Xavier Grand.

HBV Transcriptomics: Nanopore long-reads analysis.

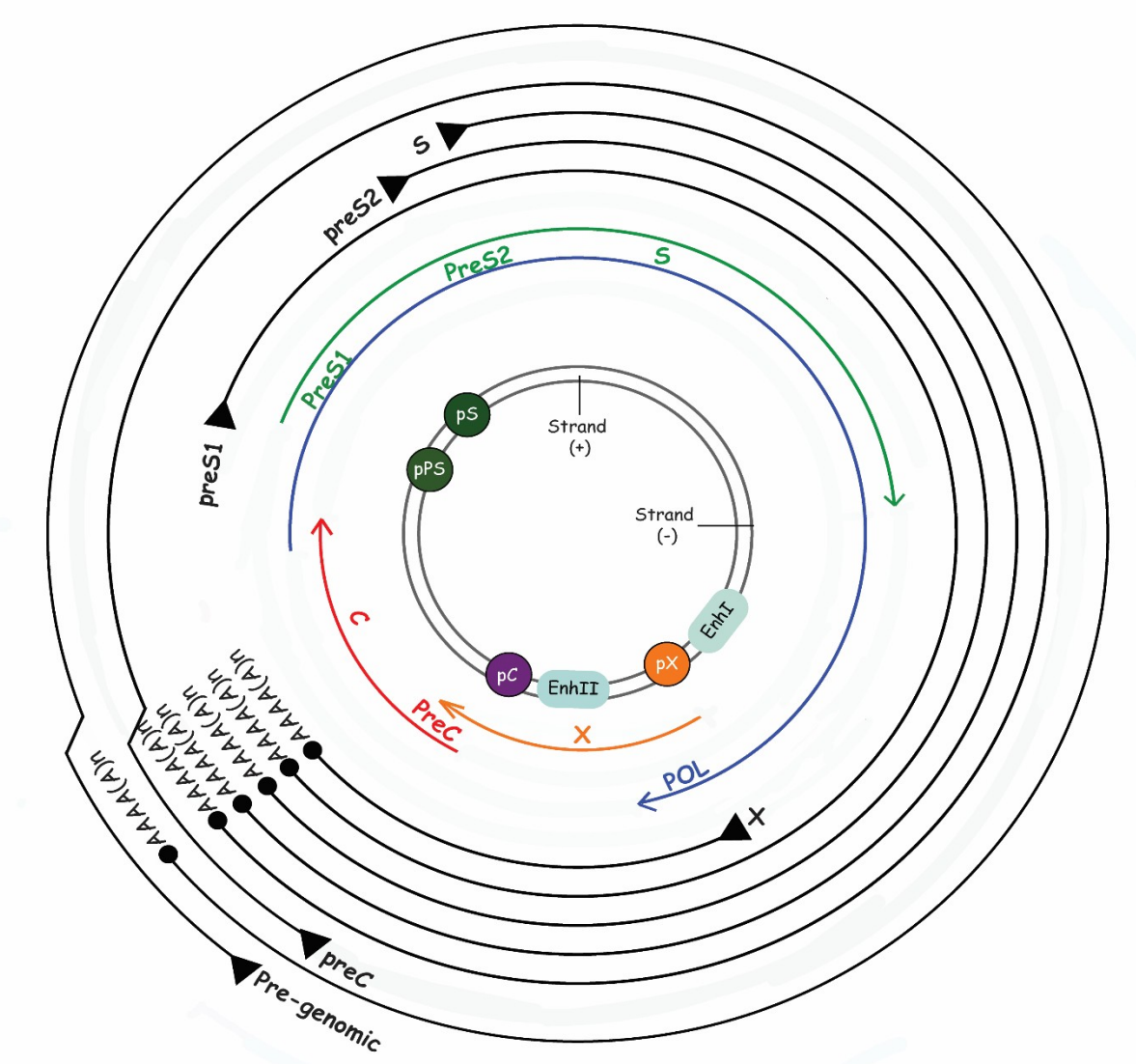
Club Bioinfo.

3rd March 2022.

CRCL CENTRE DE RECHERCHE EN CANCÉROLOGIE DE LYON
Lyon 1
UNIVERSITÉ DE LYON
ENS DE LYON
LBMC

Introduction: The biological model.

Hepatitis B Virus.

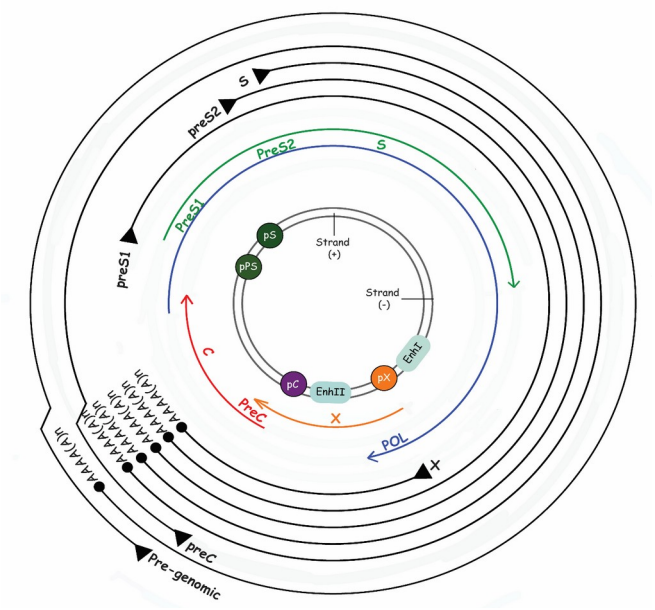


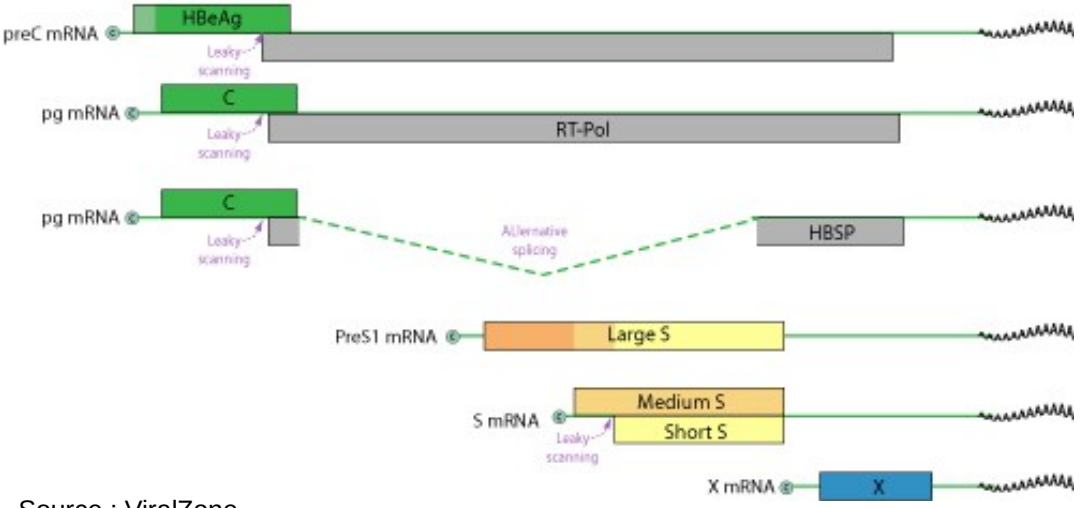
3,2kb circular genome:

_ 1 polyA site,
_ 4 Promoters,
_ 6 mRNAs,
_ 4 coded-proteins,
_ 20 Splice variants (so far),
_ Only 1 strand is coding for mRNAs.

