Quality control and characterization of LR transcriptomes

Francisco J. Pardo-Palacios
Lorena de la Fuente

SQANTI3

Genomics of Gene Expression Lab

Hands On session

Hands-on Outline

Exercises:

1. Perform QC for a subset of isoforms of PacBio's melanoma data set (chr15 isoforms) with the minimum information.
2. Incorporate to the analysis orthogonal information
   - What to do with matching short-read data?
   - Include CAGE and polyA information
3. Run SQANTI3 filter-by-rules

SQANTI3 pre-requisites:

- Build SQANTI3 conda environment
- Download gtfToGenePred
- Install cDNA_Cupcake

Instructions in https://github.com/ConesaLab/SQANTI3

Hands-on Outline

Exercises:

1. Perform QC for a subset of isoforms of PacBio's melanoma data set (chr15 isoforms) with the minimum information.
2. Incorporate to the analysis orthogonal information
   - What to do with matching short-read data?
   - Include CAGE and polyA information
3. Run SQANTI3 filter-by-rules

SQANTI3 pre-requisites:

Build SQANTI3 conda environment

Download gtfToGenePred

Install cDNA_Cupcake

Instructions in https://github.com/ConesaLab/SQANTI3

Exercise #1

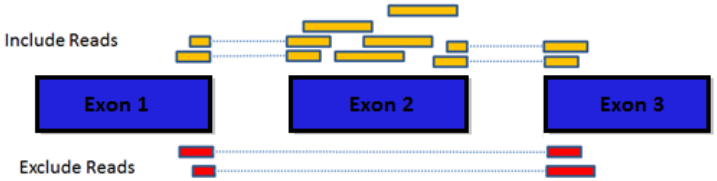
1. Locate resources on the cluster

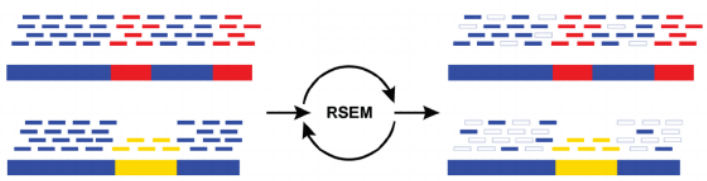
```
srun -t 02:00:00 -c 4 -n 1 --mem 2000 --partition production --account isoseq_workshop --
reservation isoseq_workshop  --pty /bin/bash
```

2. Activate SQANTI3 module
   - Activate cDNA_cupcake environment
     - Set PYHTONPATH
   - Activate SQANTI3_handsOn environment
3. Set symbolic links to data
4. Run SQANTI3

Exercise #2

- What extra information do we have about our transcriptome?
  - FL counts obtained after running IsoSeq3 + cDNA_Cupcake
  - CAGE peak data
  - List of polyA motifs
  - RNA-Seq data from Illumina sequencing
    - Are Splice-Junctions supported by Short-reads?
    - Can we quantify our transcriptome using Short-reads?





Exercise #2

Generated from *classification* and *junctions* files

- How many isoforms are already known?

Unique Genes: 409
Unique Isoforms: 2080

Gene classification

Category	# Genes
Annotated Genes	392
Novel Genes	17

Splice Junction Classification

Category	# SJs	Percent
Known canonical	4537	90.50
Known Non-canonical	1	0.02
Novel canonical	380	7.58
Novel Non-canonical	95	1.90

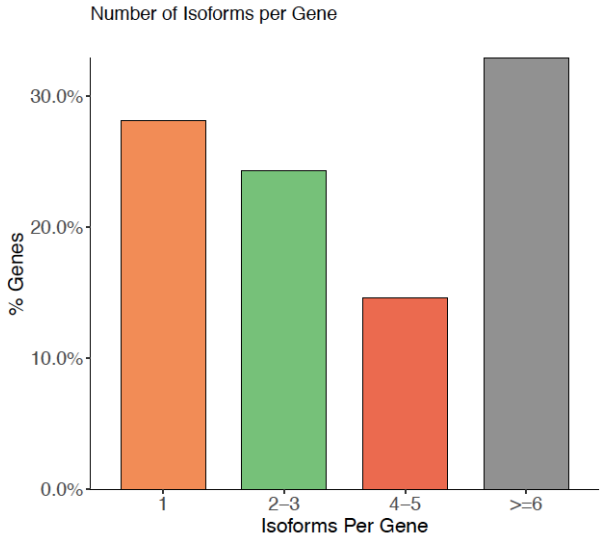
Characterization of transcripts based on splice junctions

