and economical cost. Hence, compromising between the number of tested conditions and the number of replicates tends to limit statistical power. Increasing costs are mainly due to limited throughput in library preparation and higher read depth, the latter being required in particular for transcript reconstruction and allele-specific expression analysis. In this context, improving the throughput of library preparation while optimizing sequencing depth would allow accurate analysis of large sample groups at reasonable cost.

We have implemented in our core facility the 3'-digital gene expression (3'-DGE) approach developed by the Broad institute[1,2], based on multiplexing samples and molecular indexing of mRNA molecules. We have in particular improved the fragmentation step by implementing tagmentation based on the activity of a bead-linked transposome. This technique is named 3’SRP for 3’ sequencing RNA Profiling (Fig.1).

3’SRP relies on transcript barcoding with unique molecule identifiers (UMI). As reads reporting the same gene and UMI are filtered, this strategy removes the duplicate reads, which result from multiple counts of the same DNA fragment after the PCR amplification step. Another advantage of 3’SRP is that only 5 million reads are required for transcriptome profiling, while full-length RNA sequencing (RNAseq) usually requires 50 million reads.

This technique allows sample multiplexing on 96-well plates. Libraries are then sequenced using standard procedures, e.g. on Hiseq2500 or NovaSeq 6000 SP Flow Cells. High multiplexing considerably reduces the cost of sequencing per sample.

Regarding data analysis, standard RNAseq tools can be used with some adjustments related to UMI counting. To facilitate data processing in service-oriented settings, we have modified the de-multiplexing step to obtain one fastq and one bam file per sample. This allows us to handle several projects on the same run. This file splitting step also reduces the computation time by using the parallelization power of a HPC cluster. We have also adapted the code to allow for large projects split over several runs. We have developed a snakemake[8] pipeline including every analysis step from raw fastq de-multiplexing to functional annotation of the differentially expressed genes, producing a complete HTML report for end-user.

Reagents

- RNA 6000 Nano Kit (ref 5067-1511, Agilent)
- Well-specific barcoded adapter E3V6NEXT (10µM, Integrated DNA Technologies)
- Maxima H Minus Reverse Transcriptase (ref EP0753, Life technologies)
- Universal adapter E5V6NEXT (100µM, Integrated DNA Technologies)
- Deoxynucleotide (dNTP) Solution Mix (ref N0447L, New England Biolabs)
- Nuclease-Free Water (ref 129114, Qiagen)
- DNA Clean & Concentrator-5 w/ Zymo-Spin IC Columns (Capped) 50 preps (ref ZD4013, Ozyme)
- Exonuclease I [(E. coli) + 10X reaction Buffer] (M0293S/L, New England Biolabs)
- Advantage® 2 PCR Kit (ref 639206, Ozyme)
- SINGV6 primer (10µM, Integrated DNA Technologies)
- Agencourt AMPure XP magnetics beads (ref A63881, Beckman Coulter)
- QUANT-IT dsDNA assay kit (ref Q33130, ThermoFisher scientific
- Nextera™ DNA Flex Library Prep kit (ref #20018704, Illumina)
- Nextera™ DNA CD Indexes (24 Indexes, 24 Samples) (ref #20018707, Illumina)
- i5 primer P5NEXTPT5 primer (5µM, Integrated DNA Technologies)
- High Sensitivity D1000 ScreenTape (ref 5067- 5584, Agilent)
- High Sensitivity D1000 Reagents (ref 5067- 5585, Agilent)
- P5 and P7 primers (100µM, Eurofins)

- standards Lib Quant Standards (Illumina) (ref 07960387001, Roche)
- Master Mix of kit Kapa Sybr Fast LC480 (ref KK4611, Sigma-Aldrich)
- Hiseq Rapid SBS Kit v2-50 cycles (ref FC-402-4022) if sequencing on HiSeq 2500
- Hiseq Rapid PE Cluster Kit v2 (ref PE-402-4002) if sequencing on HiSeq 2500
- NovaSeq 6000 SP Reagent Kit 100 cycles (ref #20027464, Illumina) if sequencing on NovaSeq 6000

