ChAR-seq pipeline essentials

Where to find key files and what they contain

CL – 02/07/2020

Where do we run the ChAR-seq pipeline?

Usually, for each pooled library and analysis setup, we create a folder which will Be the root folder for the pipeline:

```
PIPE_ROOT = {some_path_of_choice}/{library_name}/{analysis_setup}/
```

Example: novaseq2 data analysis

- library_name = *novchar2*
- analysis_setup =
  - Full dataset : *NOVAseq_12-02-2019*
  - First 10M reads *: NOVAseq_12-02-2019_10mReads*

Root level organization of the pipeline files

$PIPE_ROOT/pipeline.smk	→ snakemake rules
pipeline_config.yaml	→ pipeline configuration
samples_def.yaml	→ definition of samples to be processed
$PIPE_ROOT/data	→ where the processed data will be generated
$PIPE_ROOT/data/sample1/	→ individual samples (sampleID) go in their own folder
sample2/	
sample3/	
....	

Processed data for a given sample live in

```
SAMPLE_ROOT = $PIPE_ROOT/data/{sampleID}/
```

Pipeline outputs organization

All the paths from here on are relative to the sample root folder $SAMPLE_ROOT

- raw/ → symlink to raw sequencing data stored on $OAK
- chimeras/ → deduped and trimmed reads : .chimera.fastq.gz files
- Split_chimeras/ → .rna.fastq.gz and .dna.fastq.gz file (see next) files

- alignments/ → where the .rna.bam and .dna.bam files live (see next)
- pairs/ → where the contact files files live (see next)

For downstream analysis, pairs/ is essentially the only place we should look into

DNA and RNA fastq files

- RNA-to-DNA matching is maintained in all rna.fastq and dna.fastq file is the same folder (line N in dna.fastq is the same read as line N in rna.fastq)

- DNA fq lives in `split_chimeras/{mates_pairing_mode}/{filter}/dna.fastq.gz`

How the paired end read are converted to single end
- SE_merge_pear → uses PEAR to merge

Pre-alignment filtering
- Unfiltered → no filtering
- Long.decon →
  - RNA and DNA both >15bp
  - Reads where RNA aligns to rRNA removed
  - This is the one going into aligner

- Matching RNA fq lives in `split_chimeras/{mates_pairing_mode}/{filter}/rna.fastq.gz`

DNA and RNA alignment BAM files

- DNA lives in `alignments/dna/{dna_alignment_mode}/dna.bam`

Aligner configuration :
- Bowtie_hg38

- RNA lives in `alignments/rna/{rna_alignment_mode}/bytype/rna.{annotation_type}.bam`

Aligner configuration
- star_gencodeV29

Type of annotations compatible with alignment, in order of priority (see next slide)
- exons
- introns
- novel (Introns with splicing)
- intergenic
- toomany
- unmapped

