Iso-Seq
first results on transcriptomic analysis using long reads

Christophe Klopp
Genotoul bioinfo http://bioinfo.genotoul.fr

Aeschynod project

- Jean-François Arrighi (IRD)
- Philippe Leleux (IRD)
- Léo Lamy (IRD)
- ANR 2014
- 400 Mb genome
  - WGS PacBio/MiSeq
  - RNA-Seq : HiSeq/PacBio
  - GBS : HiSeq

What is Iso-Seq ?

- A PacBio trade mark.
- Produces full-length transcripts without assembly.
- The Iso-Seq method generates accurate information about alternatively spliced exons and transcriptional start sites.
- It also delivers information about poly-adenylation sites and therefore the strand.

http://www.pacb.com/applications/rna-sequencing/

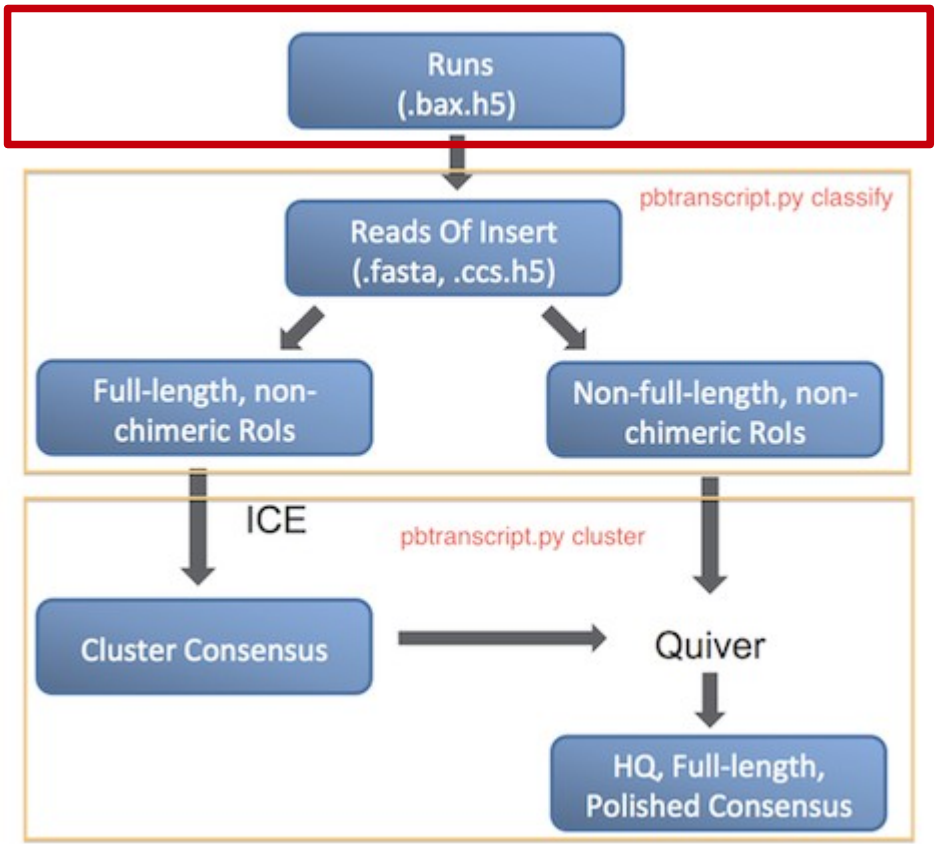
Outline

- Raw data
- pbtranscript.py : processing pipeline
  - Step one : classify
  - Step two : cluster
- Transcriptome coverage
- Detected problems

IsoSeq processing schema

Auto-correction

In cluster correction

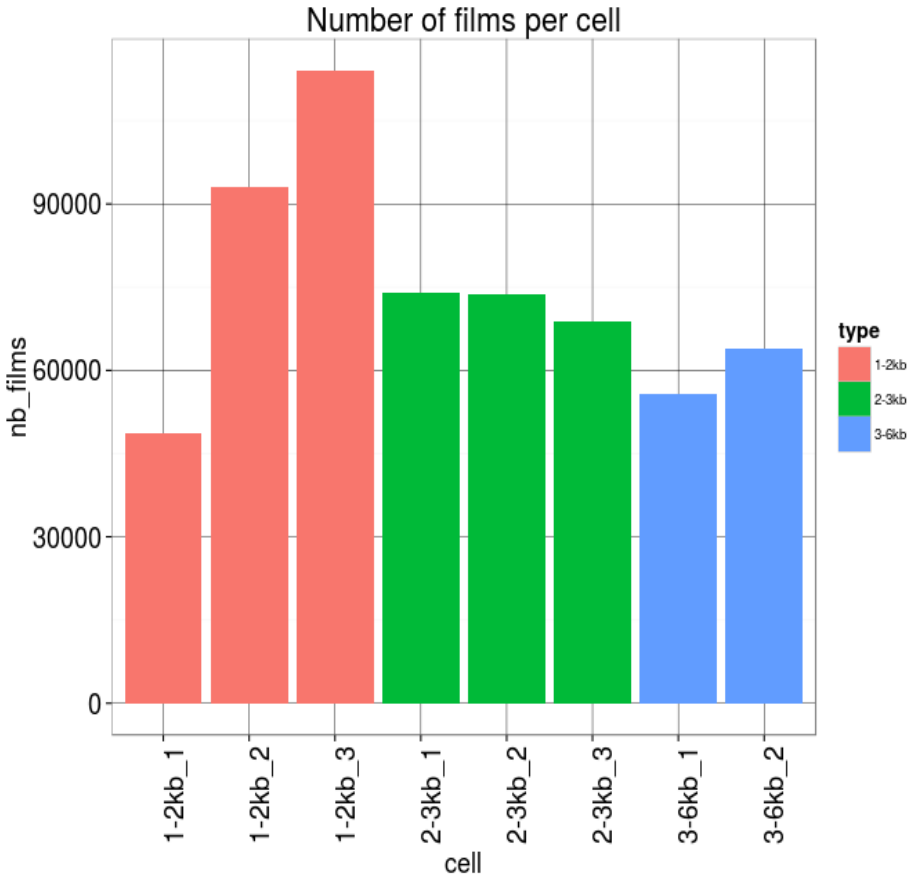


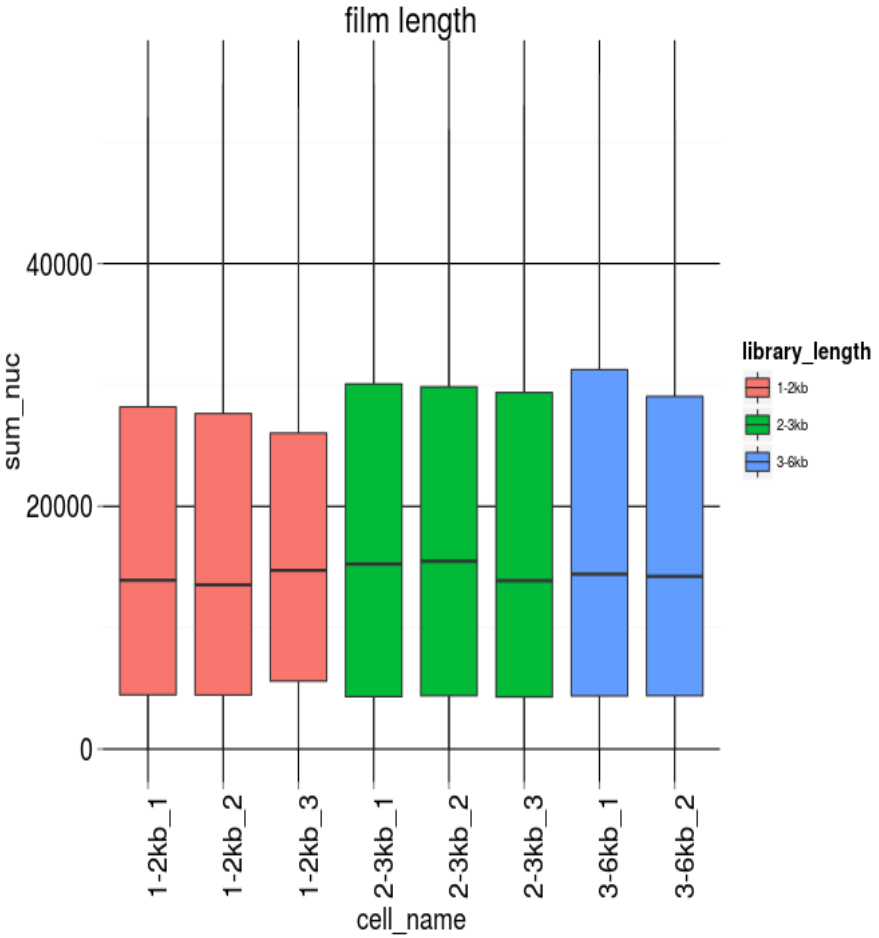
https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%2

