Quantification RNAseq dont lncRNA

2 types de RNAseq :

1/ RNAseq connus dans le GTF Ensembl

2/ RNAseq de novo reconstruits par StringTie

3 sources de lncRNA :

1/ lncRNA connus dans le GTF Ensembl

2/ lncRNA prédits par FEELnc, et connus dans le GTF Ensembl sous un autre biotype.

3/ lncRNA de novo, issus de la reconstruction StringTie et prédits par FEELnc.

Quantification RNAseq dont lncRNA

3 sources de lncRNA :
1/ lncRNA connus dans le GTF Ensembl
2/ lncRNA prédits par FEELnc, et connus dans le GTF Ensembl sous un autre biotype.
3/ lncRNA de novo, issus de la reconstruction StringTie et prédits par FEELnc.

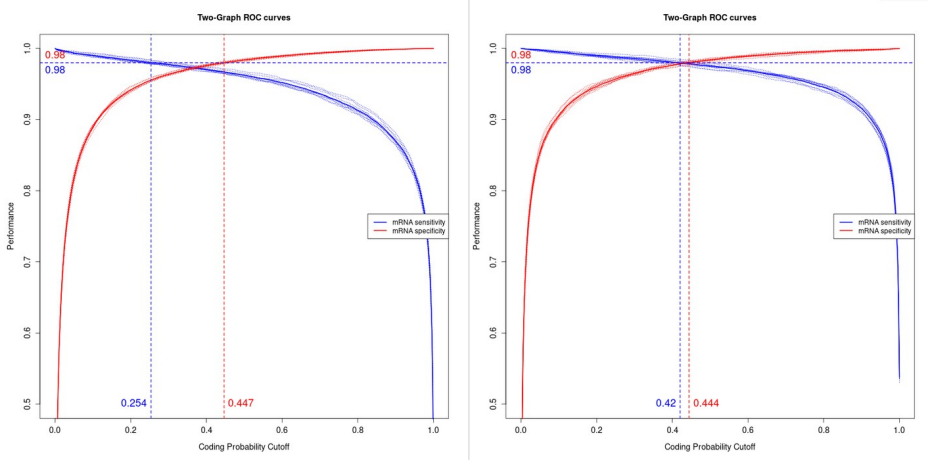
2 types de RNAseq :
1/ RNAseq connus dans le GTF Ensembl
2/ RNAseq de novo reconstruits par StringTie

Nextflow
nf-core/rnaseq
-e ou stringtie_ignore_gtf

→ StringTie merge
Change id order

→ FEELnc
avec **stringtie_ignore_gtf**
pour avoir les (M)STRG

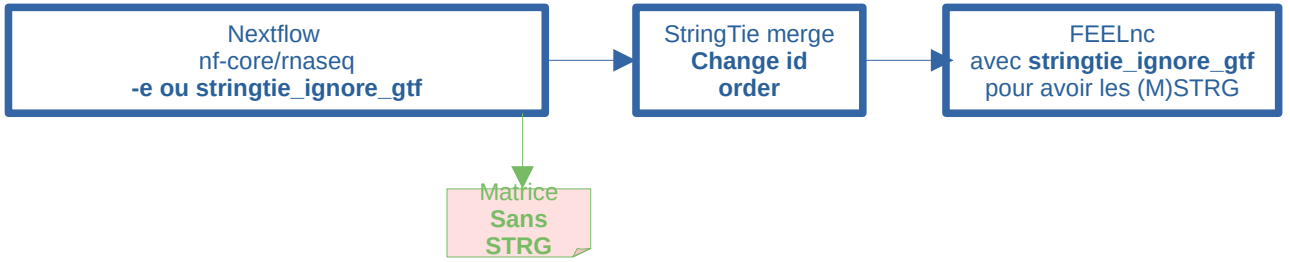
* Récupération de la liste des candidats :lst_intergenic_lncRNA_noORF_WithBiotype.gtf
* Choix entre les méthodes shuffle ou intergenique pour limiter le nombre de transcrits ambigus (mal prédits):



Quantification RNAseq dont lncRNA

3 sources de lncRNA :
1/ lncRNA connus dans le GTF Ensembl
2/ lncRNA prédits par FEELnc, et connus dans le GTF Ensembl sous un autre biotype.
3/ lncRNA de novo, issus de la reconstruction StringTie et prédits par FEELnc.