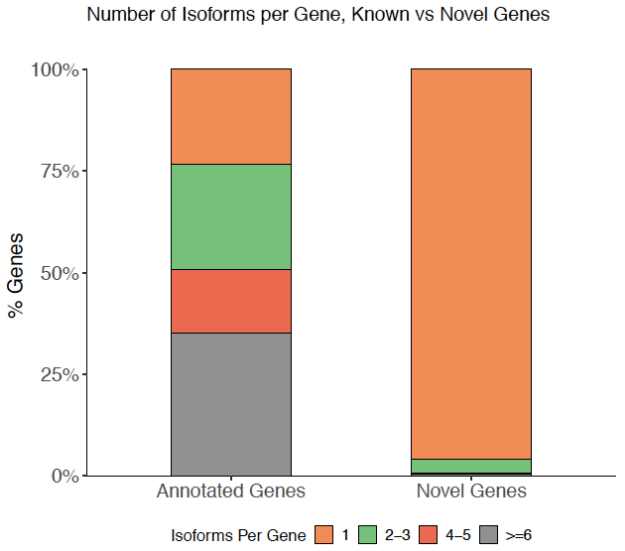
Category	# Isoforms
FSM	909
ISM	219
NIC	565
NNC	358
Genic Genomic	9
Antisense	7
Fusion	3
Intergenic	10
Genic Intron	0

Exercise #2

Generated from *classification* and *junctions* files

- How many isoforms are already known?
- How many isoforms per gene were found?

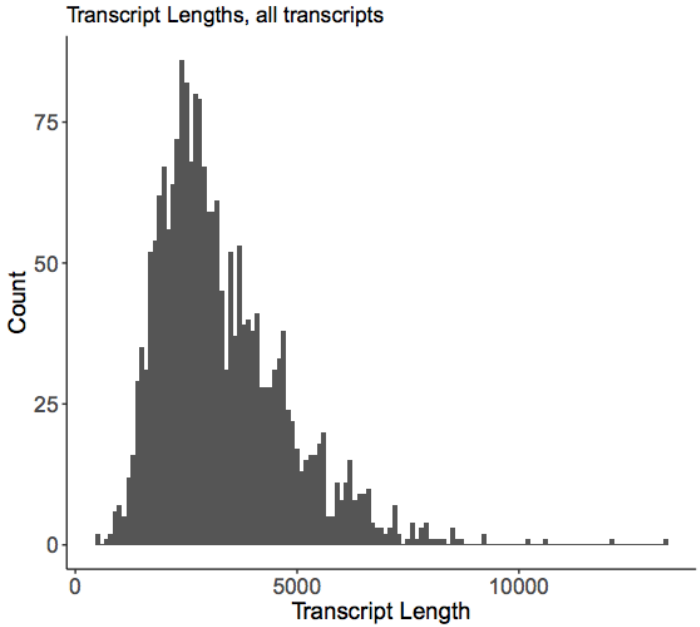




Exercise #2

Generated from *classification* and *junctions* files

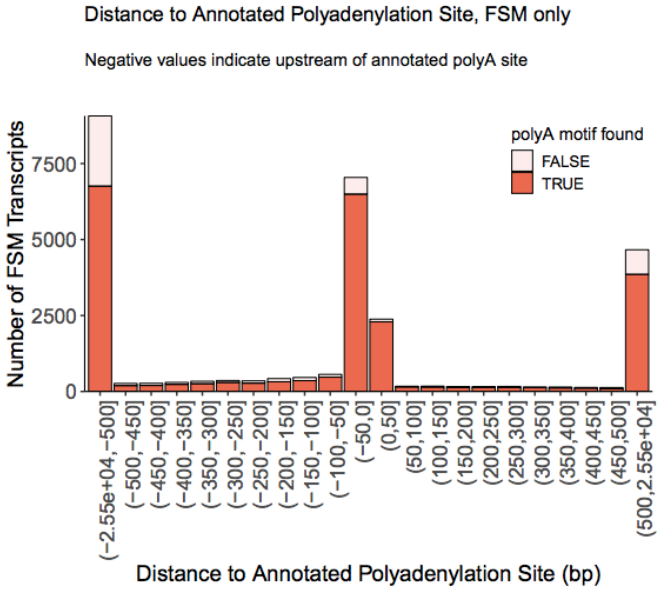
- How many isoforms are already known?
- How many isoforms per gene were found?
- How long are the transcripts detected?