A evenia IsoSeq protocol

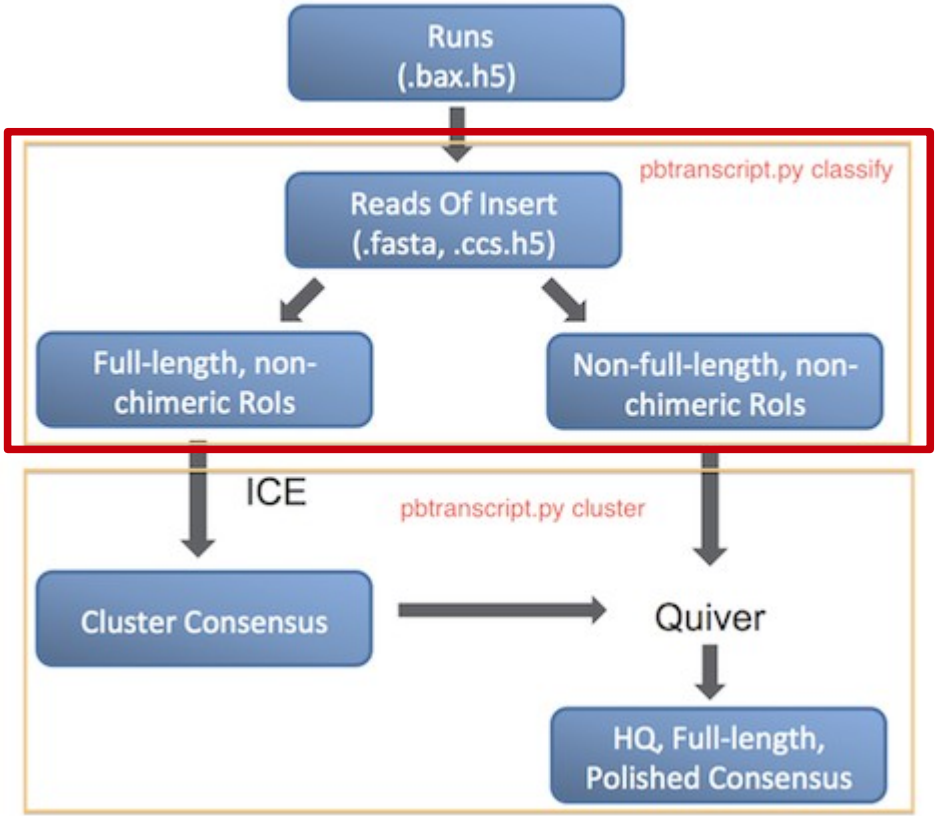
- In order to have reads of different length in the results different libraries are build.
- One or several cell can be produced per library.
- A evenia :
  - 1 kb to 2 kb library : 3 cells
  - 2 kb to 3 kb library : 3 cells
  - 3 kb to 6 kb library : 2 cells

Films


Number of films per cell
nb_films
90000
60000
30000
0
1-2kb_1
1-2kb_2
1-2kb_3
2-3kb_1
2-3kb_2
2-3kb_3
3-6kb_1
3-6kb_2
cell
type
1-2kb
2-3kb
3-6kb

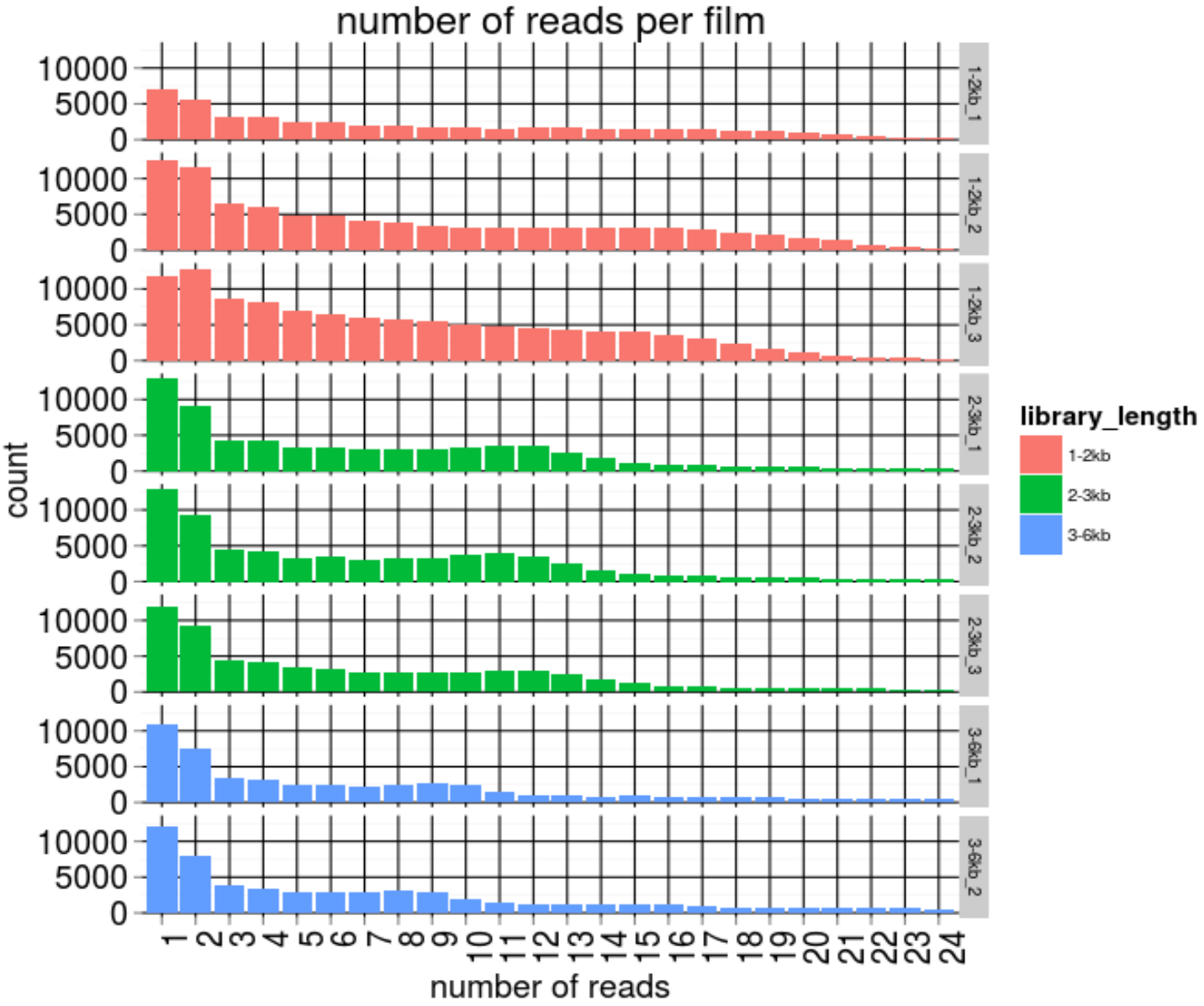

film length
sum_nuc
40000
20000
0
1-2kb_1
1-2kb_2
1-2kb_3
2-3kb_1
2-3kb_2
2-3kb_3
3-6kb_1
3-6kb_2
cell_name
library_length
1-2kb
2-3kb
3-6kb