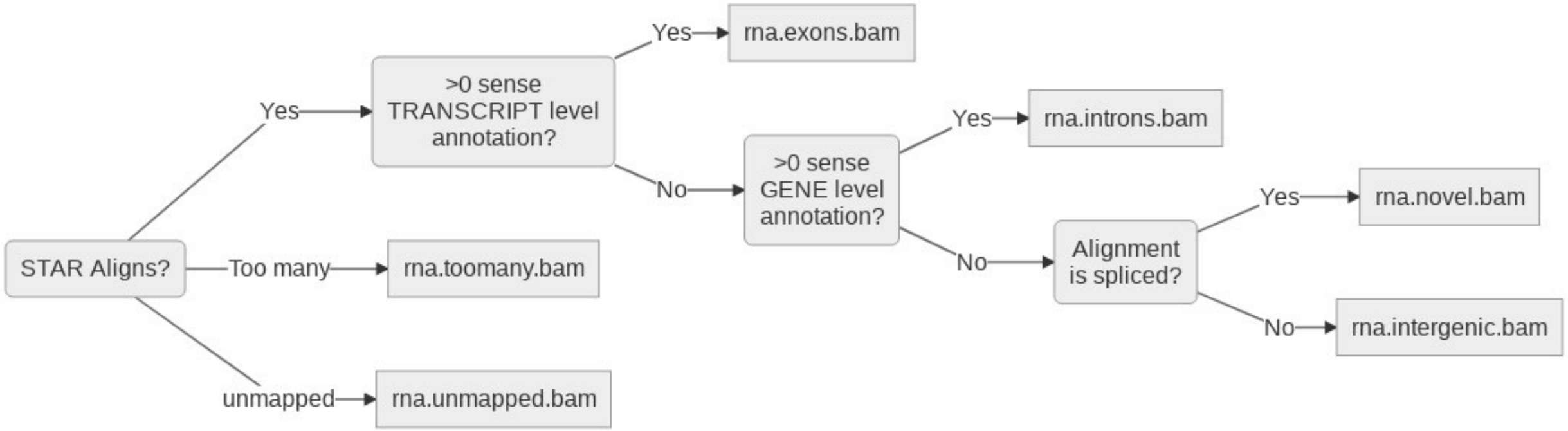
- One single alignement per readID.

- What about multimapping readas?
  - DNA: randomly selected

RNA annotation rules

Recall: RNA bam are `./alignments/rna/{rna_alignment_mode}/bytype /rna.{annotation_type}.bam`

Annotation_type decision tree



rna.all.bam = exons + introns + novel + intergenic

Contact files

- RNA-DNA contacts are represented by either .pairs or .bed files (see next slide)
- 1 single entry per readID, matching each entry in rna.{pair_type}.bam with the corresponding entry in dna.bam
- UNFILTERED contact files (include RNA/DNA multimappers) live in

```
./pairs/{pairing_mode}/{annotation_type}/
```

Defines combination of RNA and DNA aligner configurations
- gencondeV29_hg38

Type of RNA annotations
- exons
- introns
- novel (Introns with splicing)
- intergenic
- all

- FILTERED live in

```
./pairs/{pairing_mode}/{annotation_type}/filtered/DNAq15-blacklisted_RNAunq/
```

Filtering

FILTERED contact files : an entry must satisfy the following criteria to be kept

1. DNA alignment has Q>15
   - Multimappers and low quality mappers are removed

2. RNA alignment is compatible with a single gene (we call these "unambiguous" RNA annotation)
   - Intergenic alignments are removed
   - Alignments compatible with multiple transcripts of the same gene are kept
   - Multimappers with all mapping loci in the same gene are kept
   - Multimappers with loci belonging to different genes are removed
   - Annotation compatibility is always stranded
     - Alignment that is compatible with a single gene but on the wrong strand is labeled intergenic and removed

3. DNA does not align to a blacklisted genomic regions
   - This filter does not apply to RNA as we can sort that out using gene name

4. DNA does not map to chromosome Y (female cells), chrM, unassembled or unplaced scaffolds
   - This filter does not apply to RNA as we can sort that out using gene name

Contact file formats

These 3 contact files format store ~ the same information but organized differently for different usage cases

`rd.indexed.pairs.gz` → indexed .pairs file
- efficient 2D queries using **pairix**
- then convert to rna-major or dna-major bed file for interval arithmetic on RNA or DNA side

`dna.bed.gz` → indexed dna-major bed file (coordinates of DNA in bed format)
- efficient query of RNAs at a given locus using **tabix**
- make DNA side tracks using bedtools coverage, etc...
- use with bedtools for interval arithmetic

`rna.bed.gz` → indexed rna-major bed file (coordinates of RNA in bed format)
- efficient query of the DNA targets of a given RNA using **tabix**
- make RNA side tracks using bedtools coverage, etc...
- use with bedtools for interval arithmetic