Exercise #2

Generated from *classification* and *junctions* files

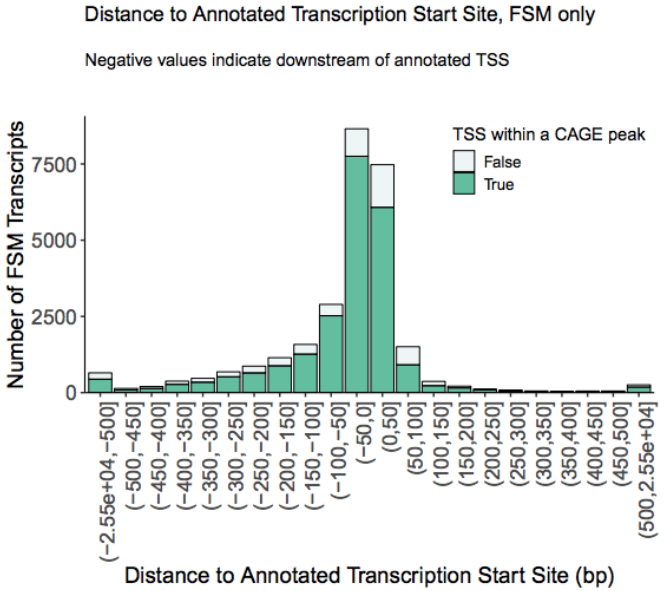
- How many isoforms are already known?
- How many isoforms per gene were found?
- How long are the transcripts detected?
- Evaluation of detected TSS and TTS…
  - Are FSM capturing known TSS and TTS?



Exercise #2

Generated from *classification* and *junctions* files

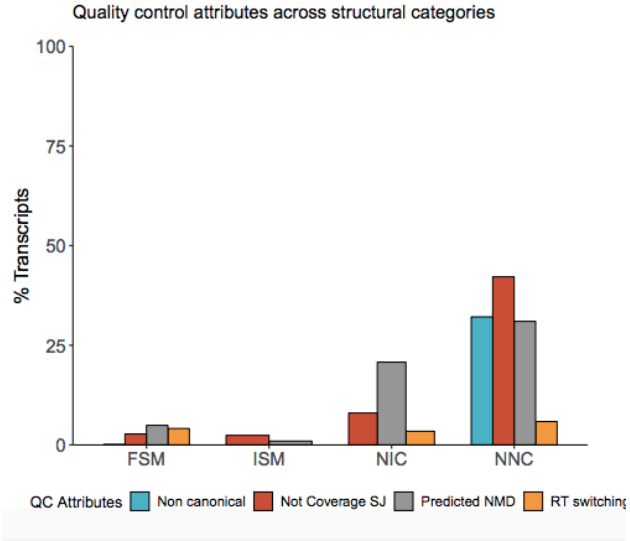
- How many isoforms are already known?
- How many isoforms per gene were found?
- How long are the transcripts detected?
- Evaluation of detected TSS and TTS…
  - Are FSM capturing known TSS and TTS?