Introduction: Overlapping ORFs.

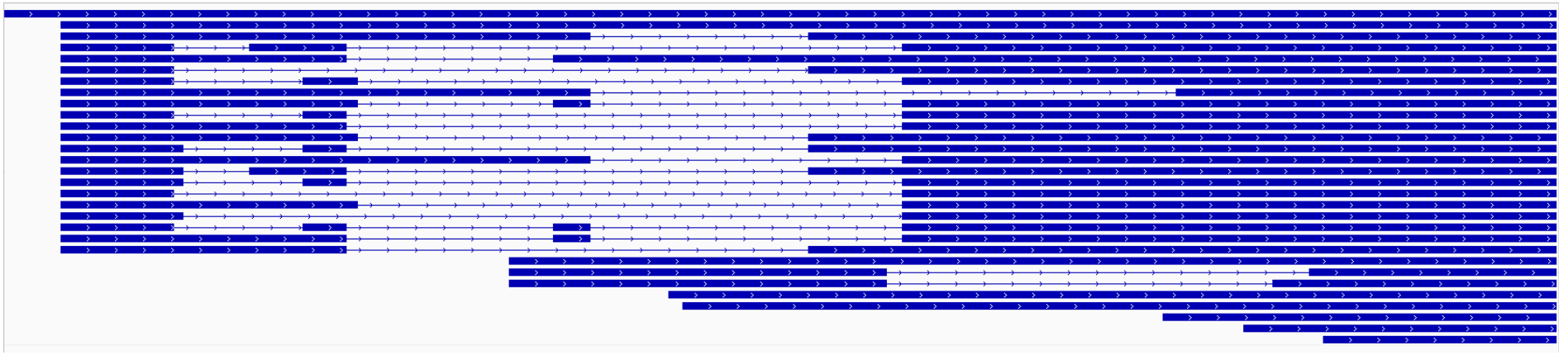
Hepatitis B Virus.





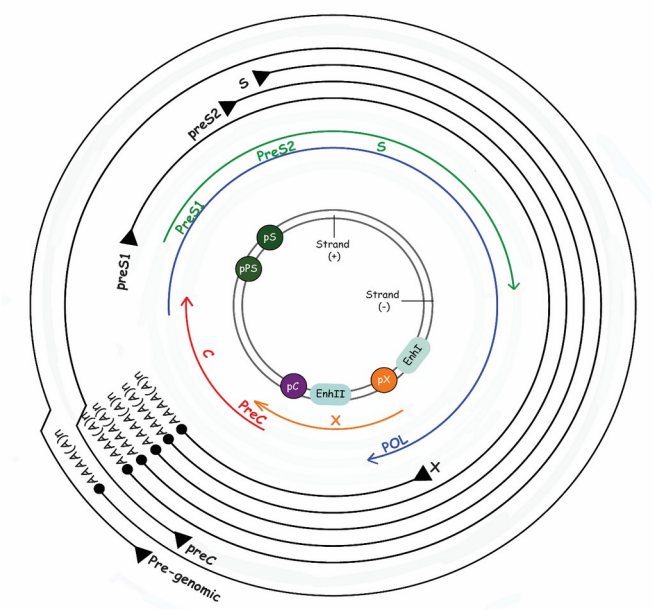
Source : ViralZone

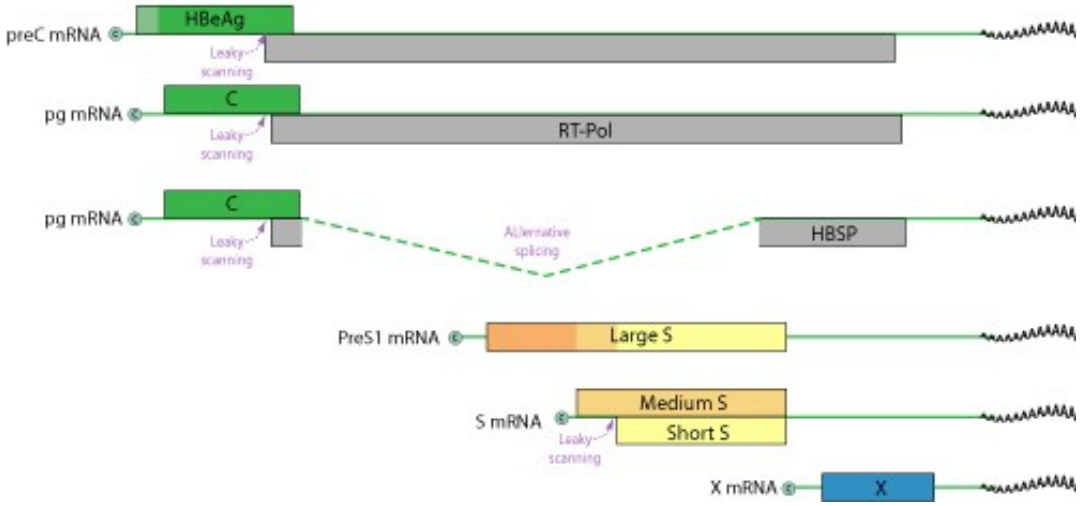
All known HBV mRNAs :



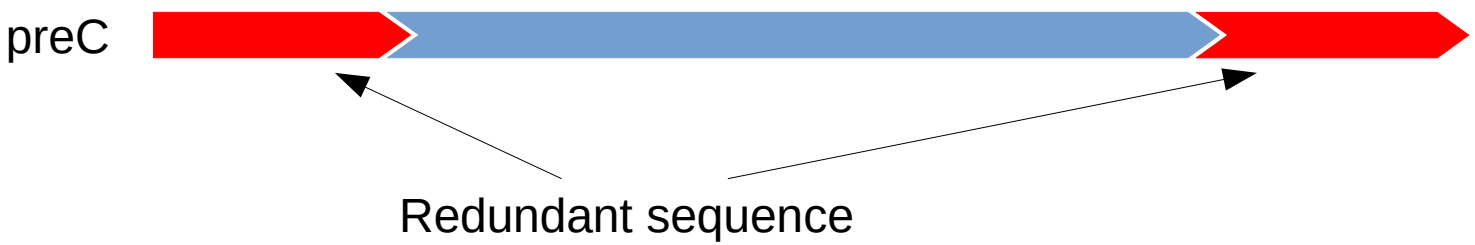
Introduction: duplicated sequence.

Hepatitis B Virus.





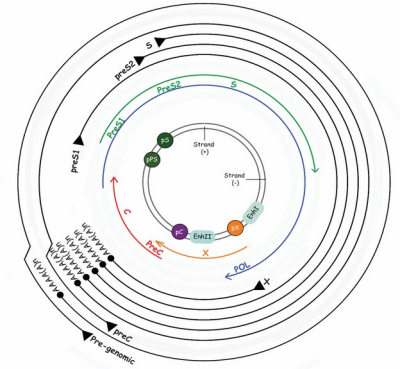
Longest mRNA is 3.5kb.
Longer than the genome.
Then redundant sequence in a single mRNA.



Questions

Questions addressed to HBV experiments:

- Question 1: Evaluation of Expression of transcript species.
- Question 2: Identification of Splice-variants.
- Question 3: Identification of New transcripts or splice-variants.