2 types de RNAseq :
1/ RNAseq connus dans le GTF Ensembl
2/ RNAseq de novo reconstruits par StringTie



hpatel 10 h 35
a répondu à un fil de discussion : **Thanks a lot @hpatel and sorry I didn't find the issue that talks about it.**
Ah, looks like this was a PR to add another tool downstream of StringTie called `prepDE.py` but as you will see explained there I didn't add it in the end:
https://github.com/nf-core/rnaseq/pull/696

> **#696 Add StringTie prepDE.py module**
> Added a local module to generate counts from StringTie results as explained in the docs. The counts will be generated separately for each sample because it is better this way for scaling purposes - especially those that run the pipeline on 100s-1000s samples.
>
> I attempted to add another downstream module to merge the counts across samples but it appears like the ordering of the gene/transcript ids isn't always the same when generated via `prepDE.py`. I think it's best to read these files individually into R and `sort` them accordingly whilst creating a counts matrix.
> **Comments**
> 6
> nf-core/rnaseq | 19 sept. 2021 | Ajouté(e) par GitHub

We decided to keep this pipeline dedicated to standard differential analysis using known annotations a while back now. StringTie should really be added to more of a splice variant detection type pipeline but we still don't have one of those. I kept it in here mainly for convenience because it is arguably the most computationally expensive step within the whole StringTie -> Ballgown type analysis. It is then up to users of the pipeline to merge the GTFs, and take the analysis from there downstream.

I have personally never used it so can't really comment about how it compares to RSEM / Salmon etc in terms of accuracy. However, you would need raw counts and not FPKM / TPM to perform quantification. I believe this is what `prepDE.py` does as explained in the PR above but there are inherent assumptions about read length which is why I chose not to add it to this pipeline.

> Or maybe it is better to merge Ensembl GTF and STringTie GTF and run a new RSEM quantification? Thanks in advance @hpatel

You could try this (untested and may break!)

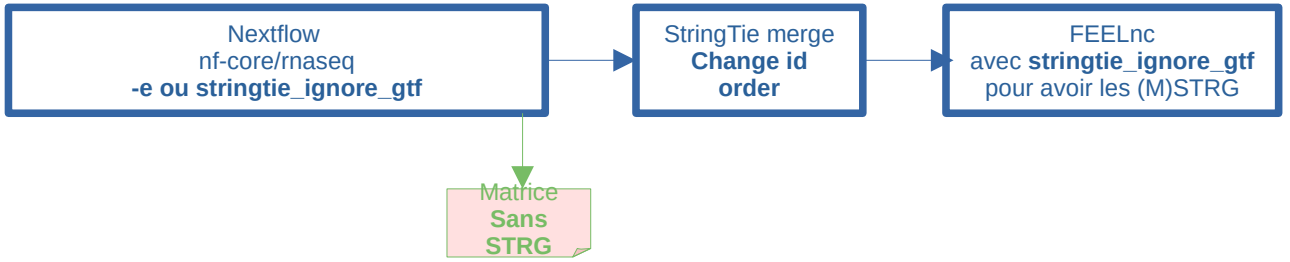
Be interested to know if it works though.

In general, unless you are interested in specifically looking at splice variants then I would just use the standard annotation.

Quantification RNAseq dont lncRNA

3 sources de lncRNA :
1/ lncRNA connus dans le GTF Ensembl
2/ lncRNA prédits par FEELnc, et connus dans le GTF Ensembl sous un autre biotype.
3/ lncRNA de novo, issus de la reconstruction StringTie et prédits par FEELnc.