Exercise #2

Generated from *classification* and *junctions* files

- Can we trust all the isoforms described?
  - Novel isoforms of known genes



Exercise #3

Once the isoforms have been classified and evaluated, we can use the rules filter to remove possible artifacts.

```
python sqanti3_RulesFilter.py chr15_classification.txt chr15_corrected.fasta
        chr15_corrected.gtf -c 3 –a 0.6
```

RULES:

- FSM isoforms pass the filter, unless intrapriming
- For the rest of categories, isoform will be discarded because of:
  - Intrapriming
  - RT-switching SJ
  - Non-canonical SJ w/o SR support below –c threshold (default: 3)

Functional annotation at the isoform-level with isoAnnot

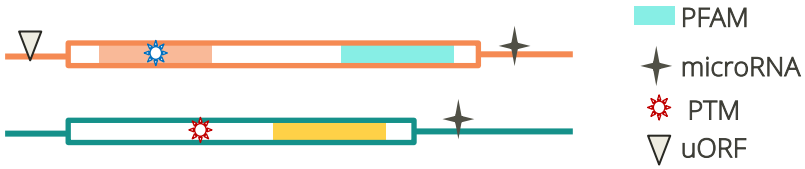
Francisco J. Pardo-Palacios

Genomics of Gene Expression Lab

- Framework for the functional annotation of isoforms.
- Collection of scripts that need to be adapted to each organism.
- Obtain a GFF3 file compatible with tappAS.

Curated full-length transcriptome

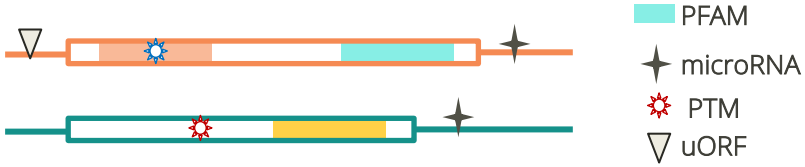
Annotated full-length transcriptome



- Framework for the functional annotation of isoforms.
- Collection of scripts that need to be adapted to each organism.
- Obtain a GFF3 file compatible with tappAS.

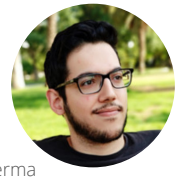
Curated full-length transcriptome

Annotated full-length transcriptome



Already annotated reference transcriptomes

Lorena de la Fuente



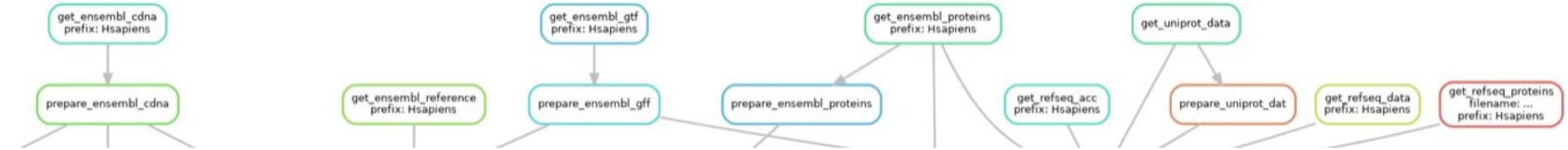
Alberto Lerma

IsoAnnot: Snakemake scheme

- Get annotation data from public databases (Ensembl, Refseq, others...)

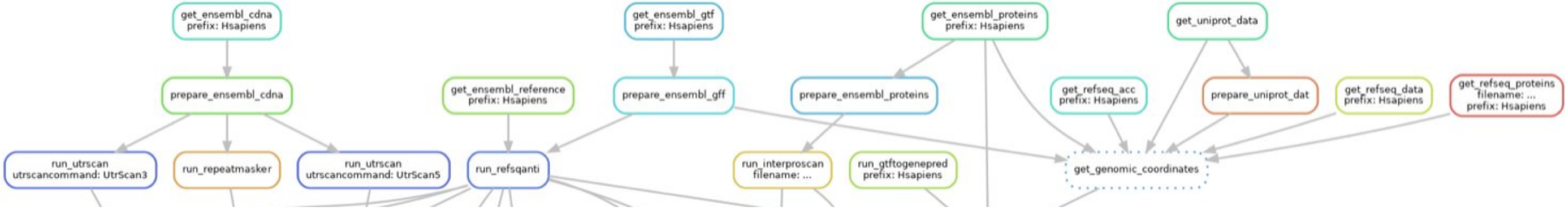
isoAnnot: Snakemake scheme

- Prepare data for function prediction and extract information from them.