Constraints:
Circular genome,
Overlapping transcripts,
Redondant sequence.

Introduction: short-reads RNA-seq issues in HBV model.

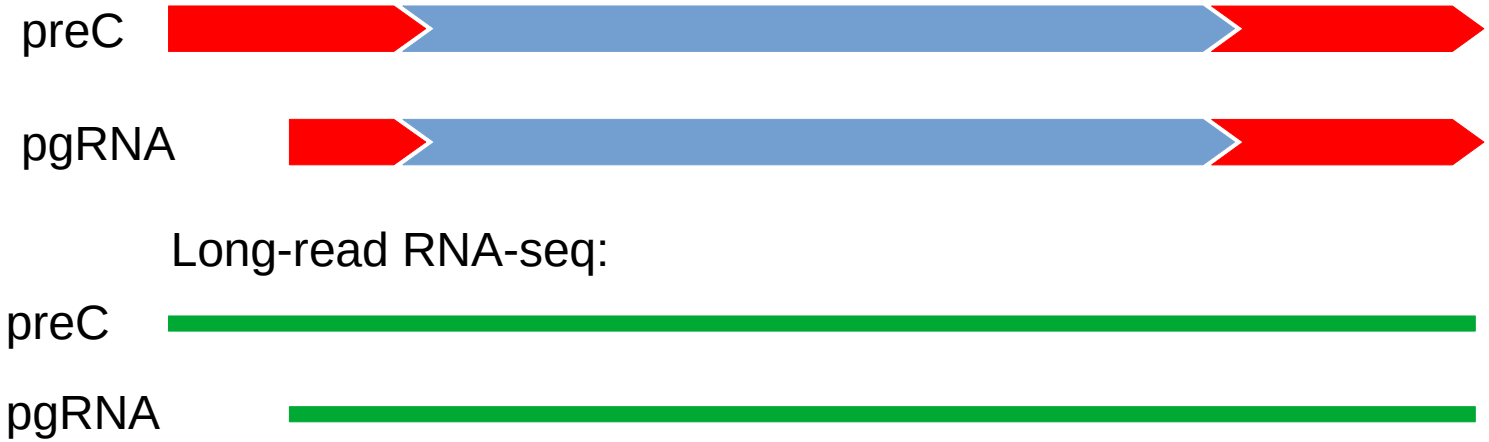
Redondant sequence:

Overlapping ORF:

Introduction: long-read sequencing of FL mRNA.

Redondant sequence:

Overlapping ORF:



Analysis steps :

I. 5’RACE products description

II. Preprocessing : Basecalling, filtering.

III. Mapping : Genome / Transcriptome.

IV. Filtering mapped results.

V. Quantification : From Genome / Transcriptome.

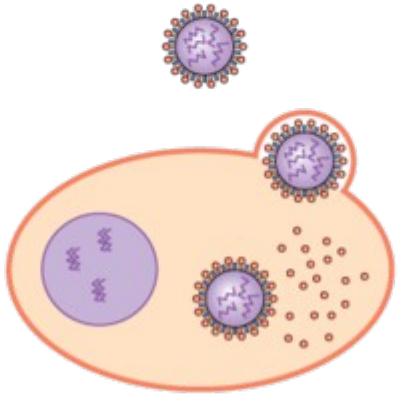
VI. Results visualization.