IsoSeq processing



https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%2

Reads per film

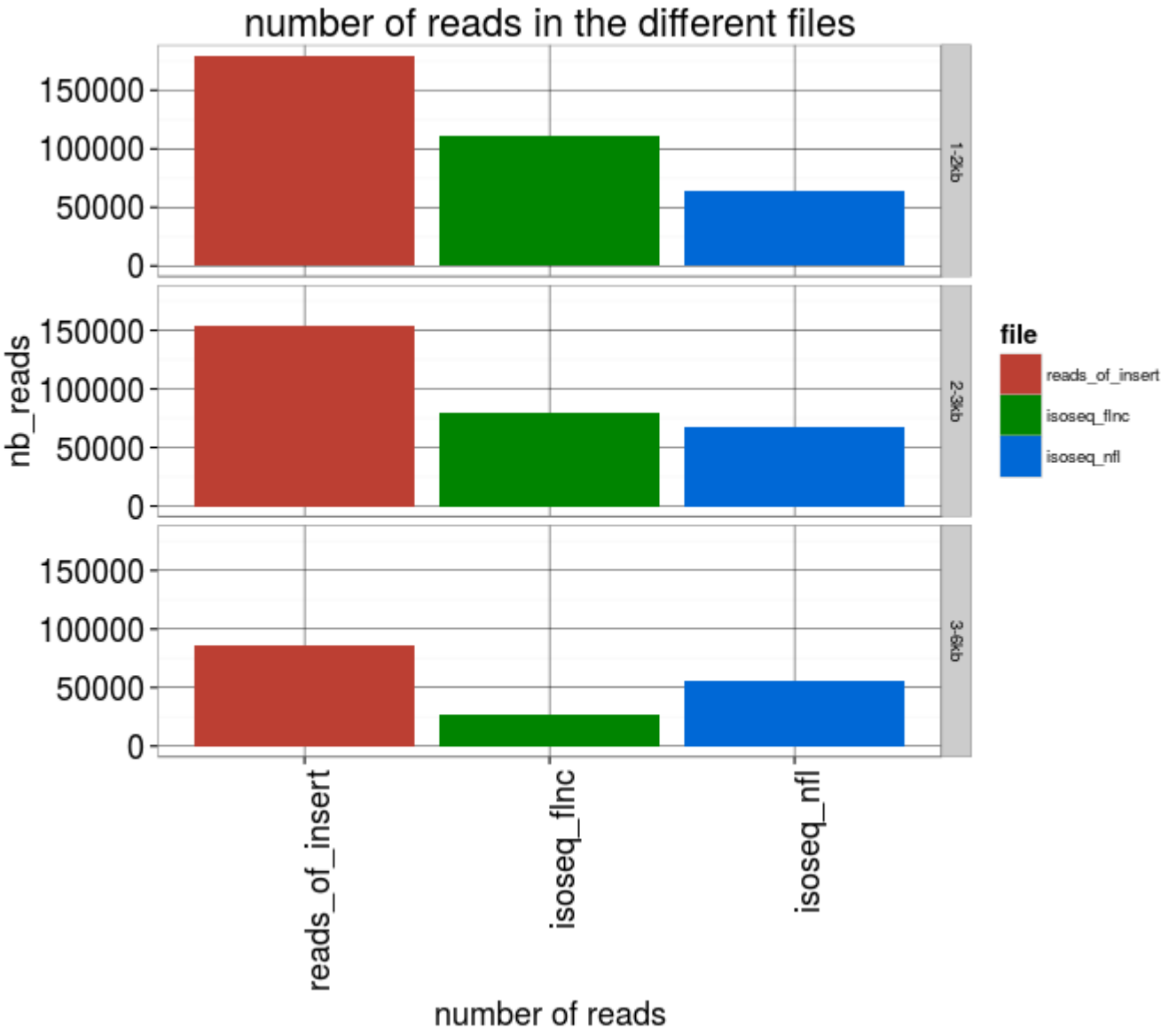


isoseq_draft.primer_info.csv

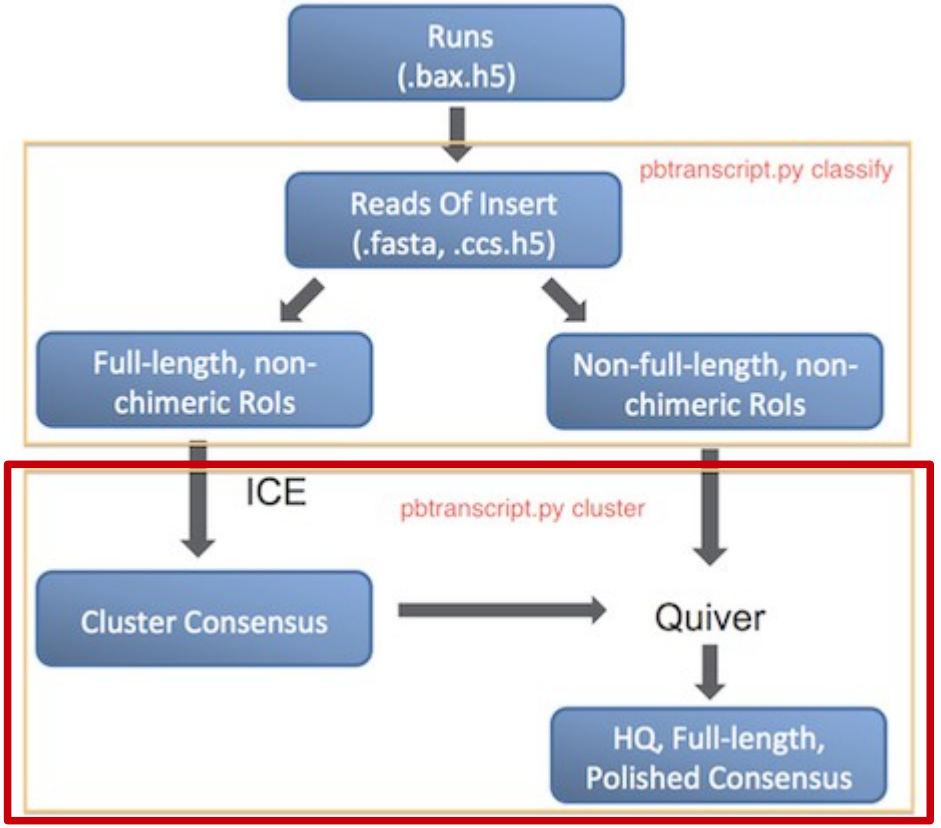
```
/work2/project/sigenae/Project_AeschyNod.451/IsoSeq/LID50178_1-2kb_results>more isoseq_draft.primer_info.csv
id,strand,fiveseen,polyAseen,threeseen,fiveend,polyAend,threeend,primer,chimera
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/31/1749_58_CCS,-,1,1,1,31,1722,1751,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/43/31_1754_CCS,+,1,1,1,31,1754,1780,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/60/2002_56_CCS,-,1,1,1,30,1976,2003,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/92/31_1933_CCS,+,1,1,1,31,1933,1966,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/135/1743_59_CCS,-,1,1,1,31,1715,1745,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/138/27_1849_CCS,+,1,1,1,27,1849,1868,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/150/31_2081_CCS,+,1,1,1,31,2081,2111,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/153/31_1864_CCS,+,1,1,1,31,1864,1890,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/160/31_2022_CCS,+,1,1,1,31,2022,2047,1,0
m151006_205116_42137_c100912202550000001823210404301625_s1_p0/162/2538_57_CCS,-,1,1,1,31,2512,2539,1,0
```

id	strand	fiveseen	polyAseen	threeseen	fiveend	polyAend	threeend	primer	chimera
10404301625_s1_p0/31/1749_58_CCS	-	1	1	1	31	1722	1751	1	0
10404301625_s1_p0/43/31_1754_CCS	+	1	1	1	31	1754	1780	1	0
10404301625_s1_p0/60/2002_56_CCS	-	1	1	1	30	1976	2003	1	0
10404301625_s1_p0/92/31_1933_CCS	+	1	1	1	31	1933	1966	1	0
10404301625_s1_p0/135/1743_59_CCS	-	1	1	1	31	1715	1745	1	0
10404301625_s1_p0/138/27_1849_CCS	+	1	1	1	27	1849	1868	1	0
10404301625_s1_p0/150/31_2081_CCS	+	1	1	1	31	2081	2111	1	0
10404301625_s1_p0/153/31_1864_CCS	+	1	1	1	31	1864	1890	1	0
10404301625_s1_p0/160/31_2022_CCS	+	1	1	1	31	2022	2047	1	0
10404301625_s1_p0/162/2538_57_CCS	-	1	1	1	31	2512	2539	1	0
10404301625_s1_p0/165/31_2407_CCS	+	1	1	1	31	2407	2439	1	0

RoI & flnc & nfl


number of reads in the different files
nb_reads
150000
100000
50000
0
1-2kb
2-3kb
3-6kb
file
reads_of_insert
isoseq_flnc
isoseq_nfl
reads_of_insert
isoseq_flnc
isoseq_nfl
number of reads

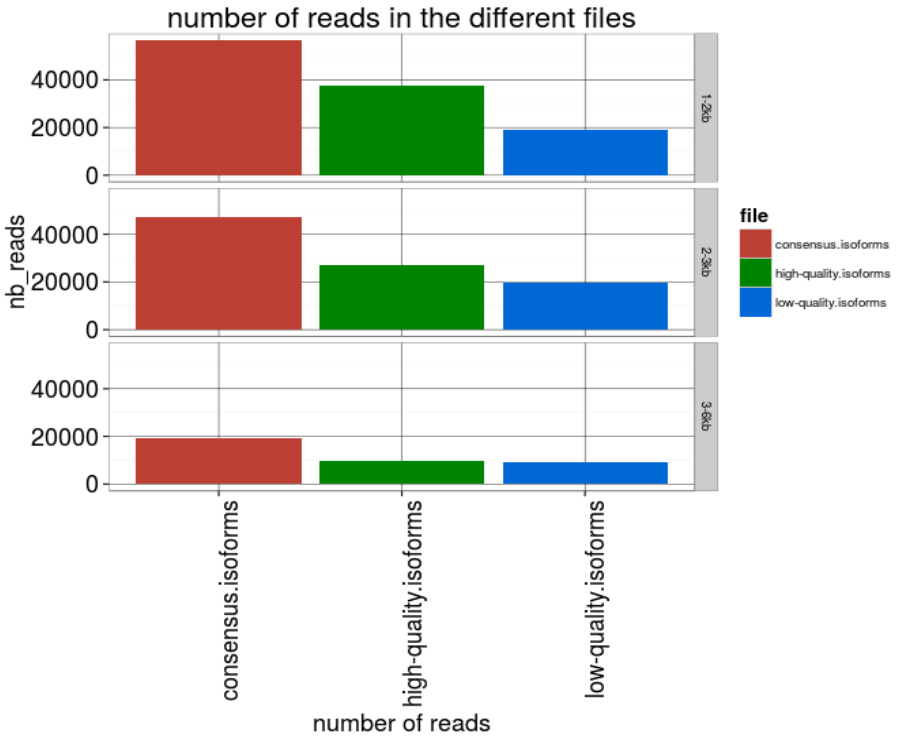
IsoSeq processing



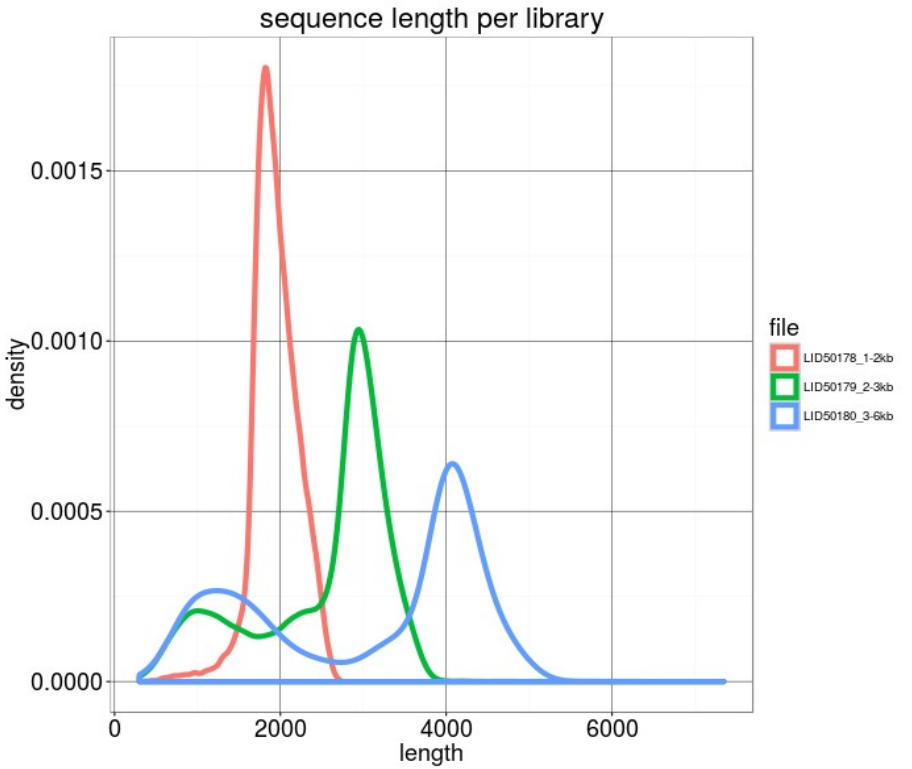
https://github.com/PacificBiosciences/cDNA_primer/wiki/RS_IsoSeq-%28v2.3%29-Tutorial-%232.-Isoform-level-clustering-%28ICE-and-Quiver%2