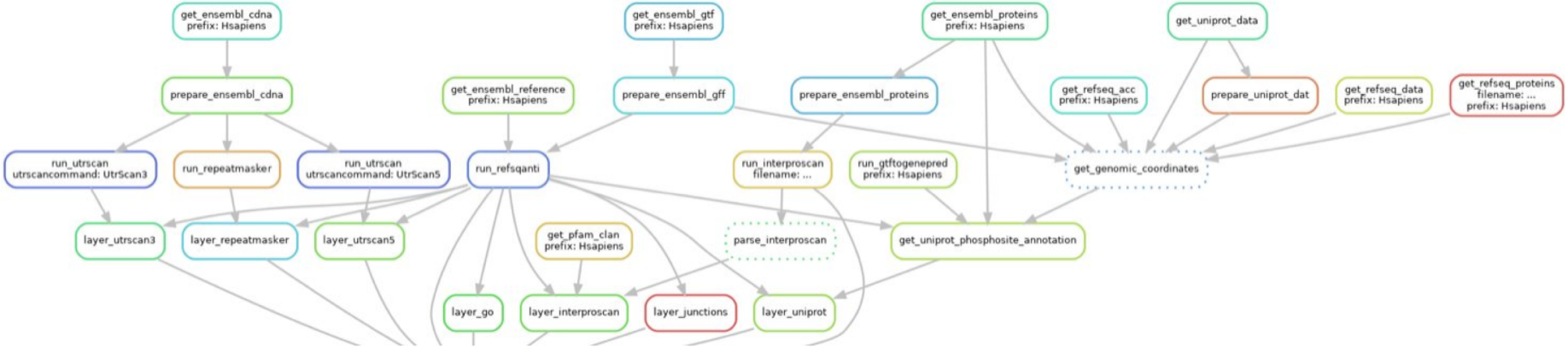
isoAnnot: Snakemake scheme

- Run annotation algorithms (UTRscan, RepeatMasker, InterproScan, SQANTI...)



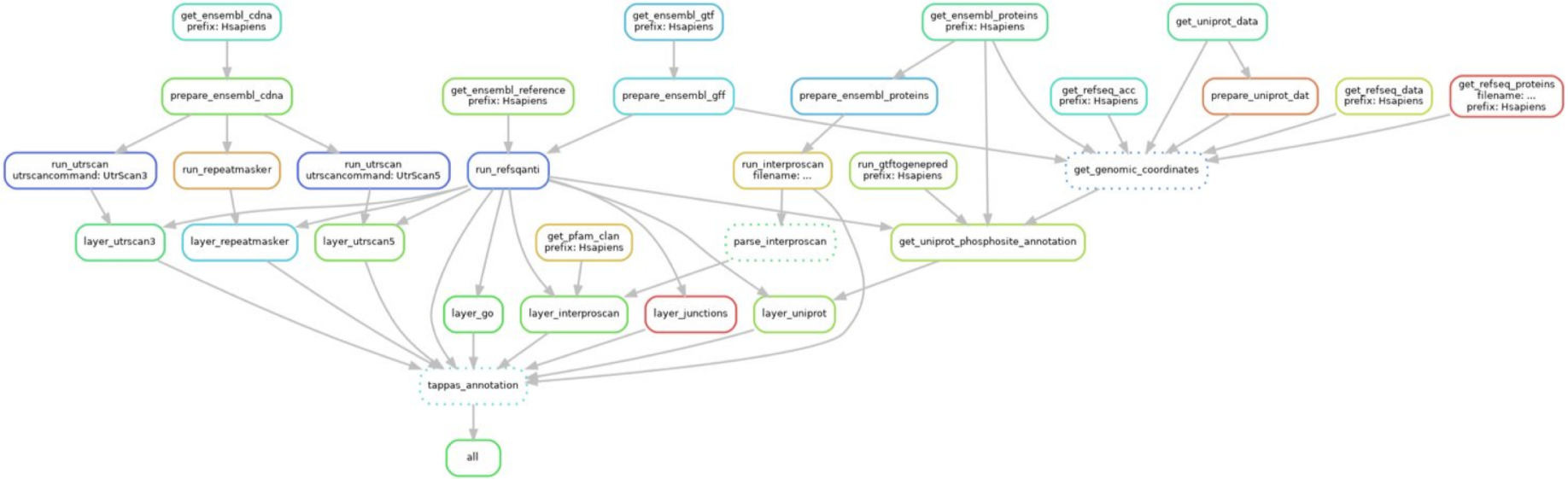
isoAnnot: Snakemake scheme

- Add and integrate annotation layers (depending of organism)