2 types de RNAseq :
1/ RNAseq connus dans le GTF Ensembl
2/ RNAseq de novo reconstruits par StringTie



hpatel 10 h 35
a répondu à un fil de discussion : **Thanks a lot @hpatel and sorry I didn't find the issue that talks about it.**
Ah, looks like this was a PR to add another tool downstream of StringTie called `prepDE.py` but as you will see explained there I didn't add it in the end:
https://github.com/nf-core/rnaseq/pull/696

> **#696 Add StringTie prepDE.py module**
> Added a local module to generate counts from StringTie results as explained in the docs. The counts will be generated separately for each sample because it is better this way for scaling purposes - especially those that run the pipeline on 100s-1000s samples.
>
> I attempted to add another downstream module to merge the counts across samples but it appears like the ordering of the gene/transcript ids isn't always the same when generated via `prepDE.py`. I think it's best to read these files individually into R and `sort` them accordingly whilst creating a counts matrix.
> **Comments**
> 6
> nf-core/rnaseq | 19 sept. 2021 | Ajouté(e) par GitHub

We decided to keep this pipeline dedicated to standard differential analysis using known annotations a while back now. StringTie should really be added to more of a splice variant detection type pipeline but we still don't have one of those. I kept it in here mainly for convenience because it is arguably the most computationally expensive step within the whole StringTie -> Ballgown type analysis. It is then up to users of the pipeline to merge the GTFs, and take the analysis from there downstream.

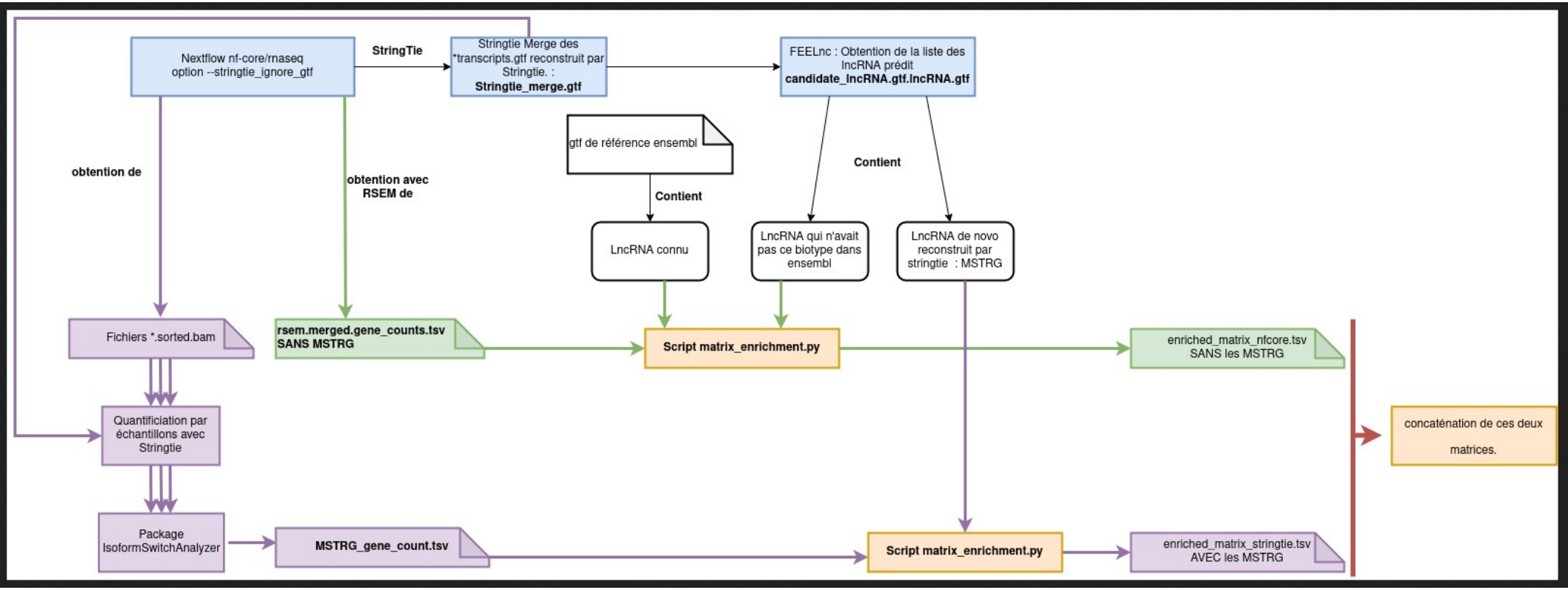
I have personally never used it so can't really comment about how it compares to RSEM / Salmon etc in terms of accuracy. However, you would need raw counts and not FPKM / TPM to perform quantification. I believe this is what `prepDE.py` does as explained in the PR above but there are inherent assumptions about read length which is why I chose not to add it to this pipeline.