I. 5’RACE products description

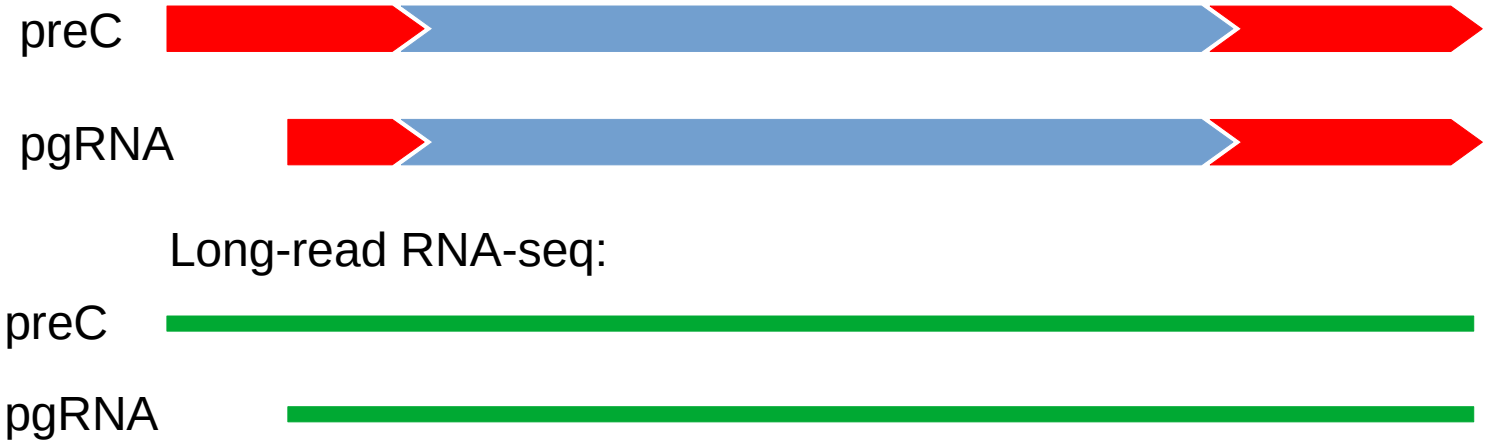
Enrichment in HBV mRNA

Filter-out Hg mRNA

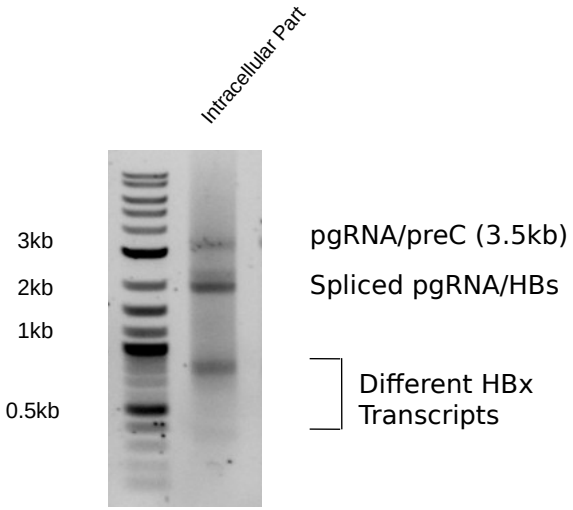
Full-length mRNA from HBV

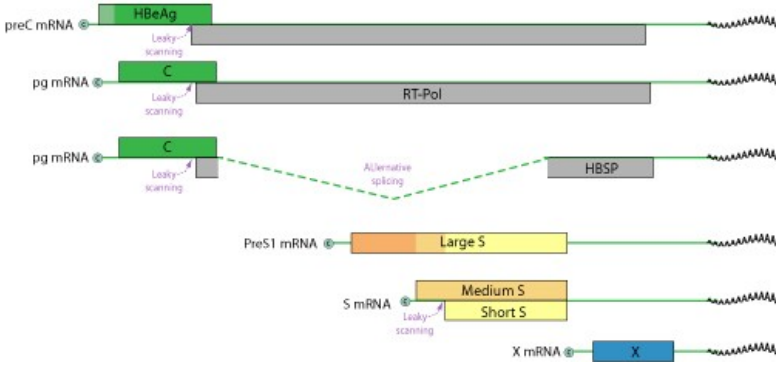


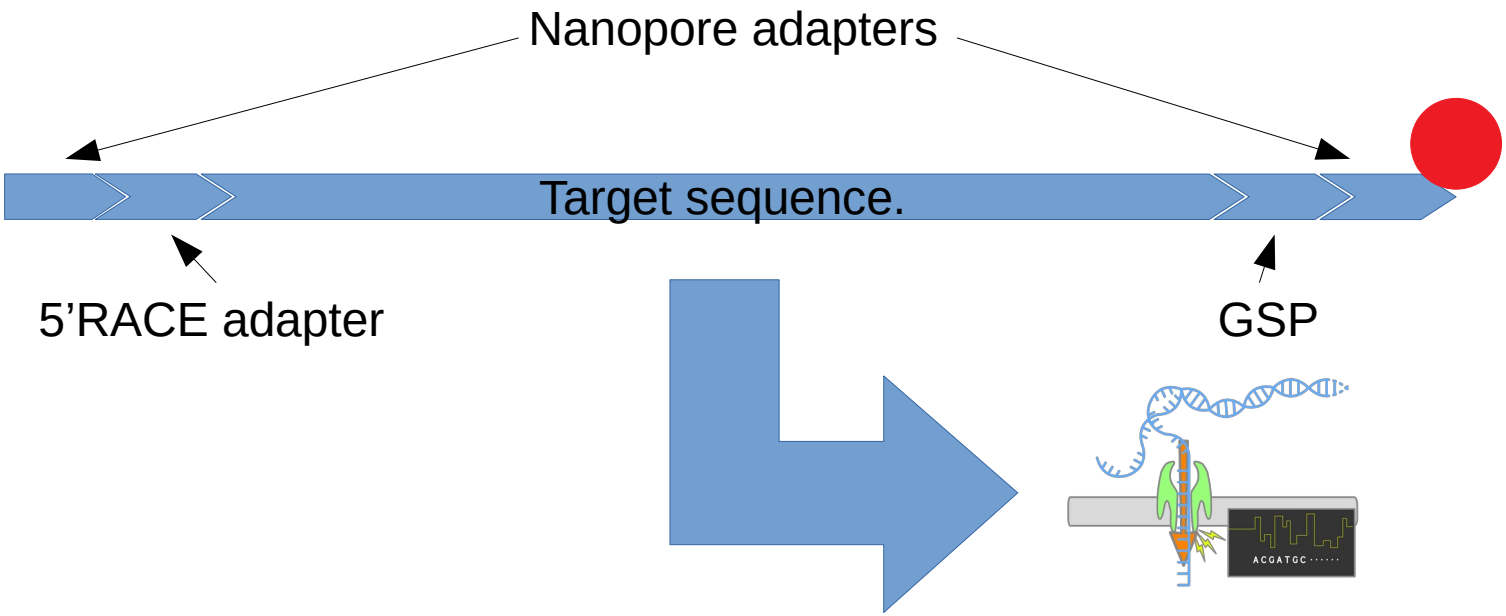
Mixed RNAs from cell and virus



I. 5’RACE products description





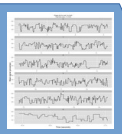


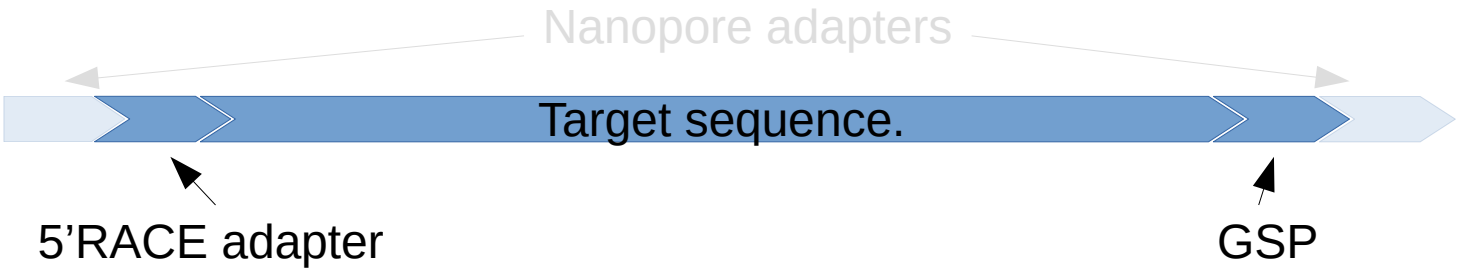
II Preprocessing :

Long-reads: Oxford Nanopore Tech.

Wetlab & sequencing

.fast5 file
Basecalling :
Guppy-basecaller





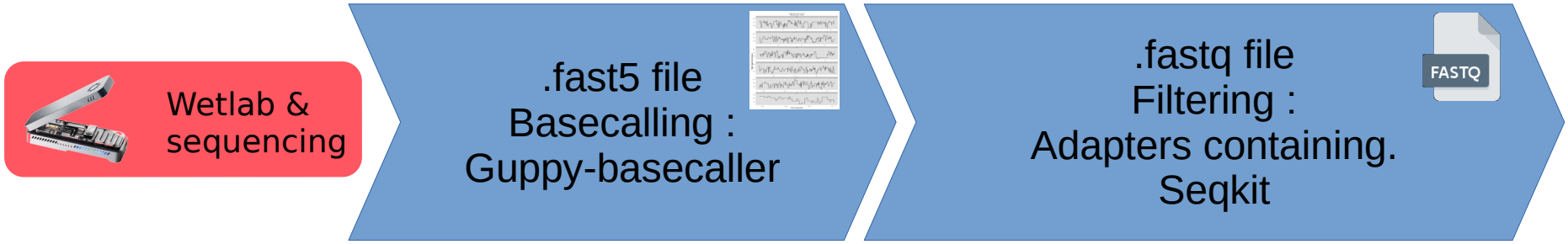
GPU optimised process.

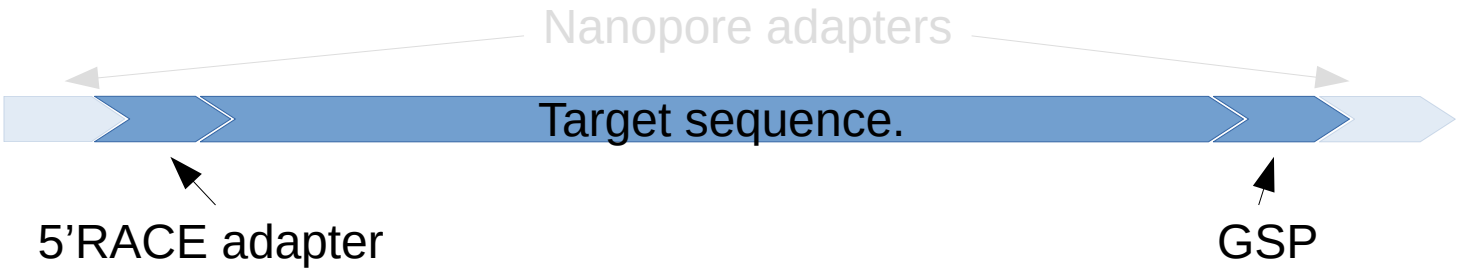
Simultaneous demultiplexing.