isoAnnot: Snakemake scheme

- Create a gff3 file with isoform-level functional labels



- Current status:
  - Running IsoAnnot is laborious and requires computational skills. Contact us if you are interested in annotate your non-model species.
- By-pass:
  - IsoAnnot Lite
    - Usage of our pre-made annotation file for selected species. So far, available:
    - Transfer annotated features to the long-read defined transcripts
    - Return a GFF3 file that can be directly loaded into tappAS

IsoAnnot Lite

- Current status:
  - Running IsoAnnot is laborious and requires computational skills. Contact us if you are interested in annotate your non-model species.
- By-pass:
  - IsoAnnot Lite
    - Usage of our pre-made annotation file for selected species. So far, available:
    - Transfer annotated features to the long-read defined transcripts
    - Return a GFF3 file that can be directly loaded into tappAS

IsoAnnot Lite → SQANTI3

Exercise #4

Use isoAnnot options to create your own tappAS-compatible GFF3 with functional annotations at the isoform level.

```
python sqanti3_qc.py chr15_seqs.fasta referenceTranscriptome.gtf referenceGenome.fasta \
        -o chr15 --fl_count FL_counts.txt --expression expression.txt \
        -c SR_support.tsv --cage_peak CAGE_peaks.bed --polyA_motif polyA.txt -n4 \
        --isoAnnotLite --gff3 tappAS_annotation.gff3
```

Exercise #4

Use isoAnnot options to create your own tappAS-compatible GFF3 with functional annotations at the isoform level.

```
python sqanti3_qc.py chr15_seqs.fasta referenceTranscriptome.gtf referenceGenome.fasta \
        -o chr15 --fl_count FL_counts.txt --expression expression.txt \
        -c SR_support.tsv --cage_peak CAGE_peaks.bed --polyA_motif polyA.txt -n4 \
        --isoAnnotLite --gff3 tappAS_annotation.gff3
```

It must have the same transcript IDs than the GFF3 annotation file

Generate a tappAS-compatible GFF3

Annotation file from isoAnnotLite webpage* used for "copy-paste" functional labels

* http://app.tappas.org/resources/downloads/gffs/

Exercise #4

Example: PB.9618.5

- Information in GTF after running SQANTI3

```
5	Homo_sapiens	transcript	172043618	172188357	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172043618	172045022	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172052929	172053042	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172054569	172054694	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172055588	172055776	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172057349	172057473	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172061139	172061268	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172064720	172064812	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172082326	172082505	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172082961	172083084	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172090232	172090362	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172093412	172093960	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172096426	172096560	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172105656	172105737	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172106620	172106814	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172107780	172107852	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172117481	172117630	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172127373	172127421	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172156624	172156788	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
5	Homo_sapiens	exon	172187887	172188357	.	-	.	transcript_id "PB.9618.5"; gene_id "PB.9618.5";
```