Quiver polishing

```
 76671246 Jan 14 08:46 all_quivered_lq.fastq
149827297 Jan 14 08:46 all_quivered_hq.100_30_0.99.fastq
 39902492 Jan 14 08:46 all_quivered_lq.fasta
 78005140 Jan 14 08:46 all_quivered_hq.100_30_0.99.fasta
```



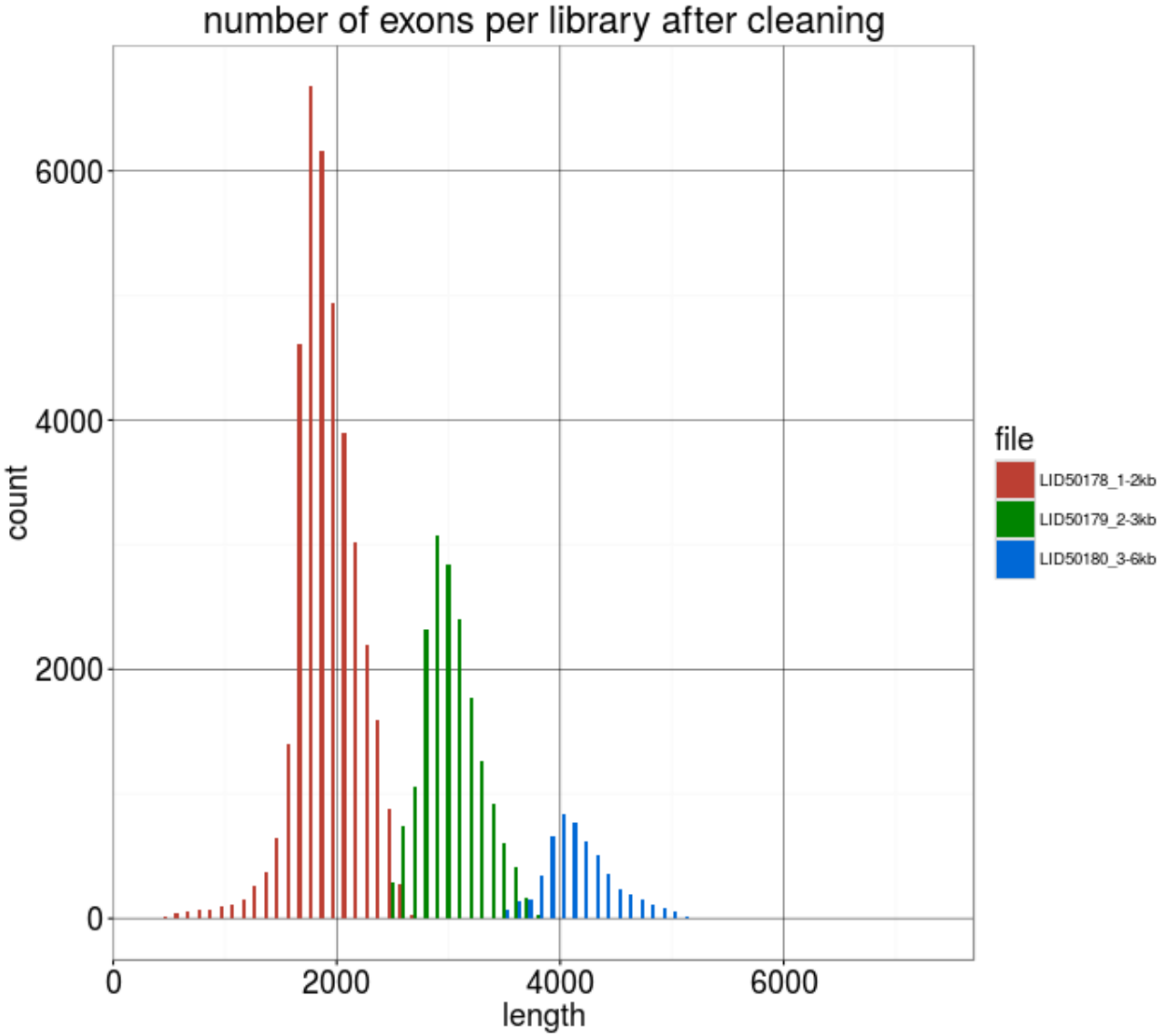
Library	nb HQ Isoforms
LID50178_1-2kb	37,615
LID50179_2-3kb	27,345
LID50180_3-6kb	9,771



Removing too short reads

```
LID50178_1-2kb 37,570
LID50179_2-3kb 17,938
LID50180_3-6kb 5,345
```