> Or maybe it is better to merge Ensembl GTF and STringTie GTF and run a new RSEM quantification? Thanks in advance @hpatel

You could try this (untested and may break!)

Be interested to know if it works though.

In general, unless you are interested in specifically looking at splice variants then I would just use the standard annotation.



Gabryelle AGOUTIN – Sarah MAMAN

Filtres du GTF par StringTie via le package R « IsoformSwitchAnalyzeR »

* 11 134 (15.33%) isoforms were removed since they were not expressed in any samples.

* **Fixing StringTie gene annotation problems : 23 830** isoforms were assigned the ref_gene_id and gene_name of their associated gene_id. This was only done when the parent gene_id were associated with a single ref_gene_id/gene_name.

* 4 095 isoforms were assigned the ref_gene_id and gene_name of the most similar annotated isoform (defined via overlap in genomic exon coordinates). This was only done if the overlap met the requriements indicated by the three fixStringTieViaOverlap* arguments.

* We were unable to assign 266 isoforms (located within annotated genes) to a known ref_gene_id/gene_name.
These were removed to enable analysis of the rest of the isoform from within the merged genes.

* 941 gene_ids which were associated with multiple ref_gene_id/gene_names were split into mutliple genes via their ref_gene_id/gene_names.

* **18 866 genes_id** were assigned their original gene_id instead of the StringTie gene_id. This was only done when it could be done unambiguous.

2 602 transcrits lncRNA ($ grep -i "lncRNA" enriched_matrix.tsv| wc -l)

1 603 + 177 gènes lncRNA ($ grep -i "lncRNA" enriched_matrix_nfcore.tsv| wc -l)

Fil de discussion #rnaseq

hpatel 🌴 21 mars à 10 h 42

Be interested to know if it works though.

1 réponse

Sarah Maman il y a une minute

Hello @hpatel,

We have tested a pipeline to quantify StringTie "de novo" transcripts .

We first run Nextflow nf-core\rnaseq with --stringtie-ignore-gtf option, then StringTie merge on *transcript.gtf files (which contains STRG transcripts), then Stringtie quantification (-e) with this stringtie_merge.gtf, on each sample.

For each sample, a ".ballgown" directory contains a t_data.ctab file on which we run SwithIsoformAnalysis (R package) to convert TPM/FPKM in raw counts and to filter transcripts.

Here are filters given by SwithIsoformAnalysis:
* Remove isoforms since they were not expressed in any samples.
* Fixing StringTie gene annotation problems : some isoforms were assigned the ref_gene_id and gene_name of their associated gene_id. This was only done when the parent gene_id were associated with a single ref_gene_id/gene_name.
* Some isoforms were assigned the ref_gene_id and gene_name of the most similar annotated isoform (defined via overlap in genomic exon coordinates). This was only done if the overlap met the requriements indicated by the three fixStringTieViaOverlap* arguments.
* We were unable to assign some isoforms (located within annotated genes) to a known ref_gene_id/gene_name. These were removed to enable analysis of the rest of the isoform from within the merged genes.
* Some gene_ids which were associated with multiple ref_gene_id/gene_names were split into mutliple genes via their ref_gene_id/gene_names.
* Some genes_id were assigned their original gene_id instead of the StringTie gene_id. This was only done when it could be done unambiguous.

At the end, we concatenate (M)STRGs counts with RSEM matrix generated by the nf-core/rnaseq pipeline.