Equipment

- NanoDrop (ThermoFisher scientific)
- 2100 Bioanalyzer system (Agilent)
- Strong magnetic field (ref A32782, Alpaqua®)
- Thermocycler
- Varioskan™ LUX (ThermoFisher scientific) or Qubit (ThermoFisher scientific)
- 2200 Tape Station (Agilent)
- Light Cycler 480 (Roche)
- HiSeq 2500 (Illumina) or NovaSeq 6000 (Illumina)

Procedure

Experimental design

A clear definition of the scientific question driving genome-wide transcriptional profiling remains the key factor to determine the optimal experimental design.

A crucial point is the specification of the number of biological replicates to ensure sufficient statistical power. The optimal number of replicates depends mainly on the magnitude of differential expression (effect size) and the complexity of experimental design (number of covariates). However, as described in Alpern et al. article[9], shifting from 20 to 5 replicates greatly reduces the ability to detect true differentially expressed genes (DEGs) (decrease statistical power). The low cost per sample is a great advantage for increasing the number of samples in an analysis and allows a high performance to detect DEGs.

In order to avoid batch effects, RNA extraction series have to be randomized according to biological conditions. A randomization of the samples on micro-well plates is also necessary to avoid rows/columns batch effects; we have developed an algorithm to randomize samples.

Sample qualification

Libraries are prepared from 10ng of total RNA per sample, diluted in 4µl of RNase free water. RNA concentration and purity are measured on NanoDrop station (ThermoFisher scientific); OD 260/230 ratios should be around 1.8 - 2.0 to ensure that concentration is properly assessed. Quality of RNA samples is controlled on 2100 Bioanalyzer system (Agilent) with RNA 6000 Nano Kit (ref 5067-1511, Agilent).

Library preparation

Library construction (Fig.2) for 3'SRP is performed according to Soumillon et al.[1]

The mRNA poly(A) tails are tagged with universal adapters, well-specific barcodes and unique molecular identifiers (UMIs) during template-switching reverse transcription: 96 RNA samples are distributed in a MicroAmp Optical 96-well Reaction Plate (ref N8010560, Applied Biosystems) with 1µl of well-specific barcoded adapter E3V6NEXT (10µM, Integrated DNA Technologies) .

E3V6NEXT = 5'-/5Biosg/ACACTCTTTCCCTACACGACGCTCTTCCGATCT[BC6]N10T30VN-3'

5Biosg = 5' biotin,

[BC6] = 6bp barcode specific to each cell/well,

N10 = Unique Molecular Identifiers

Template-switching reverse transcription was performed using Maxima H Minus Reverse Transcriptase (ref EP0753, Life technologies), Deoxynucleotide (dNTP) Solution Mix (ref N0447L, New England Biolabs) and with universal adapter E5V6NEXT (100µM, Integrated DNA Technologies). No RNA Spike-in was used in our approach.

E5V6NEXT = 5'-/5Me-isodC//iisodG//iMe-isodC/ACACTCTTTCCCTACACGACGCrGrGrG-3'

All 96 cDNAs obtained after template-switching reaction are pooled together, purified and concentrated with single DNA & Concentrator-5 column (ref ZD4013, Ozyme). Pooled and concentrated cDNAs are treated with Exonuclease I (M0293S/L, New England Biolabs) and then amplified by single primer PCR using Advantage® 2 PCR Kit (ref 639206, Ozyme) and SINGV6 primer (10µM, Integrated DNA Technologies) :

*SINGV6 = 5'-/5Biosg/ACACTCTTTCCCTACACGACG*C-3'*

12 cycles of PCR amplification are applied. PCR products are purified with Agencourt AMPure XP magnetics beads [0.6x] (ref A63881, Beckman Coulter) and quantified spefically for dbDNA on Varioskan™ LUX (ThermoFisher scientific) with QUANT-IT dsDNA assay kit (ref Q33130, ThermoFisher scientific).