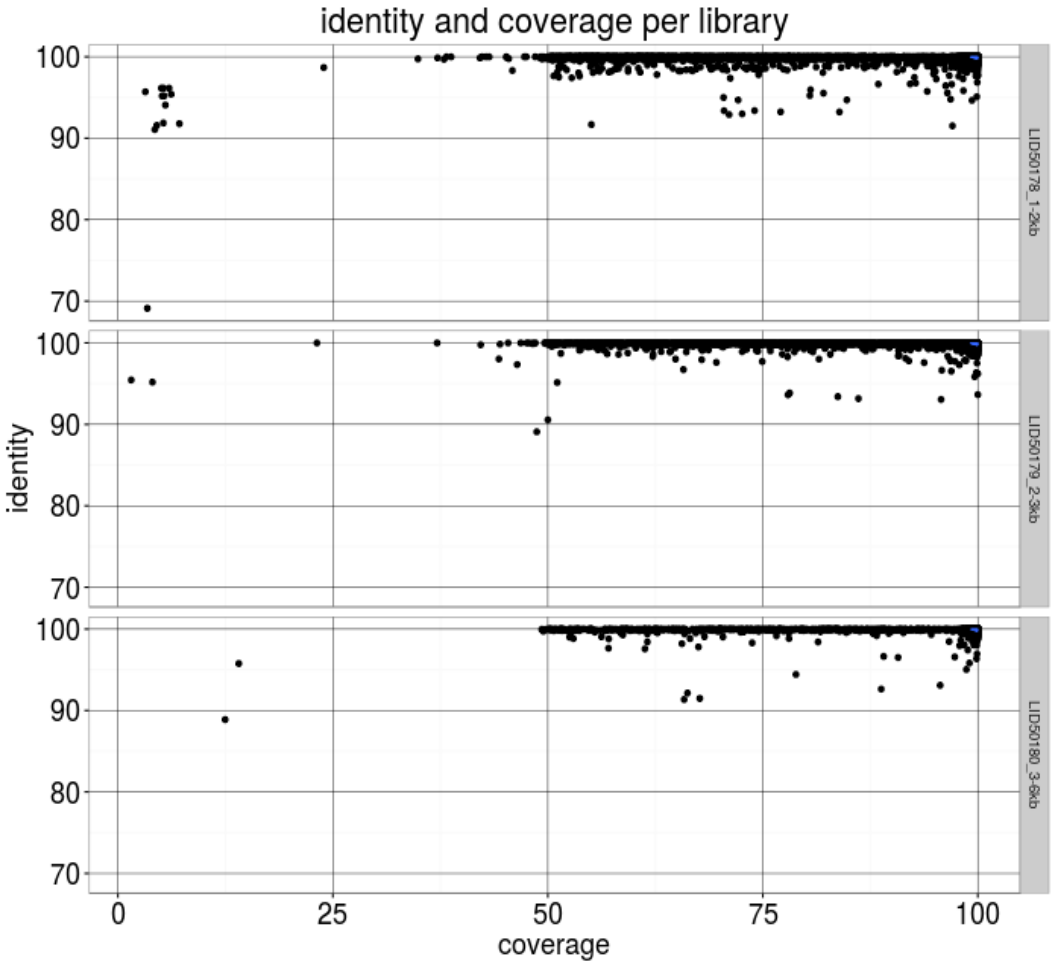
-34% for 2-3kb
-45% for 3-6kb

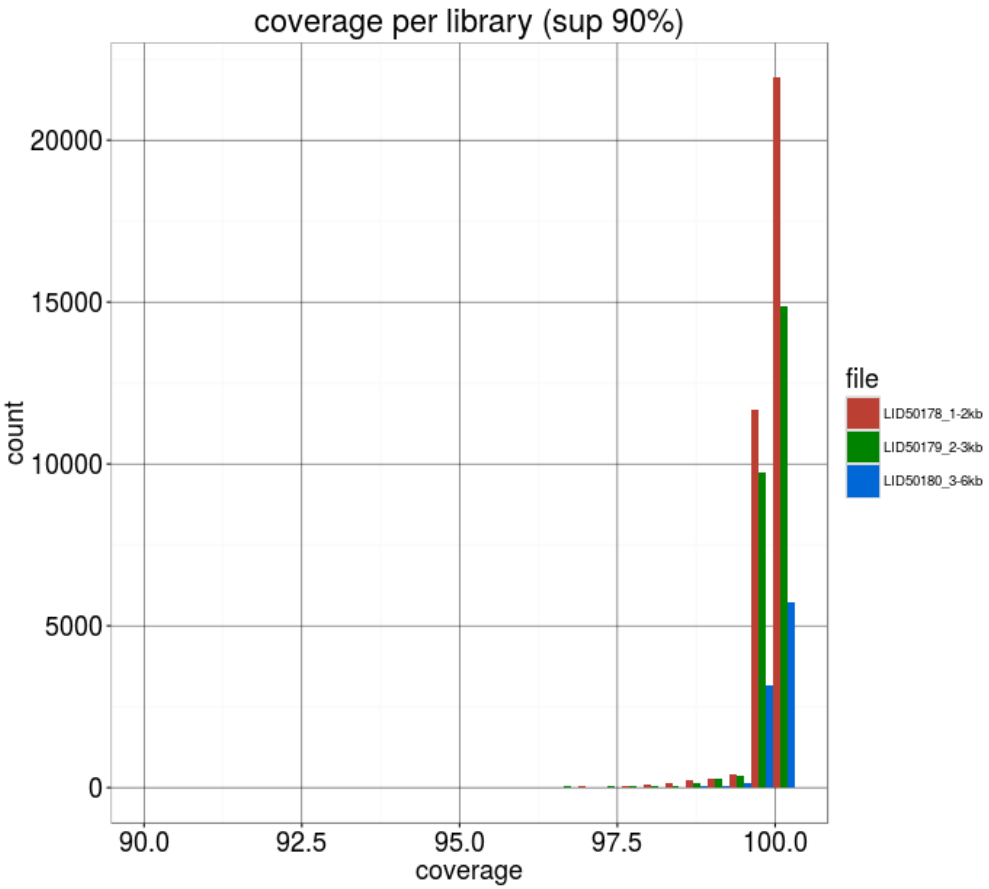


Questions

- What is the quality of the resulting data?
- How large is the Iso-Seq transcriptome coverage?
- Is there a benefit of having multiple size libraries?
- Do we see isoforms?

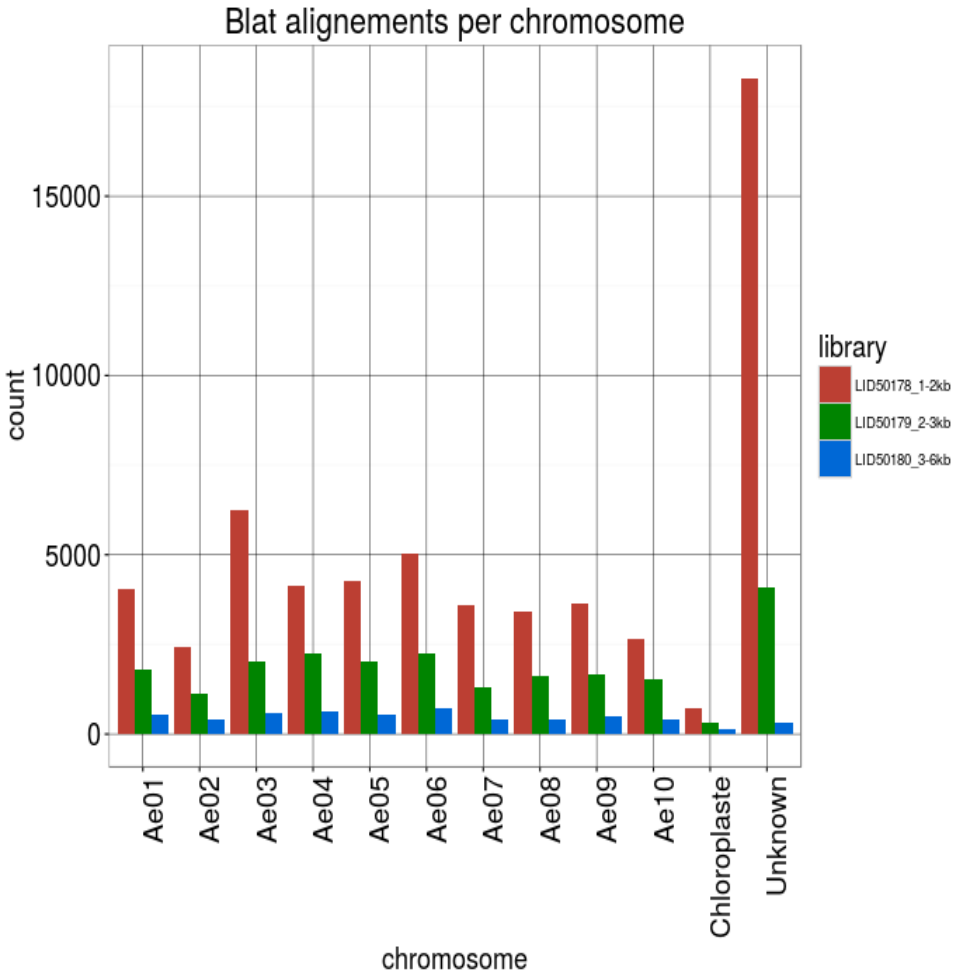
Genome alignment results

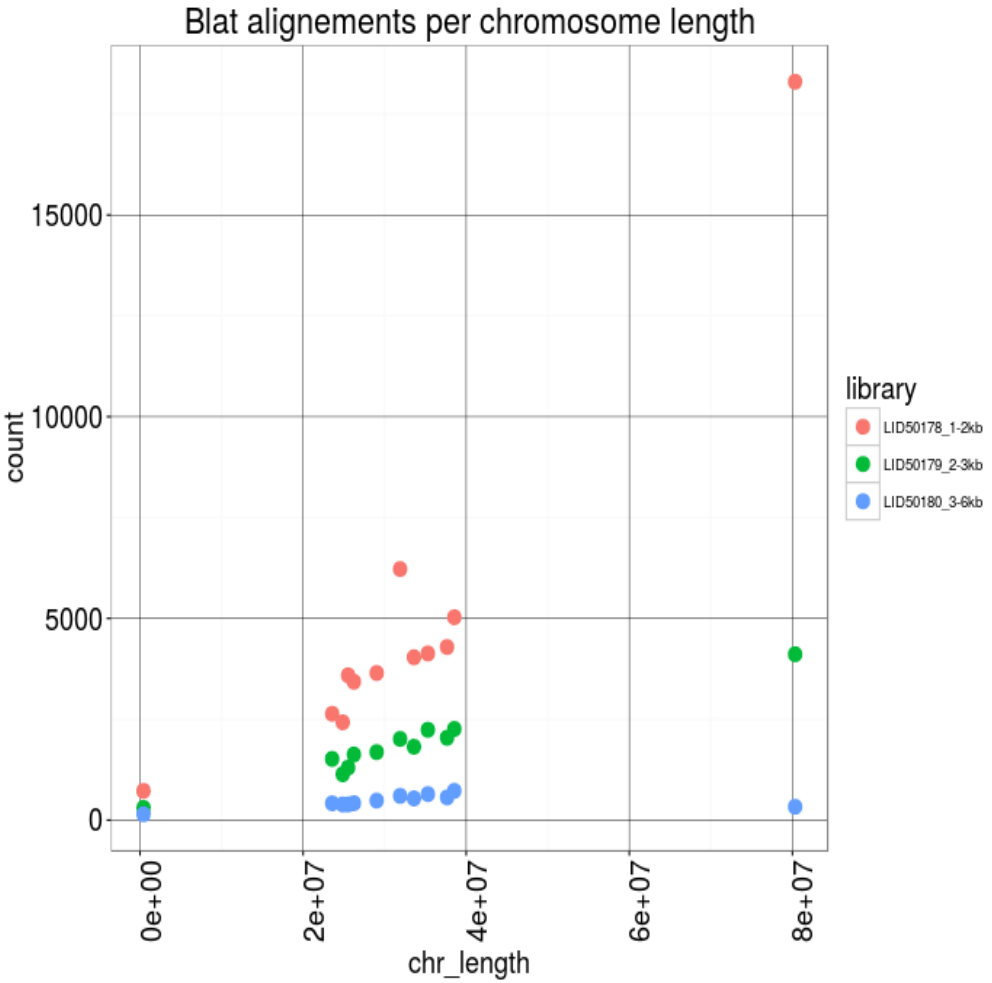




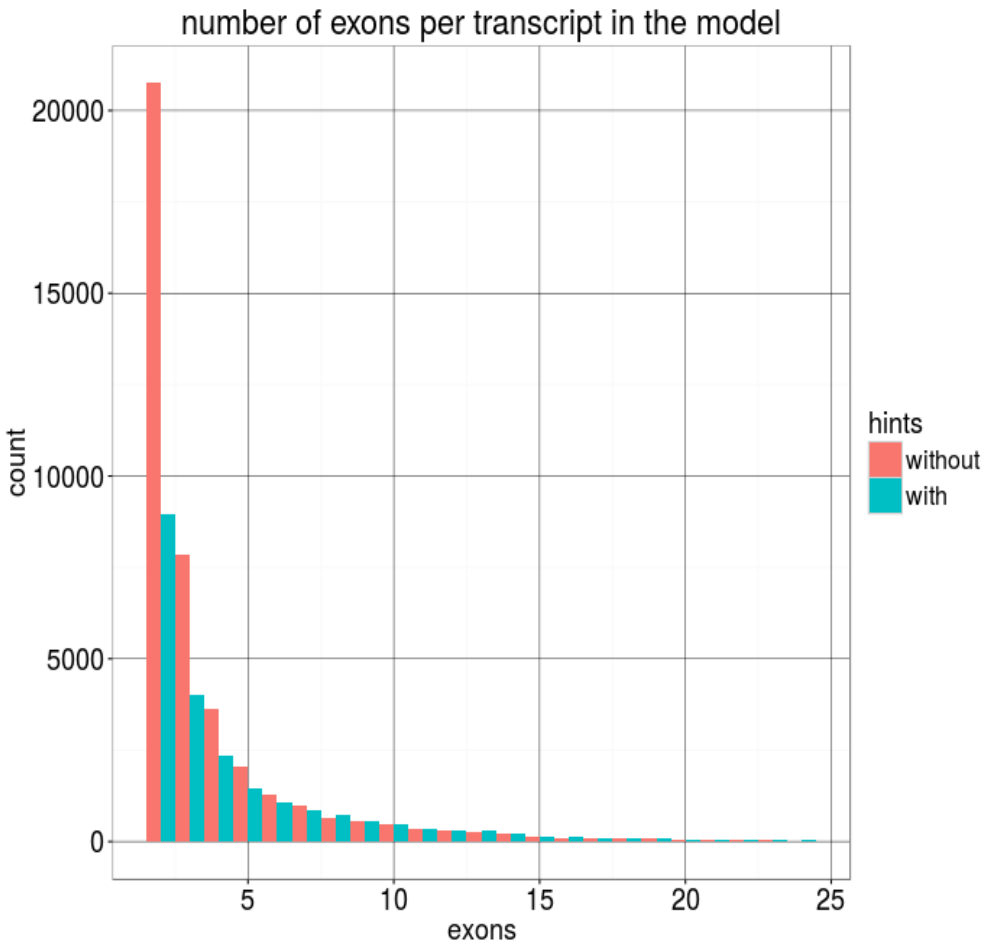
	File	Initial	Aligned	Al. rate
1	LID50178_1-2kb	37615	37570	0.9988037
2	LID50179_2-3kb	27345	27315	0.9989029
3	LID50180_3-6kb	9771	9763	0.9991813