Faithfully
Gabryelle Agoutin and Sarah Maman (modifié)

Tests TAGADA
sur projet CO-LocATION
(2021)

Control expression		FAILED: plot_gene_expression.sh \ reference_genes_TPM.tsv \ reference_genes_counts.tsv \ metadata.tsv Name .
Detect long non-coding RNAs		
Préparation du GTF	stringtie --merge -G /.../annotation/.gtf -p 8 -o /path/to /stringtie_merged.gtf /path/to/assembly_GTF_list.txt	
FEELnc filter	FEELnc_filter.pl -i /path/to/stringtie_merged.gtf -a /path/to/annotation/.gtf --biotype transcript_biotype=protein_coding --monoex=1 --size=200 -o FEELnc_filter.log --proc=12 > /path/to/filter/candidate_lncRNA.gtf;	ABORTED : FEELnc_filter.pl --mRNAfile Sus_scrofa.Sscrofa11.1.103.gtf \ --infile assembly.gff \ --biotype transcript_biotype=protein_coding \ > candidate_transcripts.gtf
FEElnc codprot	FEELnc_codpot.pl -i /path/to/filter/candidate_lncRNA.gtf -a /path/to /annotation/.gtf -b transcript_biotype=protein_coding -b transcript_status=KNOWN -g /path/to/genome/toplevel.fa --mode=shuffle --spethres=0.98,0.98	ABORTED : FEELnc_codpot.pl --genome Sus_scrofa.Sscrofa11.1.dna.toplevel.fa \ --mRNAfile Sus_scrofa.Sscrofa11.1.103.gtf \ --infile candidate_transcripts.gtf \ --biotype transcript_biotype=protein_coding \ --numtx 5000,5000 \ --kmer 1,2,3,6,9,12 \ --outdir . \ --outname exons \ --mode shuffle \
FEELnc classifier	FEELnc_classifier.pl -i /path/to/feelnc_codpot_out/intergenic /feelnc_codpot_out/lst_intergenic_lncRNA_noORF.gtf -a /path/to /annotation/.gtf > lncRNA_intergenic_classes.txt	ABORTED: FEELnc_classifier.pl --mrna Sus_scrofa.Sscrofa11.1.103.gtf \ --lncrna exons.lncRNA.gtf \ > lncRNA.txt
Quantify genes and transcripts		
Quantification des ARNlnc	Relancer nf-core/rnaseq options STAR RSEM	ABORTED : featureCounts -t exon \ -g gene_id \ -s 0 \ --primary \ -T 16 \ -a assembly.gff \ -o "assembly"_exons_counts.tsv \ 5-pos_GGCTAC_L002.bam 4-pos_GGCTAC_L002.bam
Control elements		ABORTED : detected_elements_sumstats.sh \ Sus_scrofa.Sscrofa11.1.103.gtf \ assembly.gff \ assembly_transcripts_TPM.tsv \ assembly_genes_TPM.tsv

https://partage-fichiers.inra.fr/perso/smaman/Documents/COLOCATION/comparaison_nfcore_tagada.xlsx?Web=1

Tests TAGADA
sur projet lncTrout

Indexation Timeout / Ressources

```
smaman@genologin2 /work/project/sigenae/sarah/lncTrout/TAGADA $ more TAGADA.sh
#!/bin/bash

#SBATCH -J lnc_TAGADA
#SBATCH --mem 80G
#SBATCH -o /work/project/sigenae/sarah/lncTrout/TAGADA/output.out
#SBATCH -e /work/project/sigenae/sarah/lncTrout/TAGADA/error.out
#SBATCH --chdir /work/project/sigenae/sarah/lncTrout/TAGADA/

pipelines=/work/project/sigenae/sarah/lncTrout/TAGADA/
containers=/usr/local/bioinfo/src/NextflowWorkflows/singularity-img/

module load bioinfo/Nextflow-v21.04.1
export NXF_ASSETS="$pipelines"
export NXF_SINGULARITY_CACHEDIR="$containers"

module load system/singularity-3.7.3
export SINGULARITY_PULLFOLDER="$containers"
export SINGULARITY_CACHEDIR="$containers"
export SINGULARITY_TMPDIR="$containers"

wget https://gist.githubusercontent.com/chbk/2f9122538c5db222a822cfade05f81f4/raw/63e0442914767a255f95b7d70ba2efa6afbadce0/nextflow-run
chmod +x nextflow-run

echo 'singularity.runOptions = "-B /bank -B /work2 -B /work -B /save -B /home"' > nextflow.config

./nextflow-run FAANG/analysis-TAGADA \
  --revision 1.0.2 \
  --profile slurm singularity \
  --config nextflow.config \
  --output /work/project/sigenae/sarah/lncTrout/TAGADA/ \
  --reads /work/project/sigenae/sarah/lncTrout/FASTQ/*.gz \
  --annotation /work/project/sigenae/sarah/YeastTrout/nextflow27082021/annotation/Oncorhynchus_mykiss.Omyk_1.0.104.gtf \
  --genome /work/project/sigenae/sarah/YeastTrout/nextflow27082021/genome/Oncorhynchus_mykiss.Omyk_1.0.dna.toplevel.fa
```