Tagmentation has been adapted from the original protocol. Amplified barcoded cDNAs are tagmented using a bead-linked transposome (Illumina). Between 100 and 200ng of full-length cDNAs was used as input to the Nextera™ DNA Flex Library Prep kit (ref #20018704, Illumina) and Nextera™ DNA CD Indexes (24 Indexes, 24 Samples) (ref #20018707, Illumina) according to the manufacturer's protocol (Nextera DNA Flex Library Document, ref #1000000025416 v04, Illumina) with exception that :

- the i5 primer was replaced by P5NEXTPT5 primer (5µM, Integrated DNA Technologies)

P5NEXTPT5 = 5'-
*AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC*G*A*T*C*T*-3'*

*where * = phosphorothioate bonds*

- number of PCR cycles was defined at 8

- final elution of purified library was realized in 15µl of RSB (Resuspension Buffer - included Nextera™ DNA Flex Library Prep kit, Illumina)

Library quantification

The size of the resulting sequencing library was controlled on a 2200 Tape Station (Agilent) with High Sensitivity D1000 ScreenTape (ref 5067- 5584, Agilent) and High Sensitivity D1000 Reagents (ref 5067-

5585, Agilent). Expected size of library is between 400 and 800pb.

Library is specifically P5-P7 quantified by qPCR on Light Cycler 480 (Roche) according to standards Lib Quant Standards (Illumina) (ref 07960387001, Roche). P5 and P7 primers (100µM, Eurofins) are added to Master Mix of kit Kapa Sybr Fast LC480 (ref KK4611, Sigma-Aldrich).

P5 = 5'-AATGATACGGCGACCACCGAGAT-3'

P7 = 5'-CAAGCAGAAGACGGCATACGA-3'

The final concentration must be greater than 2nM in 10µl (for HiSeq sequencing) or greater than 15nM in 10µl (for NovaSeq sequencing) to be sequenced.

Sequencing

Libraries are sequenced either on an Illumina HiSeq 2500 using a Hiseq Rapid SBS Kit v2-50 cycles (ref FC-402-4022) and a Hiseq Rapid PE Cluster Kit v2 (ref PE-402-4002) with 16-58 cycles reads according to manufacturer's protocol (Denaturing and Diluting Libraries for the HiSeq® and GAIIx, Part # 15050107v03, Illumina), or on a NovaSeq 6000 using NovaSeq 6000 SP Reagent Kit 100 cycles (ref #20027464, Illumina) with 17-8-105 cycles reads (Illumina recommends to add one cycle for reads 1 and 3 on NovaSeq sequencing) according to the manufacturer's protocol (NovaSeq 6000 Sequencing System Guide Document #1000000019358v11 Material #20023471, Illumina). i7 indexes are read on NovaSeq sequencing due to injection of two plates of 96 libraries on a SP Flow Cell.

Raw fastq pairs used for analysis matched the following criteria:

- the 16 bases of the first read (forward reads) correspond to 6 bases for a designed well-specific barcode and 10 bases for a unique molecular identifier (UMI).

- the second read (reverse reads), 104 bases for NovaSeq (58 bases for HiSeq runs), corresponds to the captured poly(A) RNAs sequence.

Bioinformatics analysis (Fig.1) can be divided in three parts: 1) primary analysis which includes demultiplexing, alignment and counting steps; 2) secondary analysis that performs the differential

analysis; 3) tertiary analysis which is focused on the functional annotation and the representation of differentially expressed genes.

Primary Analysis

The primary analysis (Fig.3) consist in generating the expression matrix containing the raw counts for each sample and for each gene, from raw sequencing data files.

- Illumina basecall files are transformed into paired-end fastq files with Illumina bcl2fastq converter.

- Samples are demultiplexed according to their respective barcodes described in a sample sheet to create a single-end fastq file per sample.