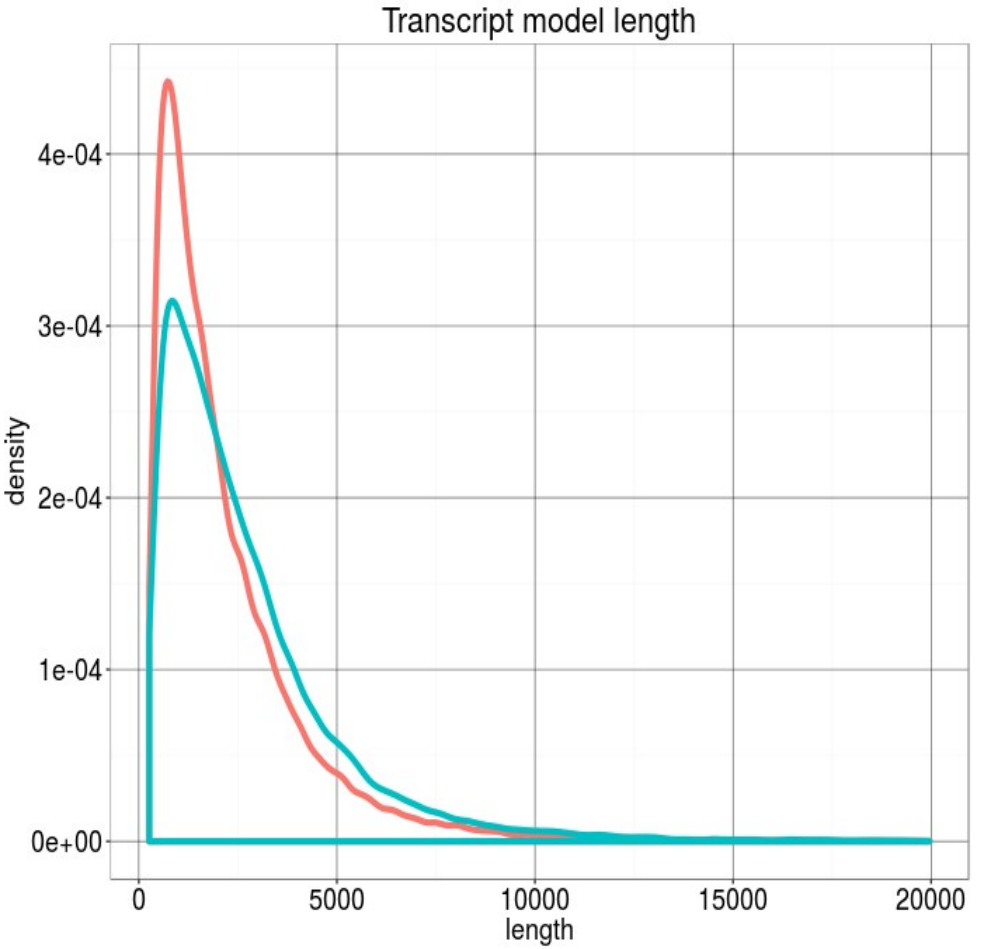
Blat alignments





Gene model transcripts

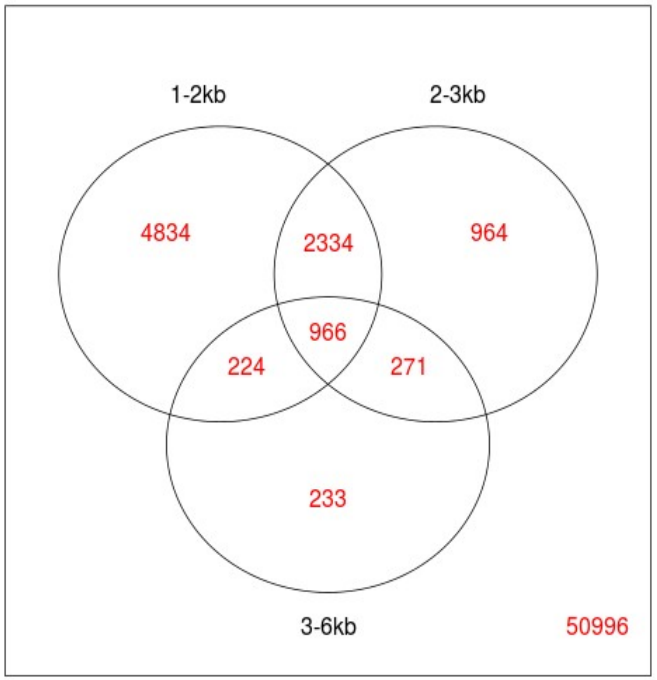




Transcripts with hints	22,506
Transcripts without hints	40,104

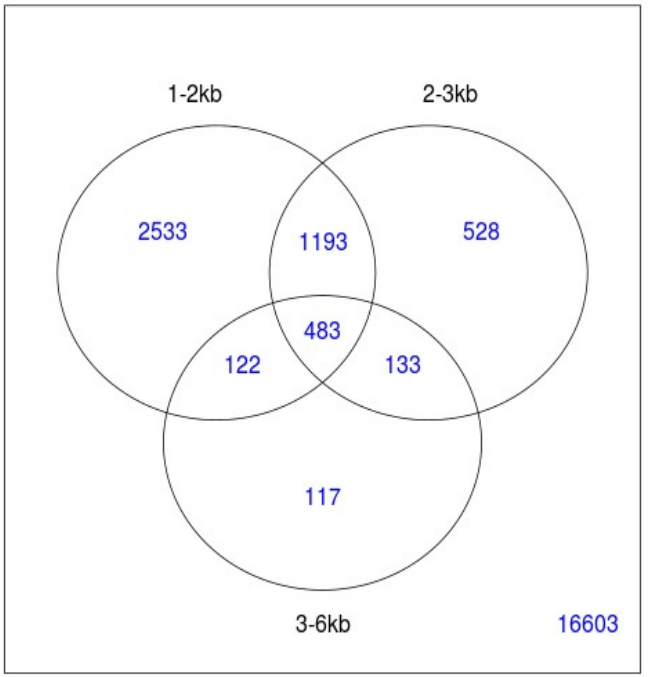
Correspondence with the model

sequences overlapping genes of the models



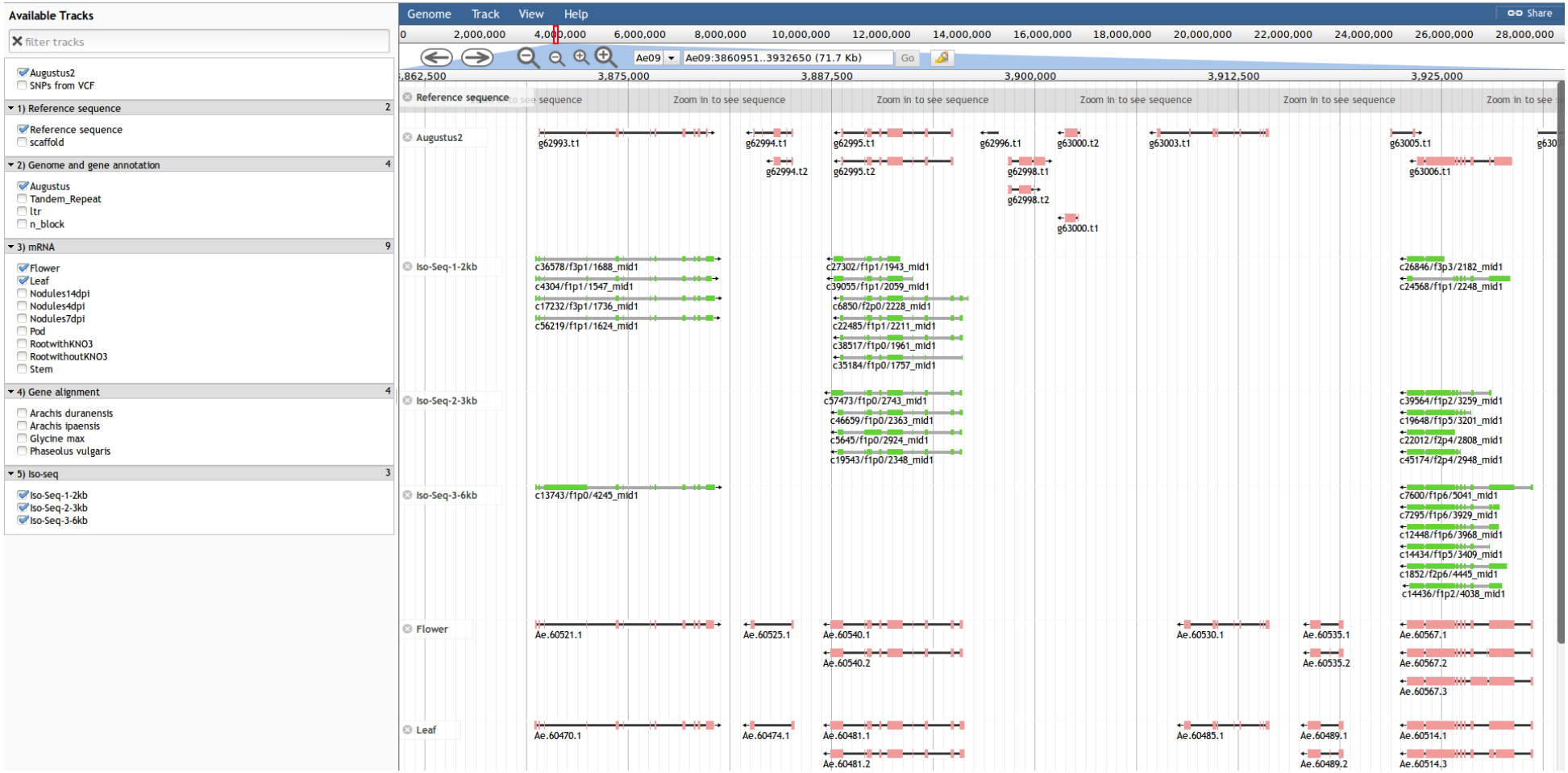
9,826 genes