Simultaneous qulity filtration (Phred ≥ 7).

Trimming ONT barcodes adapters.

Long-reads: Oxford Nanopore Tech.





I. Filtering with adapters:
5' gene racer adapter:
CGACTGGAGCACGAGGACACTGACATGGACTGAAGGAGTAGAAA
3’ primer:
TTAGGCAGAGGTGAAAAAGTTG

Different strategies:
1. Extract reads containing 5’ AND 3’ adapters.
2. Exctract reads containing 5’ adapter.

Long-reads: Oxford Nanopore Tech.

.fast5 file
Basecalling :
Guppy-basecaller

.fastq file
Filtering :
Adapters containing.
Seqkit

1 000 000 reads

I. Filtering with adapters:

Extract reads containing 5’ AND 3’ adapters. (2 mm allowed).

250 000 reads

Different strategies:

1. Extract reads containing 5’ AND 3’ adapters. (2 mm allowed).

2. Extract reads containing at least 5’ adapter. (2 mm allowed).

Long-reads: Oxford Nanopore Tech.

Wetlab & sequencing

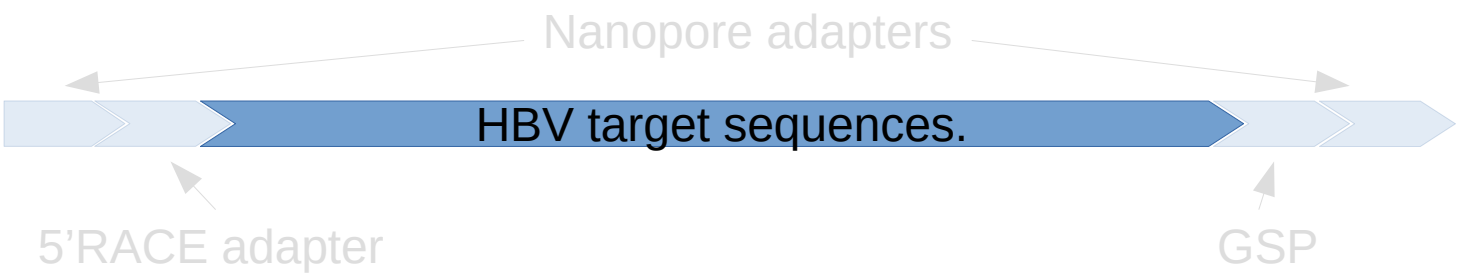
.fast5 file
Basecalling :
Guppy-basecaller

.fastq file
Filtering :
Adapters containing.
Seqkit

.fastq file
Trimming :
Porechop & Cutadapt

Trimming Nanopore Adapters

Trimming 5’RACE Adapter



Long-reads: Oxford Nanopore Tech.

Wetlab & sequencing

.fast5 file
Basecalling :
Guppy-basecaller

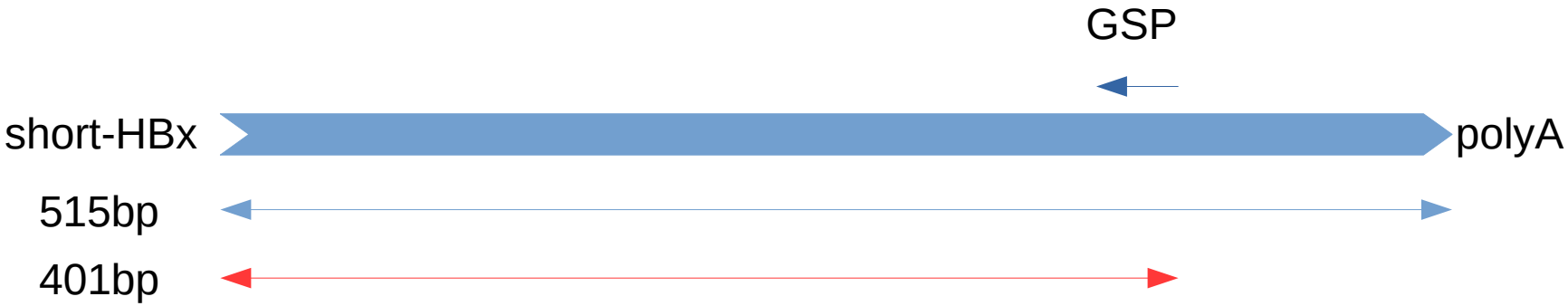
.fastq file
Filtering :
Adapters containing.
Seqkit

II. Filtering by length:

Expected smallest mRNA from HBV is 515bp long.

Then, consider only reads longer than 390bp (allowing indels produced by ONT).

Taking in account the GSP position on HBV genome, in consequence, the short-HBx expected length is 401bp.



Long-reads: Oxford Nanopore Tech.

Wetlab & sequencing

.fast5 file
Basecalling :
Guppy-basecaller

.fastq file
Filtering :
Adapters containing.
Seqkit

.fastq file
Trimming :
Porechop & Cutadapt

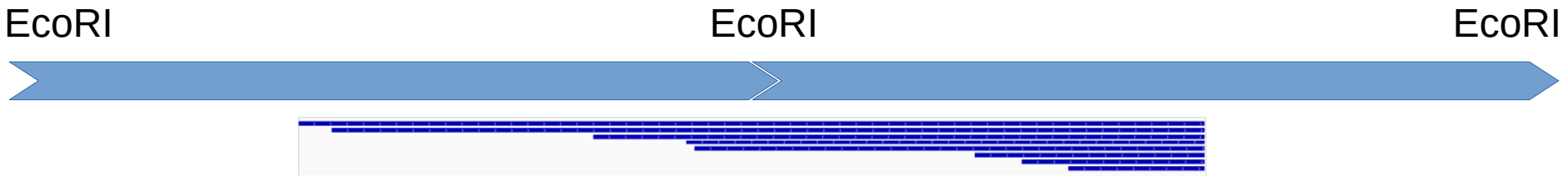
.fastq file
Filtering :
Minimum length.
Seqkit

Ready to be mapped...

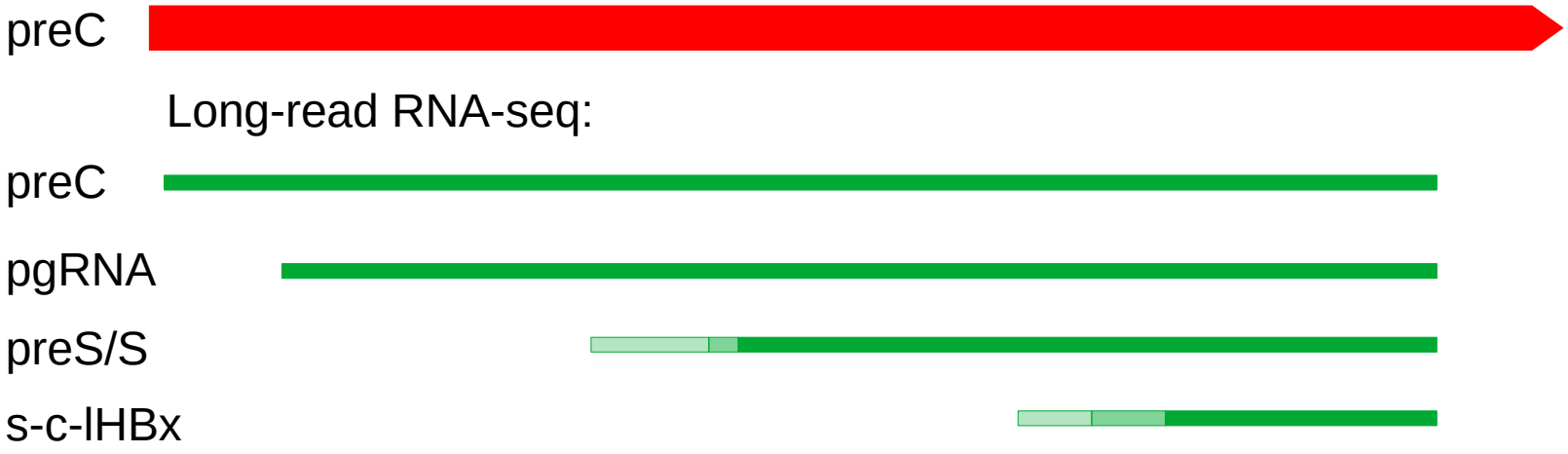
II Mapping :

III. Mapping : How to map reads on a circular genome ?

Usually Linearized and doubled reference genome (DNA sequencing).