ATTENTION: Pour information, l'option --workdir de la commande sbatch a été remplacée par la commande --chdir.

Lancement du job:

```
smaman@genologin2 /work/project/sigenae/sarah/lncTrout/TAGADA $ sbatch TAGADA.sh
Submitted batch job  30118052  / TIMEOUT - INDEXATION TROP LONGUE  ==> DONC A SORTIR ?
sigenae@genologin2 /work/project/sigenae/sarah/lncTrout $ seff 30118052
Job ID: 30118052
Cluster: genobull
User/Group: smaman/SIGENAE
State: TIMEOUT (exit code 0)
Cores: 1
CPU Utilized: 00:03:17
CPU Efficiency: 0.06% of 4-00:00:21 core-walltime
Job Wall-clock time: 4-00:00:21
Memory Utilized: 1.30 GB
Memory Efficiency: 1.63% of 80.00 GB
```

Tests TAGADA
sur projet MONOPOLY

- Tests en Avril 2021 : reports de bugs
- Run avril 2022 sur données bovins

$ wc -l reference_genes_counts.tsv

27 608

$ wc -l novel_genes_counts.tsv

33 645

$ grep -i strg novel_genes_counts.tsv | wc -l

22 148

```
smaman@genologin1 /work/project/sigenae/sarah/Project_MONOPOLY.222/TAGADA_test6 $ ls RESULTATS/*
RESULTATS/alignment:
SRR13949332.bam  SRR13949334.bam  SRR13949336.bam  SRR13949338.bam  SRR13949340.bam  SRR13949343.bam  SRR13949345.bam  SRR13949347.bam  SRR13949349.bam  SRR13949351.bam
SRR13949333.bam  SRR13949335.bam  SRR13949337.bam  SRR13949339.bam  SRR13949341.bam  SRR13949344.bam  SRR13949346.bam  SRR13949348.bam  SRR13949350.bam  SRR13949352.bam

RESULTATS/annotation:
lnc_classification  novel.gtf

RESULTATS/control:
annotation  exons  expression  lnc  multiqc_report.html  splicing

RESULTATS/coverage:
SRR13949332.bed  SRR13949334.bed  SRR13949336.bed  SRR13949338.bed  SRR13949340.bed  SRR13949343.bed  SRR13949345.bed  SRR13949347.bed  SRR13949349.bed  SRR13949351.bed
SRR13949333.bed  SRR13949335.bed  SRR13949337.bed  SRR13949339.bed  SRR13949341.bed  SRR13949344.bed  SRR13949346.bed  SRR13949348.bed  SRR13949350.bed  SRR13949352.bed

RESULTATS/index:
chrLength.txt  chrNameLength.txt  chrName.txt  chrStart.txt  exonGeTrInfo.tab  exonInfo.tab  geneInfo.tab  Genome  genomeParameters.txt  Log.out  SA  SAindex  sjdbInfo.txt  sjdbList.fromGTF.out.tab  sjdbList.out.tab  transcriptInfo.tab

RESULTATS/quantification:
novel_genes_counts.tsv  novel_genes_TPM.tsv  novel_transcripts_counts.tsv  novel_transcripts_TPM.tsv  reference_genes_counts.tsv  reference_genes_TPM.tsv  reference_transcripts_counts.tsv  reference_transcripts_TPM.tsv
```