Gene coverage 16,15 %

sequences overlapping genes with hints of the models



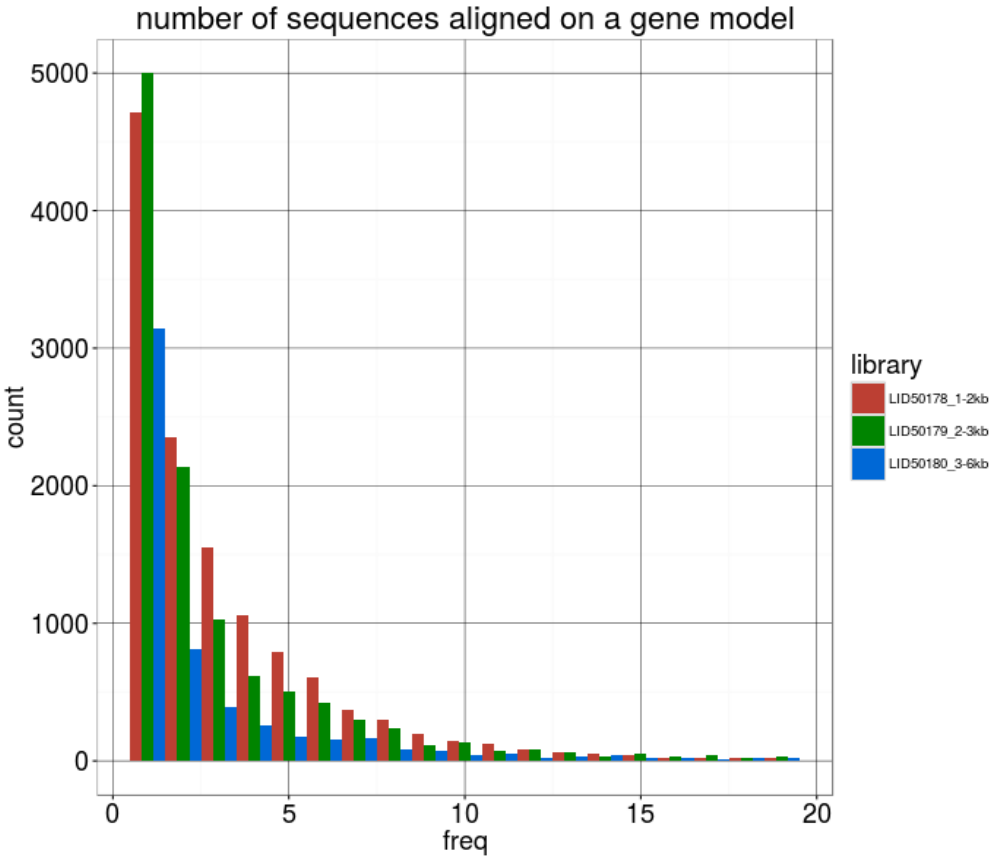
5,109 genes
Gene coverage 23,53 %

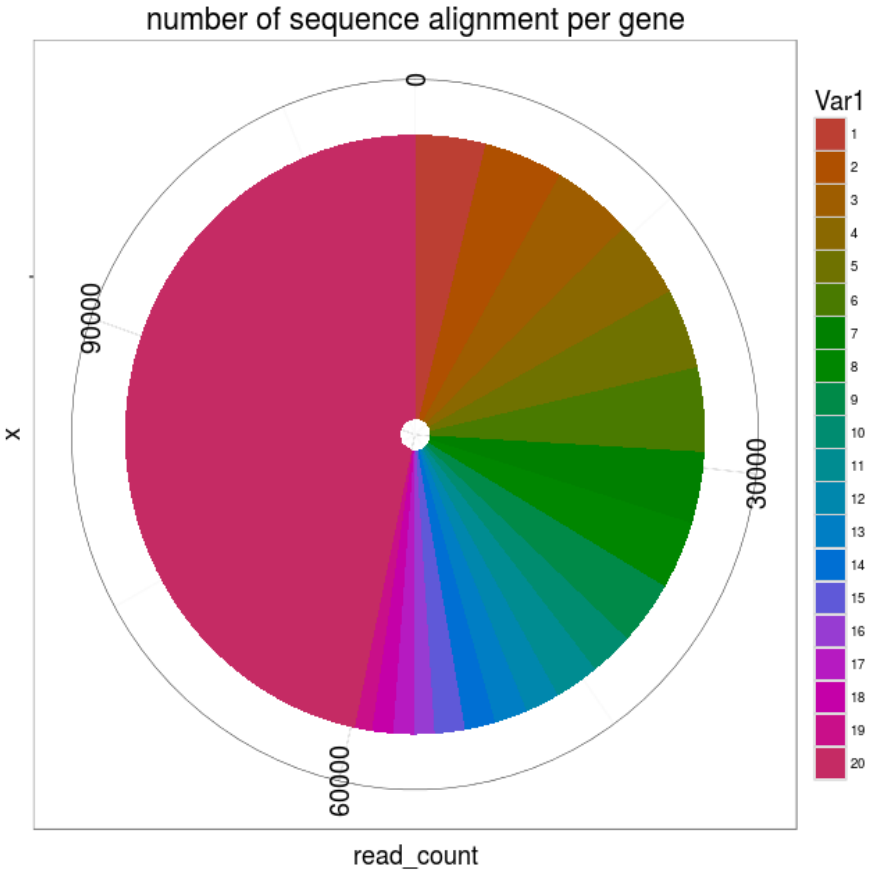
IsoSeq vs Illumina example



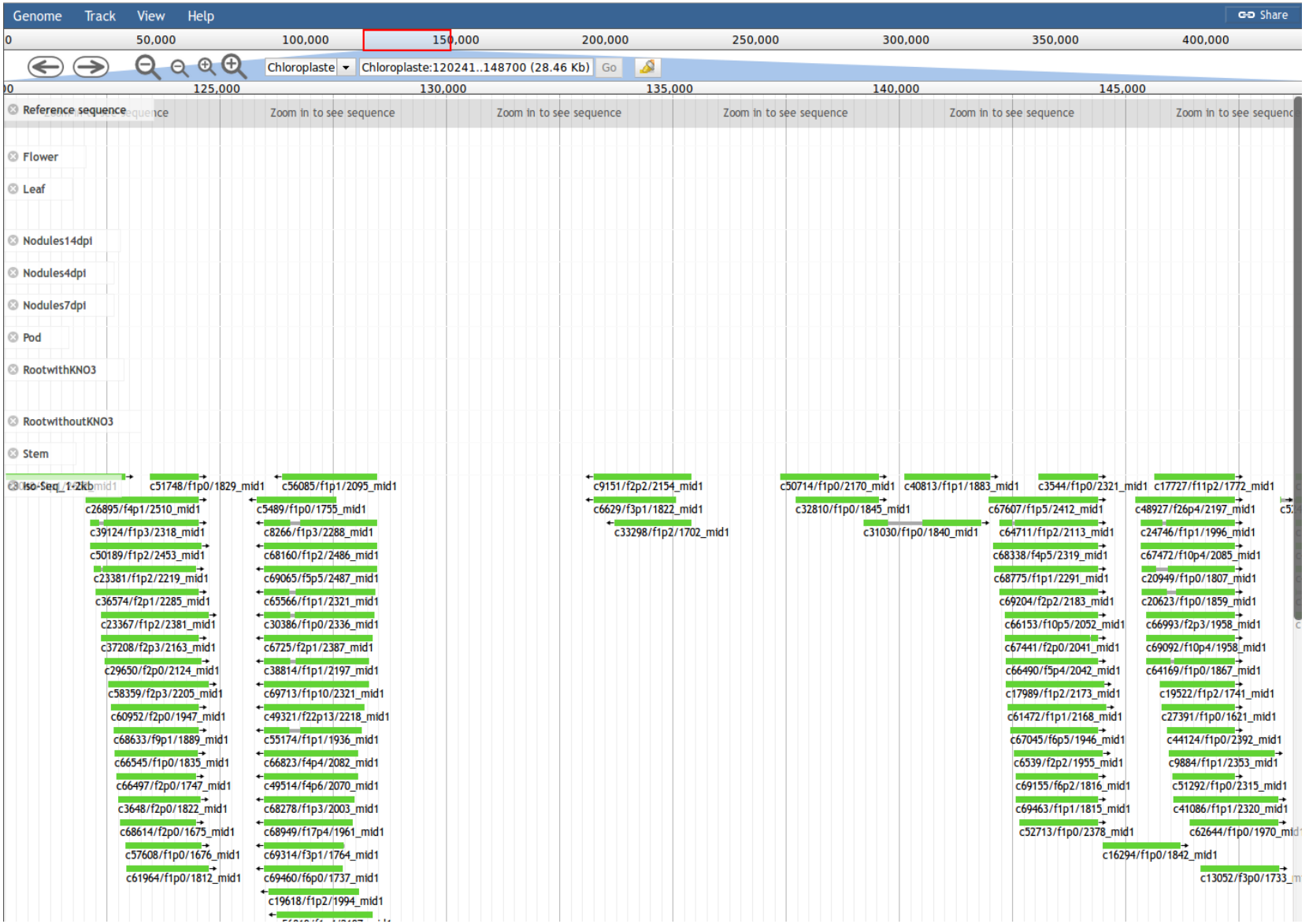
Inter-library duplication removal

`60,853 reads => 122,986 ovelaps between reads and model genes`