- A step of read trimming is performed in order to remove any poly A tail, which can occur when the transposase cuts a DNA fragment in close proximity to the poly A tail. By using cutadapt on these fragments, all the bases following the poly A tails (reverse sequences of the UMI, sample barcode and illumina adapter) are also removed.

- The cDNA sequences from each sample are then aligned on the reference transcriptome and the mitochondrial genomic reference sequence with BWA[5] aligner. Since a reference transcriptome is used, there is no need for RNA aligner (such as STAR or bowtie).

- Expression profiles are generated by parsing the alignment files (bam) then counting the number of unique UMIs associated with each gene for each sample.

- Reads are removed if one of the following cases occur:

* the UMI contains 'N's in its sequence.

* there are more than 3 mismatches in the alignment.

* the sequence aligns with the same score on transcripts of more that one gene.

- Finally, the raw expression matrix is produced, which contains the values of abundance for every gene in all samples.

Secondary Analysis

The secondary analysis consists in identifying the differentially expressed genes between conditions from the raw expression matrix.

The pipeline is mainly written in the R programming language.

- Samples and genes are filtered out of the raw expression matrix according to the parameters specified in the configuration file:

* Samples are removed if they have less than –minReads reads assigned (default 200k) or if they have less than –minGenes genes detected (i.e. number of genes with at least 1 count) (default 5k).

* Genes are removed if not detected (counts = 0) in at least the number of samples forming the smallest condition.

- Counts are then normalized by DESeq2[6] method.

- Differentially expressed genes are found by DESeq2. By default, genes with an adjusted p-value < 0.05 and a log2 fold-change > 0.58 (i.e. fold-change > 1.5) are defined as differentially expressed.

Tertiary Analysis

Using ClusterProfiler[10] and StringDB[8] tools provides some hints on the interpretation of the differentially expressed genes.

- Enrichment tests are performed on Gene Ontology annotations and KEGG pathways.

- GSEA analysis

- StringDB networks are shown if the number of differentially expressed genes is below 50.

Running the 3'SRP pipeline

The pipeline is entirely implemented as a Snakemake workflow. Scripts used in the pipeline are mainly written in the python and R programming language.

Requirements

Installing git and conda are mandatory before running the pipeline.

All necessary tools are then listed within a conda environment recipe. This environment can be created with a basic conda command.

$ conda env create -n srp -f CONDA/srp.yml

Testing the installation

A minimal sample size test data as well as a sample sheet are available in the "TESTDATA" folder. In order to check the installation, run the pipeline on this data.

This test data is made up of six samples classified in three conditions (2 samples per condition). The fastq files are real reads taken on a small portion of the human chromosome 22. The reference is a fasta file of the human chromosome 22.

$ conda activate srp

 # Under the srp-pipeline directory:

$ python SCRIPTS/make_srp_config.py -s TESTDATA/samplesheet.tsv -r TESTDATA/REFERENCES/ -w RESULTS -f TESTDATA/fastqFiles.txt -c TESTDATA/conditions.tsv –minGenes 0 –minReads 0 > config.json

$ snakemake –config conf="config.json" -rp -j 1

Input files

The pipeline takes as input a samplesheet describing samples, one or multiple pairs of fastq files and eventually a file listing the comparisons to test.

–The samplesheet :

A samplesheet describing the samples is necessary to run the pipeline. This file has to be tab delimited (without header) containing six columns: well, index, name, project, condition, species.

–The fastq files :

The fastq files have to be in paired-end mode. The first file should contain the sequences of the sample indexes as well as the unique molecule identifiers (UMI) (6+10 bases). The second file should contain the DNA sequences of the captured polyA RNAs (N bases). There are two ways of specifying the fastq files to the program:

- Creating a file listing the fastq paths

- Specifying a Illumina directory (bcl2fastq output folder)

–The comparisons file :

In order to perform secondary analysis for all or some projects, a file listing the project name and comparisons has to be create.