Exercise #4

Example: PB.9618.5

- Information in GFF3 after running SQANTI3+isoAnnotLite

```
PB.9618.5	tappAS	transcript	1	4486	.	-	.	ID=ENST00000176763; primary_class=full-splice_match; PosType=T
PB.9618.5	tappAS	gene	1	4486	.	-	.	ID=ENSG00000072786; Name=ENSG00000072786; Desc=ENSG00000072786; PosType=T
PB.9618.5	tappAS	CDS	316	3222	.	-	.	ID=Protein_PB.9618.5; Name=Protein_PB.9618.5; Desc=Protein_PB.9618.5; PosType=T
PB.9618.5	UTRsite	3UTRmotif	4066	4073	.	-	.	ID=U0023; Name=K-BOX; Desc=K-Box; PosType=T
PB.9618.5	UTRsite	3UTRmotif	3119	3123	.	-	.	ID=U0035; Name=MBE; Desc=Musashi binding element; PosType=T
PB.9618.5	UTRsite	3UTRmotif	4386	4390	.	-	.	ID=U0035; Name=MBE; Desc=Musashi binding element; PosType=T
PB.9618.5	UTRsite	PAS	4463	4486	.	-	.	ID=U0043; Name=PAS; Desc=Polyadenylation Signal; PosType=T
PB.9618.5	RepeatMasker	repeat	253	278	.	-	.	ID=Simple_repeat; Name=Simple_repeat; Desc=(TCC)n; PosType=T
PB.9618.5	RepeatMasker	repeat	296	323	.	-	.	ID=Simple_repeat; Name=Simple_repeat; Desc=(AGCCCG)n; PosType=T
PB.9618.5	RepeatMasker	repeat	3208	3468	.	-	.	ID=SINE/Alu; Name=SINE/Alu; Desc=AluSx; PosType=T
PB.9618.5	RepeatMasker	repeat	3469	3587	.	-	.	ID=SINE/Alu; Name=SINE/Alu; Desc=FLAM_A; PosType=T
PB.9618.5	miRWalk	miRNA_Binding	4316	4324	.	-	.	ID=hsa-miR-1324; Name=hsa-miR-1324; Desc=3UTR; PosType=T
PB.9618.5	GeneOntology	C	.	.	.	.	.	ID=GO:0005886; Name=plasma membrane; PosType=N
PB.9618.5	GeneOntology	C	.	.	.	.	.	ID=GO:0005737; Name=cytoplasm; PosType=N
PB.9618.5	GeneOntology	P	.	.	.	.	.	ID=GO:0071593; Name=lymphocyte aggregation; PosType=N
PB.9618.5	GeneOntology	F	.	.	.	.	.	ID=GO:0004672; Name=protein kinase activity; PosType=N
PB.9618.5	KEGG	pathway	.	.	.	.	.	ID=04151+2.7.11.1; Name=PI3K-Akt signaling pathway; PosType=N
PB.9618.5	KEGG	pathway	.	.	.	.	.	ID=04150+2.7.11.1; Name=mTOR signaling pathway; PosType=N
PB.9618.5	PFAM	CLAN	.	.	.	.	.	ID=CL0016; Name=PKinase; Desc=Protein kinase superfamily; PosType=N
PB.9618.5	UniProtKB/Swiss-Prot_Phosphosite	PTM	954	954	.	.	.	ID=Phosphoserine; Name=Phosphoserine; Desc=Phosphoserine_PhosphositePlus; PosType=P
PB.9618.5	UniProtKB/Swiss-Prot_Phosphosite	PTM	184	184	.	.	.	ID=PTM_other; Name=PTM_other; Desc=Ubiquitination_PhosphositePlus; PosType=P
PB.9618.5	UniProtKB/Swiss-Prot_Phosphosite	BINDING	65	65	.	.	.	ID=Binding; Name=Binding; Desc=Atp_SwissProt; PosType=P
PB.9618.5	UniProtKB/Swiss-Prot_Phosphosite	PTM	195	195	.	.	.	ID=Phosphothreonine; Name=Phosphothreonine; Desc=Phosphothreonine_PhosphositePlus; PosType=P
PB.9618.5	TranscriptAttributes	3UTR_Length	3223	4486	.	-	.	ID=3UTR_Length; Name=3UTR_Length; Desc=3UTR_Length; PosType=T
PB.9618.5	TranscriptAttributes	5UTR_Length	1	316	.	-	.	ID=5UTR_Length; Name=5UTR_Length; Desc=5UTR_Length; PosType=T
PB.9618.5	TranscriptAttributes	CDS	316	3222	.	-	.	ID=CDS; Name=CDS; Desc=CDS; PosType=T
PB.9618.5	TranscriptAttributes	polyA_Site	4486	4486	.	-	.	ID=polyA_Site; Name=polyA_Site; Desc=polyA_Site; PosType=T
```

Exercise #4

Example: PB.9618.5

- Information in GFF3 after running SQANTI3+isoAnnotLite loaded into tappAS