Multi-position of reads… Problem to quantification.



2 mapping strategies, consequent quantification tools.

1. Map reads against genome.

    Need the genomic sequence.
    Fasta file containing « complete » genome.

    HBV constraints: circular and redondance.

    « complete » genome:
    _ preCore Full length (3421bp).
    _ From TSS to polyA site. (Stadelmeyer *et al.*, 2020)

2. Map reads against transcriptome.

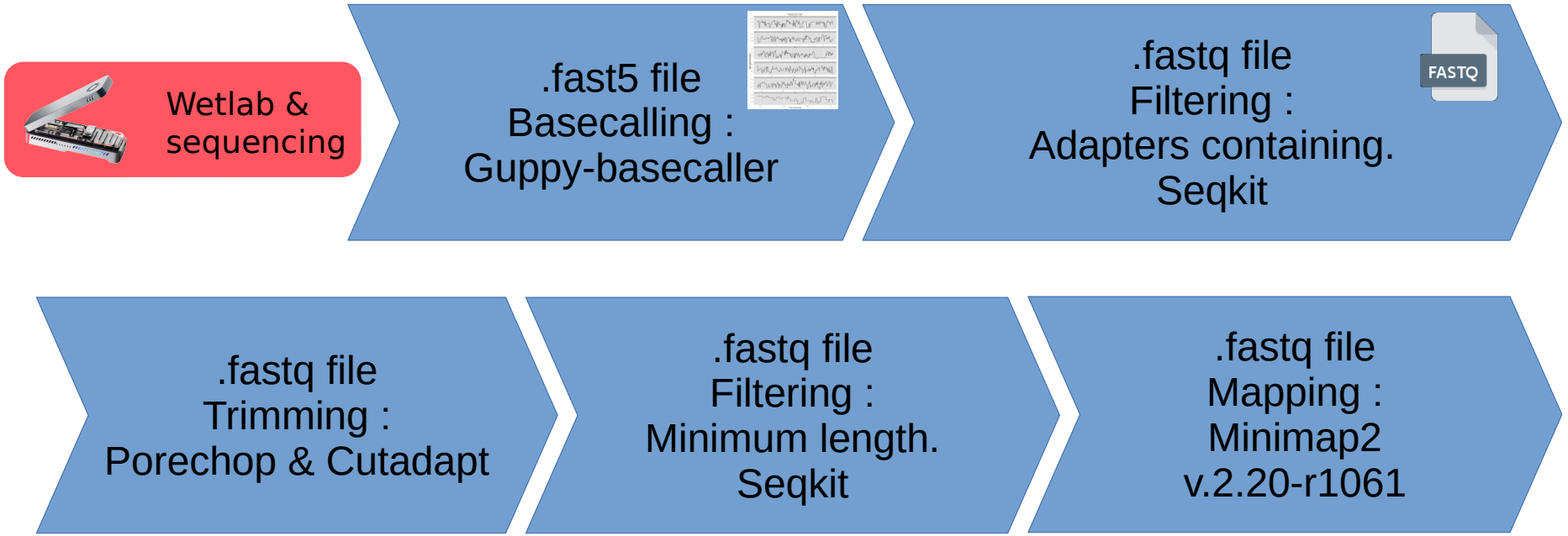
    Need the transcriptome sequences.
    Fasta file containing « all » transcripts sequences (known transcripts).

III. Mapping : Genome or transcriptome.

2 mapping strategies:

1. Map reads against genome.

2. Map reads against transcriptome.



IV Filtering of mapping results

IV. Filtering of mapping results.

Filter mapped reads:

1. MAPQ: mapping quality. Usual for short-reads, unusual for ONT... ✗

2. AS: Alignment score. Empirical adjustment… Not satisfying and not reproducible. ✗

3. Gap Max: 1650 bp (longest known splicing exclusion). Avoid spliting short reads on redundant sequences. ✓

4. Filtration of FLAG: mapped reads. ✓

Genomic :

Keep only first (best) results.

Transcriptomic :

Keep secundary alignments to Fine quantification.

V Quantification

V. Quantification : From Genome / Transcriptome.

Genome mapped reads:

StringTie2 guided by GTF file. ✅

Splice junctions count.

Over-estimation of SPxx transcripts.
« Everything » is preCore or Spxx.

Possibility to identify new transcript species.

Transcriptome mapped reads:

Salmon ✅
NanoCount ❌

Salmon counts all transcripts,
Ude of secundary alignments.

Seems to not distinguish preC and pgRNA.
Potentially adjustable by options but could disturb other alignments...

Then, the best choise could be to use Salmon and do not consider preC and pgRNA separately. When displaying reads in IGV, pgRNAs are clearly predominent.

Final pipeline analysis, 5’RACE dedicated to HBV.

Long-reads: Oxford Nanopore Tech.

Wetlab & sequencing

.fast5 file
Basecalling :
Guppy-basecaller
v.5.0.11

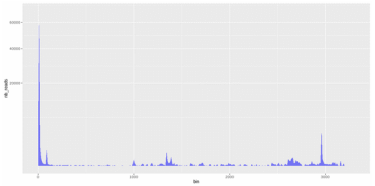
.fastq file
Filtering :
Adapters containing.
Seqkit
v.2.1.0
*

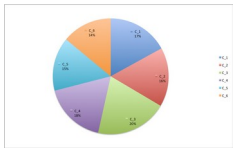
.fastq file
Trimming :
Porechop v. 0.2.4 &
Cutadapt v.1.15

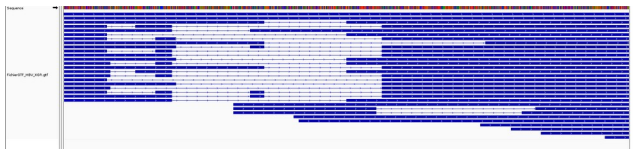
.fastq file
Mapping :
Minimap2
v.2.20-r1061

.sam file
Filtering, binarizing & indexing :
Samtools.
v.1.7
*

.bam file
quantification :
Salmon
v.1.6.0







pycoQC
Quality check
*

VI Results visualization.