Running the pipeline

1- Clone this repository and move to it

___

$ git clone "https://gitlab.univ-nantes.fr/bird_pipeline_registry/srp-pipeline.git"

$ cd srp-pipeline

___

2- Activate the conda environment

___

$ conda activate srp

___

3- Create the configuration file necessary for the pipeline.

The script used to create the configuration file is make_srp_config.py in the SCRIPTS folder. The help can be visualized with:

___

$ python SCRIPTS/make_srp_config.py -h

—

The mandatory options are **-s for the samplesheet**, either **-f for a file** listing the fastq paths or **-i for an illumina directory** and **-r for the directory containing the genome files** for the different assemblies.

It is recommended to specify a working directory where the files will be **output with option -w** to keep the srp-pipeline git clone clean. The program outputs a json object on stdout which can be redirected to a file.

—

$ python SCRIPTS/make_srp_config.py -s <my_samplesheet> -r <path_to_reference_folder> -w <path_to_workdir> -i <path_to_fastqs> > config.json

—

For secondary analysis, use **option -c** to specify the comparisons to perform.

—

$ python SCRIPTS/make_srp_config.py -s <my_samplesheet> -r <path_to_reference_folder> -w <path_to_workdir> -i <path_to_fastqs> -c <comparisons_file> > config.json

—

The generated configuration file should be checked to see if everything seems alright.

—

$ cat config.json

—

4- Launch the snakemake pipeline.

Test the launch with a **dry run**:

—

$ snakemake –config conf="config.json" -rpn

—

All rules and commands that will be run should be displayed. Otherwise errors are returned.

Launch the run:

$ snakemake –config conf="config.json" -rp -j 1

In order to launch the pipeline **on a cluster**, specify a script to encapsulate the jobs to snakemake. example for SGE:

$ snakemake –config conf="config.json" –cluster "qsub -e ./logs/ -o ./logs/" -j 33 –jobscript SCRIPTS/sge.sh –latency-wait 100 -rp

---

Troubleshooting

Note on sequencing run qualification in Sequence Analysis Viewer (SAV, Illumina):

- Preliminary quality control of the data is carried out on SAV (Illumina). % cluster PF is calculated on the first 25 bases, but only 16 or 17 bases are read for R1. Thus, % Cluster PF obtained is not relevant.

Note on samplesheet:

- The file should be tab delimited without trailing or leading spaces.

- Only characters a-zA-Z-_ are allowed.

- There should be no empty lines.

- There should be no header.

Note on fastq files list:

- The file should be tab delimited without trailing or leading spaces.

- Fastq files listed can be .fastq or .fastq.gz

- There should be no empty lines.

- There should be no header.

Note on comparison file:

- The project name and conditions must match one or more samples in the samplesheet.

- There should be no empty lines.

- There should be no header.

- Same condition can be specified in column 2 and 3 to perform only the first part of the secondary analysis (all but comparisons). The project has to appear only once in the file.

- The first condition column is the test and the second is the control.

Note on the conda environment:

- If the environment for this pipeline has already been created the environment for this pipeline, just “conda activate srp”.

Note on snakemake launch:

- Specify the number of jobs with -j <N>.

- □ Even if multiple jobs are not specified, two scripts in the pipeline are still threaded. .Thus it is advised not to run the pipeline on a real dataset without an HPC.

- The path to the log output files must exist ($ mkdir ./logs).

Time Taken

Experimental protocol

- Quantitative and quality control of RNA: 3 hours

- Normalisation and randomization of samples in 96 well-plate : ~2-3 hours

- Day 1 (RT, cDNA pooling, Exonuclease, PCR, quantification) : ~7 hours

- Day 2 (Tagmentation, Indexing PCR, QC, qPCR) : ~6 hours

- Sequencing run : ~13 hours (NovaSeq 6000) ~16 hours (HiSeq 2500)

Bioinformatics

- Snakemake tiny dataset analysis on a PC (16Go RAM, 8 cores): 5 minutes

- Real analysis with 96 samples on HPC cluster with 35 parallel jobs: 2 hours

Anticipated Results

The main strength of 3’SRP is that it uses UMI barcoding. UMI barcoding has been shown to prevent PCR amplification, improving the accuracy of quantification of mRNA molecules. However, unlike RNAseq, transcriptome analysis by 3’SRP is limited to quantitative studies. Indeed, only the 3' end of the genes is captured and sequenced, excluding isoform reconstruction, novel gene discovery or SNP studies.