Fields definition in contact files

FIELD	Example	.pairs	.dna.bed	.rna.bed	Field description
QNAME	"A00564:124:HL3LWDSXX:3:1101:10004:10802"	1	4	4	Query template name.
RNAME_RNA	R_chrX	2	7	1	Name of the reference sequence in genomic space to which the RNA query maps.
POS_RNA_START	136842010	3	8	2	Mapping position of the RNA query on the reference sequence in genomic space.
POS_RNA_STOP	136842011		9	3	
RNAME_DNA	chr8	4	1	1	Name of the reference sequence in genomic space to which the DNA query maps
POS_DNA_START	85298996	5	2	8	Mapping position of the DNA query on the reference sequence in genomic space.
POS_DNA_STOP	85298997		3	9	
STRAND_RNA	-	6	11	6	Which strand of the genomic reference to which the RNA query maps
STRAND_DNA	+	7	6	11	Which strand of the genomic reference to which the DNA query maps
MAPQ_RNA	255	8	10	5	Score of the reported alignment for the RNA query
MAPQ_DNA	32	9	5	10	Score of the reported alignment for the DNA query
FLAG_RNA	16	10			Flag of the reported alignment for the RNA query
FLAG_DNA	0	11			Flag of the reported alignment for the DNA query
BRIDGEGAP_RNA	9	12			Number of bp on the RNA side of the bridge that are not part of the reported alignment for the RNA query
BRIDGEGAP_DNA	0	13			Number of bp on the DNA side of the bridge that are not part of the reported alignment of the DNA query
TLEN_RNA	95	14			Number of bp of the reported alignment for the RNA query
TLEN_DNA	29	15			Number of bp of the reported alignment for the DNA query
NH_RNA	1	16			Number of reported alignments that contains the RNA query in the current record
NH_DNA	1	17			Number of reported alignments that contains the DNA query in the current record. This number is always 1 with Bowtie2
ANNOT_RNAME_RNA	ENST00000435597.1	18	12	12	Name of the reference sequence in annotation space to which the RNA query maps. When the annotation space is a transcriptome, this is the name of the transcript to which the RNA query maps
ANNOT_POS_RNA	1843	19	13	13	Mapping position of the RNA query on the reference sequence in annotation space. When the annotation space is a transcriptome, this is the position from the 5'end of the transcript of the mapped segment.
ANNOT_ENGLISH_NAME_RNA	AL683813.1	20	14	14	Meaninful "english" name of the reference sequence in annotation space to which the query maps. When the annotation space is a transcriptome, this is the name of the gene for this annotation.
ANNOT_TYPE_RNA	lincRNA	21	15	15	Type of annotation to which the RNA query maps
gS_RNA	1	22			Number of reported alignments that contains the RNA query in the current record and that are compatible with an annotation. When the annotation space is a transcriptome, this is the number of distinct genomic alignments which are compatible with at least one transcript.
aS_RNA	1	23			Total number of annotations compatible with one or more of reported alignments that contains the RNA query in the current record. When the annotation space is a transcriptome, this is the number of transcripts the RNA query maps to (counting only those in the sense direction)
ai_RNA	1	24			Transcript level ambivalence group ID. A value of 1 indicates that the RNA query maps to a single transcript. A positive value larger than one indicates that the RNA query maps to mulitple transcripts. When the RNA query also maps to a genomic locus which is not compatible with any annotation, the reported number is multplied by -1
aI_RNA	1	25	16	16	Gene level ambivalence group ID. A value of 1 indicates that the RNA query maps to a single gene (but possibily multiple transcripts for the same gene) A positive value larger than one indicates that the RNA query maps to mulitple genes When the RNA query also maps to a genomic locus which is not compatible with any annotation, the reported number is multplied by -1
ANNOT_GENEID_RNA	ENSG00000232611.1	26	17	17	ID of the GENE corresponding to annotation of field ANNOT_RNAME_RNA
IS_CIS_CHR	0		18	18	1 if interation in CIS, 0 if interaction in TRANS
FLIGHT	-51543014		19	19	Travel distance RNA-DNA in bp. Equals POS_DNA_START-POS_RNA_START. Only useful when the interaction is in CIS.