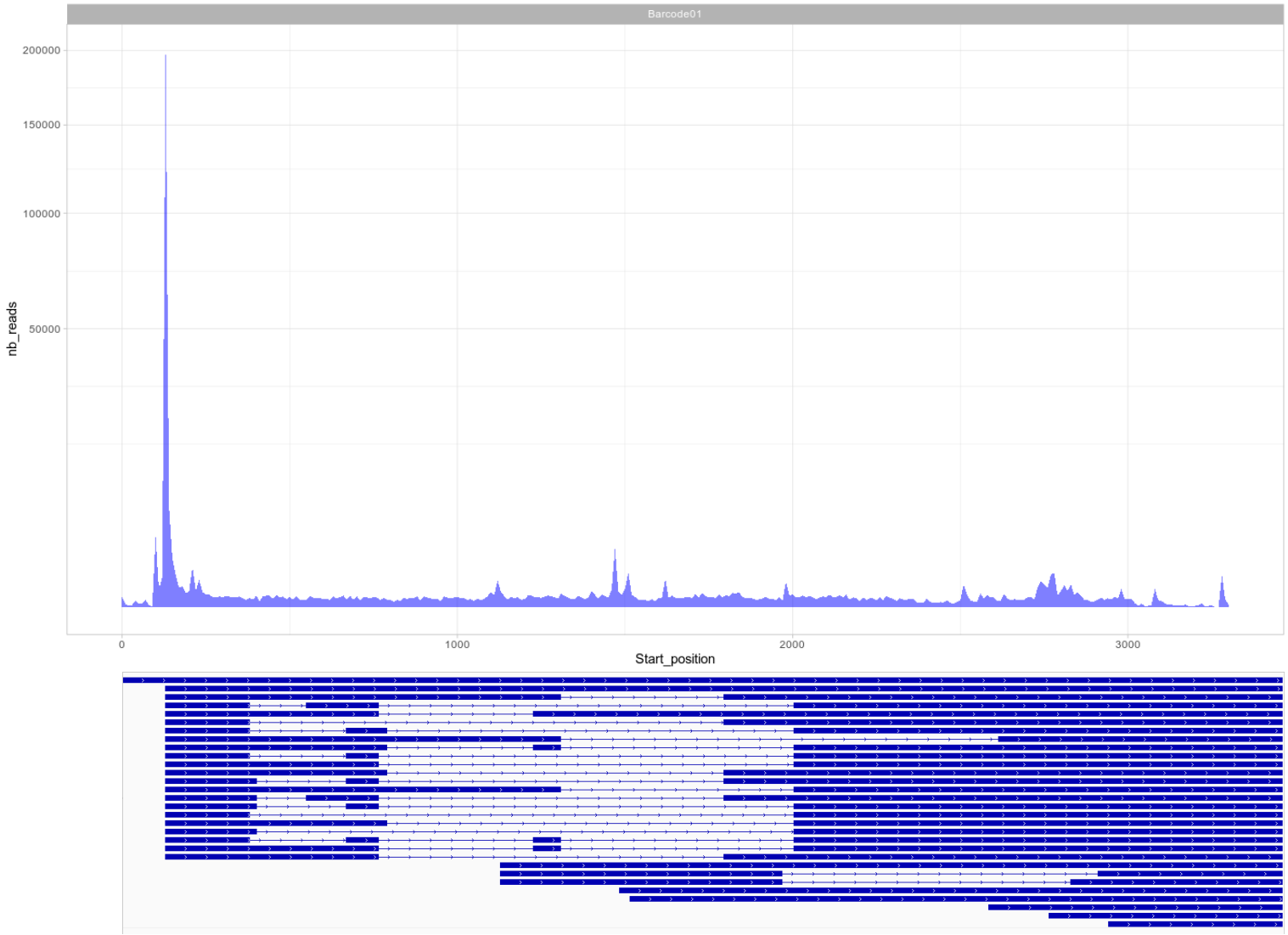
Visualization: IGV.

Specific illustration of SP, transcript species...

VI. Results visualization.

Visualization: Start positions of reads density.

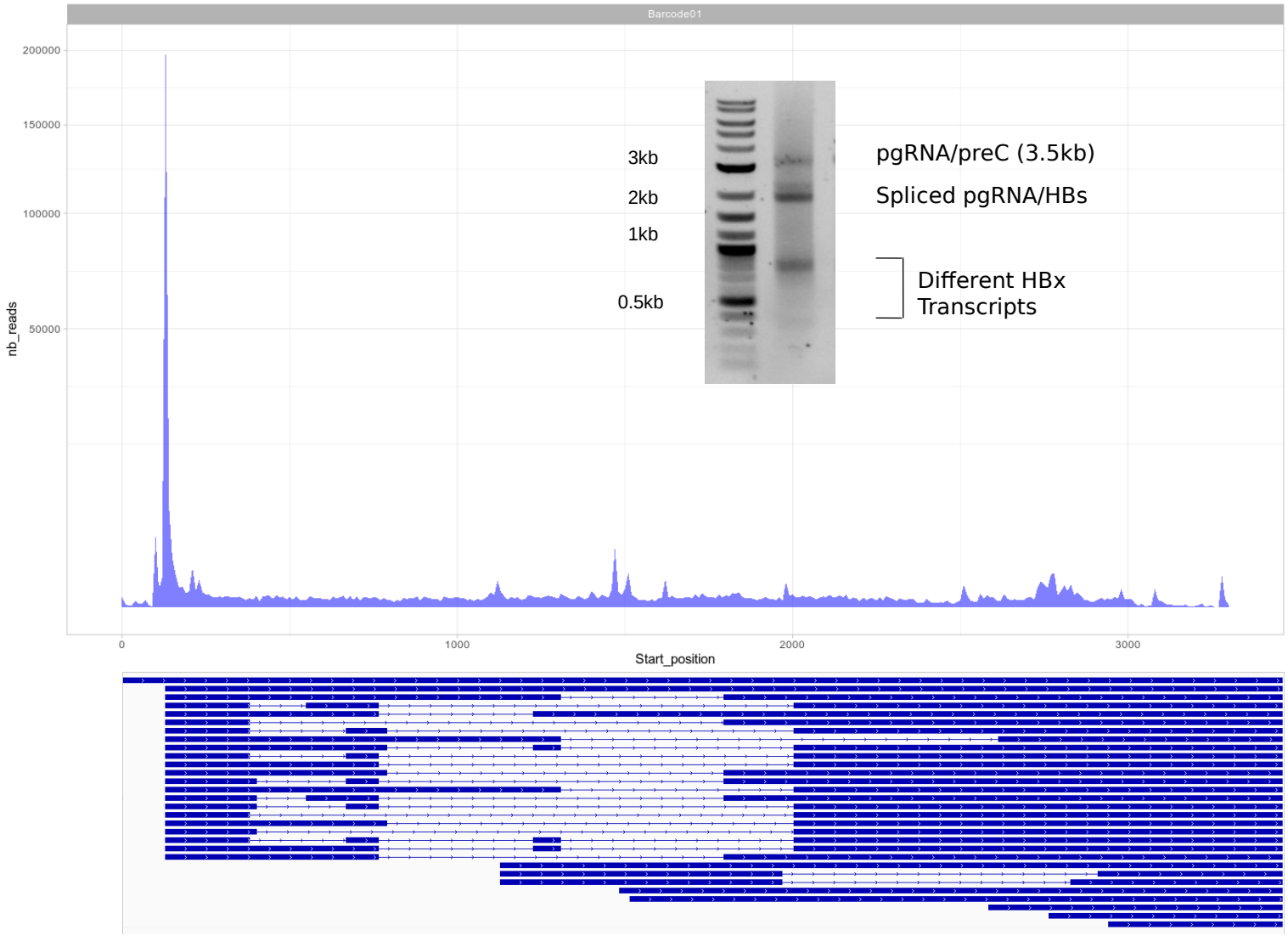
Bin = 10



VI. Results visualization.

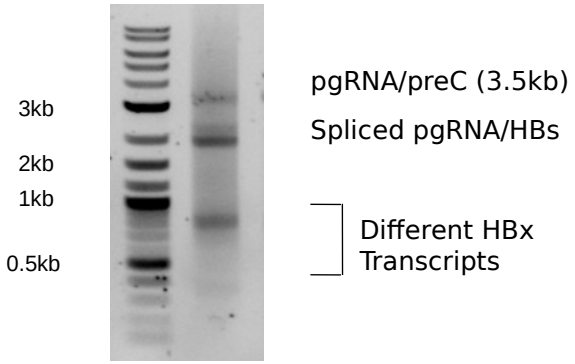
Visualization: Start positions of reads density.