As demonstrated in the article by Xiong et al.[3] there is good concordance of outcomes in terms of differential expressed genes and biological interpretation between these two methods, 3’SRP being slightly less sensitive on gene detection (15%). However, this technique shows a good balance between benefits in costs and shortcomings.

We have carried out more than 60 projects on different types of samples. These projects represent more than 6000 samples analyzed. Different types of samples were tested, with RNA extracted from cell culture, primary cells, cell sorting, or biopsies. On average, about 300 million reads are generated by a Hiseq rapid run multiplexing 96 samples and about 1 billion on a SP NovaSeq run multiplexing 2*96 samples. From a total of 27k human genes referenced in RefSeq, a maximum of about 16k expressed genes are detected per sample. This plateau is reached with 5 million reads (Fig.4). Our experience based on these projects has allowed us to optimise library preparation and improve data outputs. Each aspect is detailed below.

Library preparation

- Size library qualification:

The expected library size recommended by Illumina is between 350 and 800 pb. In this protocol, we recommend a library size greater than 400 pb to avoid sequencing the polyA tails. Nevertheless, a trimming of these polyA tails is applied in the primary analysis.

The common size obtained is ~550-650 pb.

- Quantity of library injected:

The quantity of library injected is ~9-10 pM on HiSeq 2500 and ~380 pM on NovaSeq 6000. However, this quantity may vary depending of the sequencer used.

Bioinformatics analysis

The snakemake pipeline is part of our registry and is based on FAIR practices, taking advantage of virtual environments (conda, docker) and continuous integration (jenkins). These best practices allow for pipeline deployment in multiple environments for reproducibility and scalability issues.

The gitlab repository provides a short example dataset, which is quick and easy to use. This small dataset mimes 3 conditions with 2 replicates each aligned on hg19 reference.

More details are available on the gitlab wiki.

Output files & folders:

The main output folder corresponds to the -w specified for the creation of the configuration file or to a default folder named "RESULTS". One folder per project specified in the samplesheet is created.

For each analysis step, a folder with the name of the tool or stage is created containing corresponding result files.

The **CUTADAPT** folder contains the single-end demultiplexed fastq files (one per sample).

The **FASTQC** and **MULTIQC** folders hold the results files of these tools applied on the fastq files.

The **ALIGNMENT** folder contains the aligned reads on the reference transcriptome bam files, as well as their indexes bai files.

The **EXPRESSION** folder contains the result files of the counting step. A naming convention, in four parts, has been defined for the different matrices in this folder.

Prefix name is the name of the project. The second part is either "all" (multiple alignments) or "unq" (align on one gene). The third part of the name defines the information to be found i.e counts matrices, summary... The fourth part of the name, only found in count matrices files can either be "total" or "umi":

- "total" is the number of total reads of the counts without taking into account the UMIs. Counts can be artificially increased during PCR amplification.

- "umi" is the number of unique molecules of RNA, this taking into account the UMIs for each read. If two reads of the same sample map to the same gene and have the same UMI, it will only be counted for 1.

Thus, the expression matrix usually used for secondary analysis is “ProjetcName.unq.refseq.umi.dat".

The **DE** folder contains the result files of the secondary analysis (i.e. the differential expression analysis) as well as the tertiary analysis (i.e. the functional annotation analysis). The folder is composed of the following files:

- filtered, normalized and log-transformed matrices.