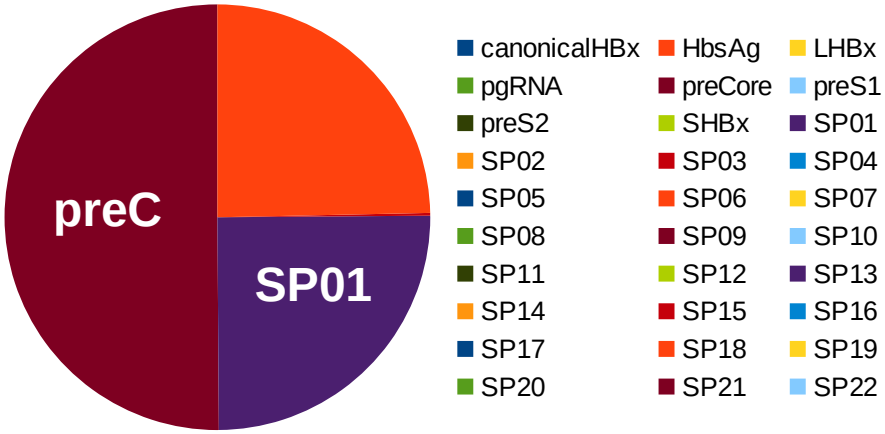
Bin = 10



VI. Results visualization.

Visualization: Transcript Per Million, StringTie2. (Genome alignment).



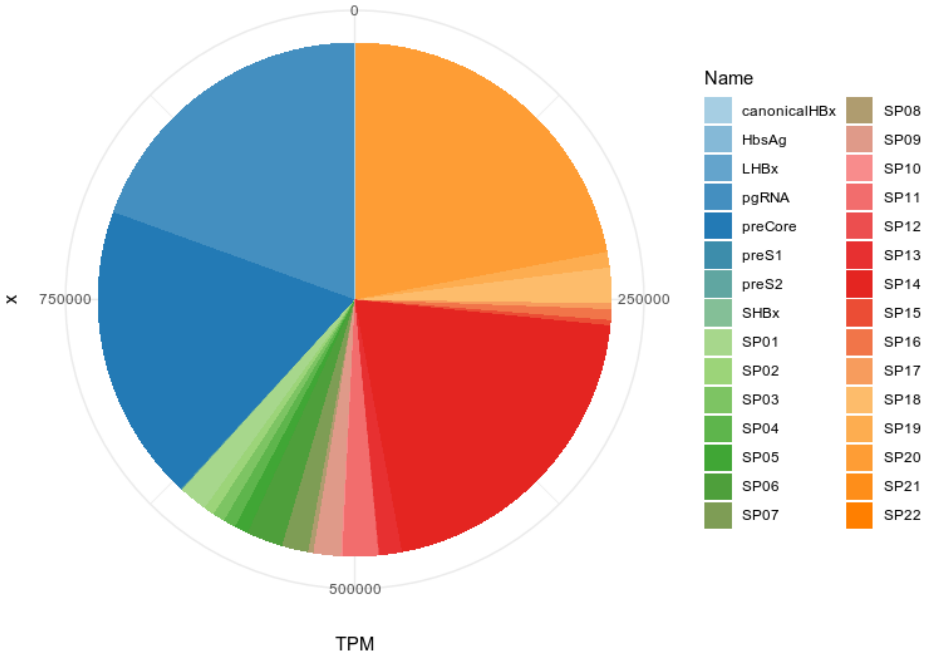


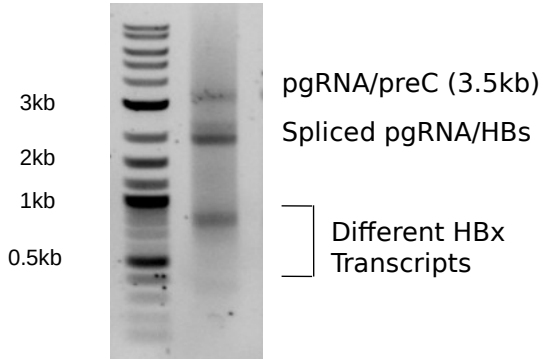
Name	TPM
canonicalHBx	0
HbsAg	0
LHBx	0
pgRNA	0
preCore	501112,375
preS1	0
preS2	0
SHBx	0
SP01	250050,45313
SP02	0
SP03	2101,647217
SP04	0
SP05	0
SP06	246735,54688
SP07	0
SP08	0
SP09	0
SP10	0
SP11	0
SP12	0
SP13	0
SP14	0
SP15	0
SP16	0
SP17	0
SP18	0
SP19	0
SP20	0
SP21	0
SP22	0

VI. Results visualization.

Visualization: Transcript proportions (%), Salmon quantification. (Transcriptome alignment).

Name	Length	EffectiveLength	TPM
canonicalHBx	690	440	430,856072
HbsAg	1926	1676	141,390716
LHBx	868	618	76,689592
pgRNA	3295	3045	193997,013835
preCore	3421	3171	186946,15277
preS1	2308	2058	621,789374
preS2	1956	1706	166,685233
SHBx	515	265	357,691832
SP01	2072	1822	17844,346412
SP02	1790	1540	6616,718249
SP03	1692	1442	7690,870526
SP04	1811	1561	7833,244922
SP05	1713	1463	8617,121443
SP06	2096	1846	22926,864523
SP07	2156	1906	16610,33795
SP08	1874	1624	3239,379703
SP09	2279	2029	17215,131447
SP10	2018	1768	911,426305
SP11	2303	2053	22462,01915
SP12	2134	1884	528,278941
SP13	2610	2360	14218,250345
SP14	2841	2591	204447,940371
SP15	1906	1656	3319,881326
SP16	1814	1564	6818,214694
SP17	1899	1649	3563,903469
SP18	2180	1930	22346,472929
SP19	2006	1756	9905,273129
SP20	2817	2567	219818,839831
SP21	1460	1210	117,506202
SP22	1380	1130	209,708708
		Total	**1000000**





To conclude

Conclusion & Remaining questions

_ Identification of potential new transcript species, StringTie2 option. But, still difficult to identify known species...

_ mRNA from HBV integrated genome. Chimeric mRNA detection and filtering.

_ TPM value depends on mRNA length. Question about full-length mRNA reads. (RPKM/FPKM is not pertinent).

_ Over-estimation/quantification of preC mRNA due to high similarity between pgRNA and preC.

_ Possible over-estimation/quantification of splice variants, due to unique junctions.