- UMIs per million matrix

- sample correlation heatmap

- principal component analysis

- one folder per comparison containing:

* the DESeq2 differential expression output files

* a MA and Volcano plots

* the GO, KEGG and STRINGDB enrichment result files

* the GSEA result files

* a heatmap of the DEG.

Report

The pipeline generates a HTML report using JavaScript/jinja/python technologies. This report, intended for end-users, displays project summary, raw and processed data quality controls and differential analysis results (PCA, differential expressed genes, sample clustering...)

The first part of the report presents the primary analysis: Samples and QC summary, visualization of samples variability. Graphs display the number of total sequences, the read distribution and the number of gene detected (Fig.5A). The HTML report embeds the MultiQC[4] report. All FastQC reports are summarized in one report, containing quality scores, GC and adapter content, overrepresented sequences. The samples variability is shown by a reduction dimension graph (PCA) and a sample clustering (Fig.5B,C).

For each comparison, as described in the config file, Differentially Expressed Genes (DEGs) are visualized by a table (over/under expressed), a MA-plot, a vulcano plot and a clustering of genes (Fig.6).

The DEGs can be displayed dynamically or the full table of DESeq2 results can be downloaded. The available data table is as follow: Gene symbol; base Mean (mean counts of all samples in the project); Log FC (fold-change log transformed); stat (value of DESeq2 statistical test); pvalue; padj (ajusted pvalue by BH method); meanInComp (mean counts of the selected sample in the comparison).

Besides, an analysis with STRING is performed. STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations. The proteins network from STRING depicts how genes, through their proteins, interacts together according to the following evidence sources: Neighborhood in the Genome, Gene Fusions, Co-occurrence Across Genomes, Co-Expression, Experimental/Biochemical Data, Association in Curated Databases, Co-Mentioned in PubMed Abstracts.

The network is displayed in the report, and user can open the interactive view and explore protein interactions (Fig.7B). STRING's list of functional annotation on over-represented and under-represented genes gives another insight from several sources. An interactive table allows to explore each term showing description, statistics, and the list of the under or over-expressed genes matching this term (Fig.7C).

References

1. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. bioRxiv; 2014.

2. Kilens S, Meistermann D, Moreno D, Chariau C, Gaignerie A, Reignier A, Lelièvre Y, Casanova M, Vallot C, Nedellec S, Flippe L, Firmin J, Song J, Charpentier E, Lammers J, Donnart A, Marec N, Deb W, Bihouée A, Le Caignec C, Pecqueur C, Redon R, Barrière P, Bourdon J, Pasque V, Soumillon M, Mikkelsen TS, Rougeulle C, Fréour T, David L and The Milieu Intérieur Consortium. Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. Nat Commun. 2018 Jan 24;9(1):360.

3. Xiong, Y., Soumillon, M., Wu, J. et al. A Comparison of mRNA Sequencing with Random Primed and 3'-Directed Libraries. Sci Rep 7, 14626 (2017).

4. Philip Ewels, Måns Magnusson, Sverker Lundin and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016).

5. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

6. Love MI, Huber W and Anders S (2014). Moderated estimation of fold-change and dispersion for RNA-seq data with DESeq2. Genome Biology ,15, pp. 550.

7. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607-D613.

8. Johannes Köster, Sven Rahmann, Snakemake-a scalable bioinformatics workflow engine, Bioinformatics, Volume 28, Issue 19, 1 October 2012, Pages 2520–2522.

9. Alpern, D., Gardeux, V., Russeil, J. et al. BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. Genome Biol20, 71 (2019).

10. Yu G, Wang L, Han Y and He QY. clusterProfiler: an R package for comparing biological themes among gene clusters.OMICS: A Journal of Integrative Biology. 2012, 16(5):284-287.

Acknowledgements

We are most grateful to Magali Soumillon and Tarjei Mikkelsen at the Broad Institute for sharing their experience on 3’ RNA sequencing. We acknowledge Pierre Lindenbaum for his plate randomizer tool. We also thank Audrey Donnart and Raluca Teusan for their constructive advice and feedback on this protocol.