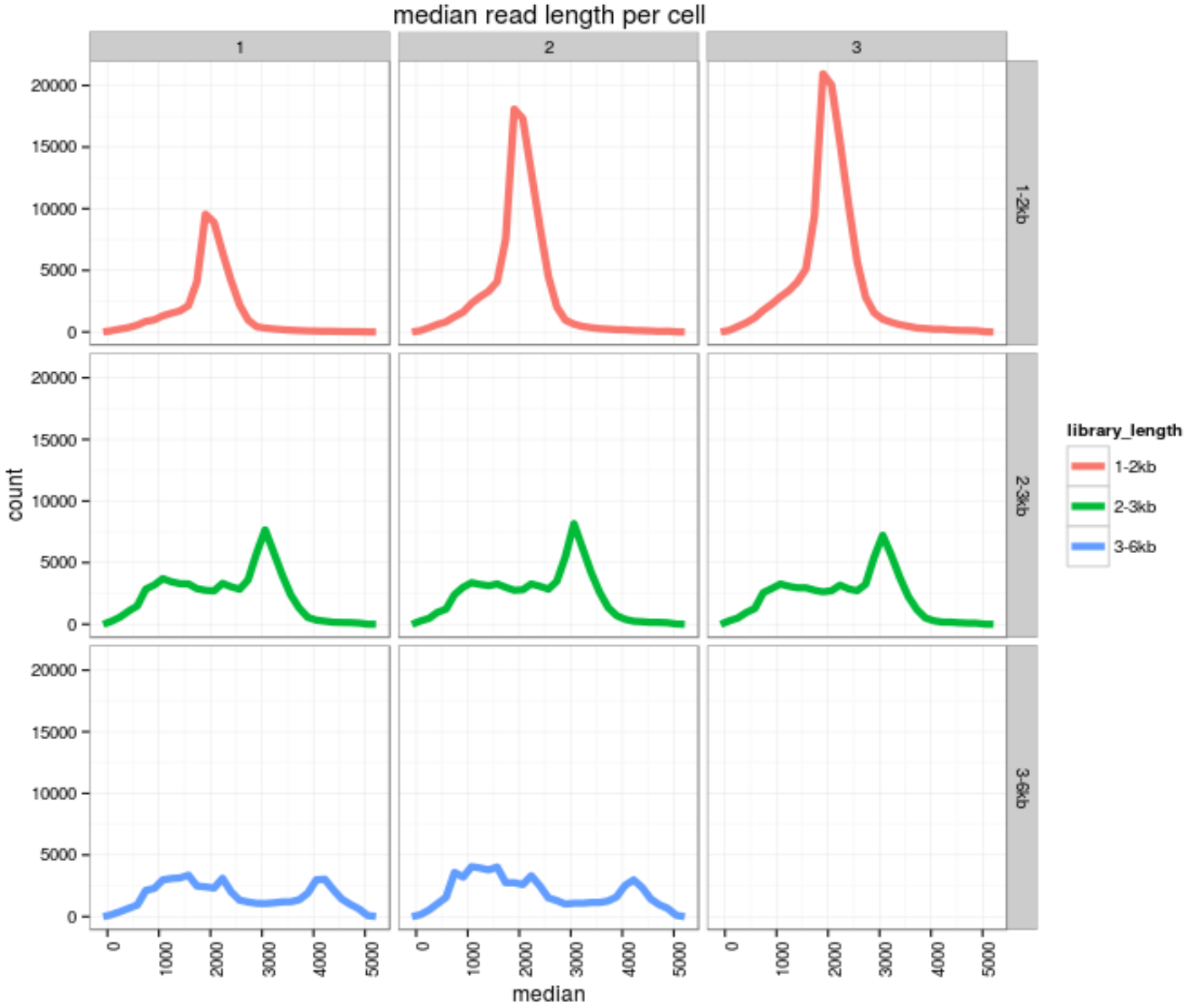
Not like Illumina! (chloroplast repeats)



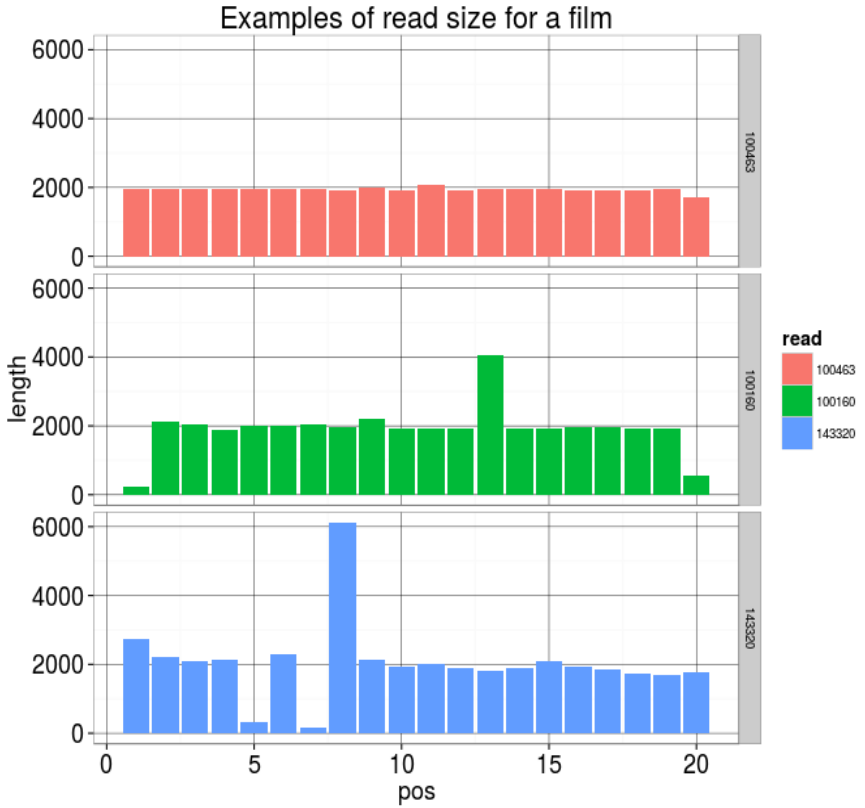
Detected problems

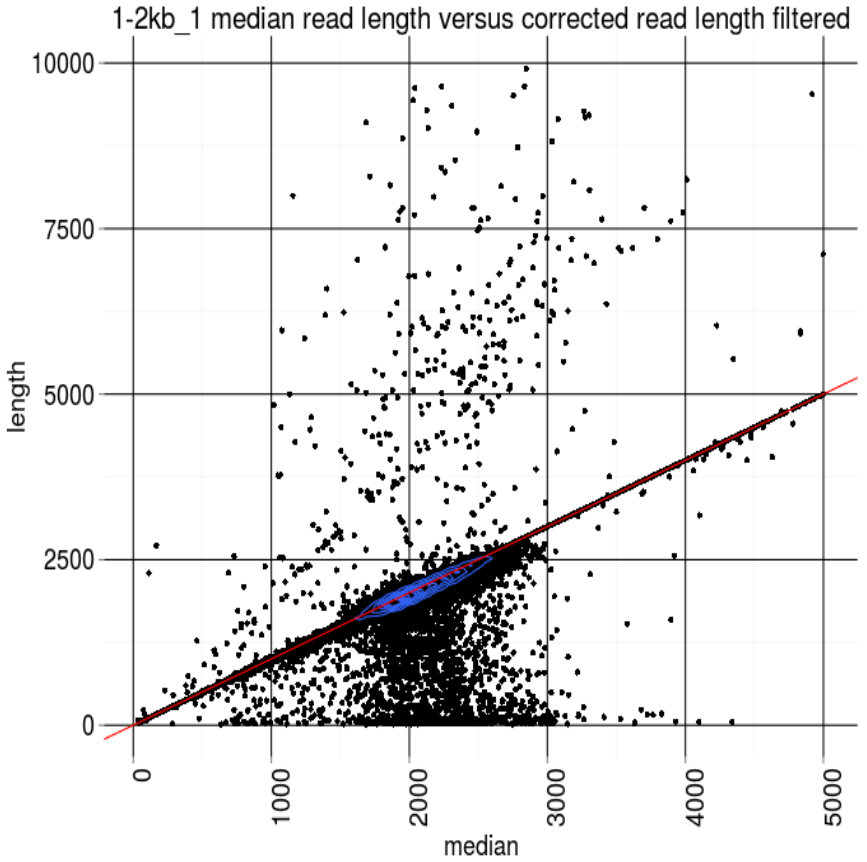
- Longer libraries have less full length transcripts.
- Film splitting is sometimes wrong
- RoI selection is sometimes faulty
- RoI production is biased

read length distributions

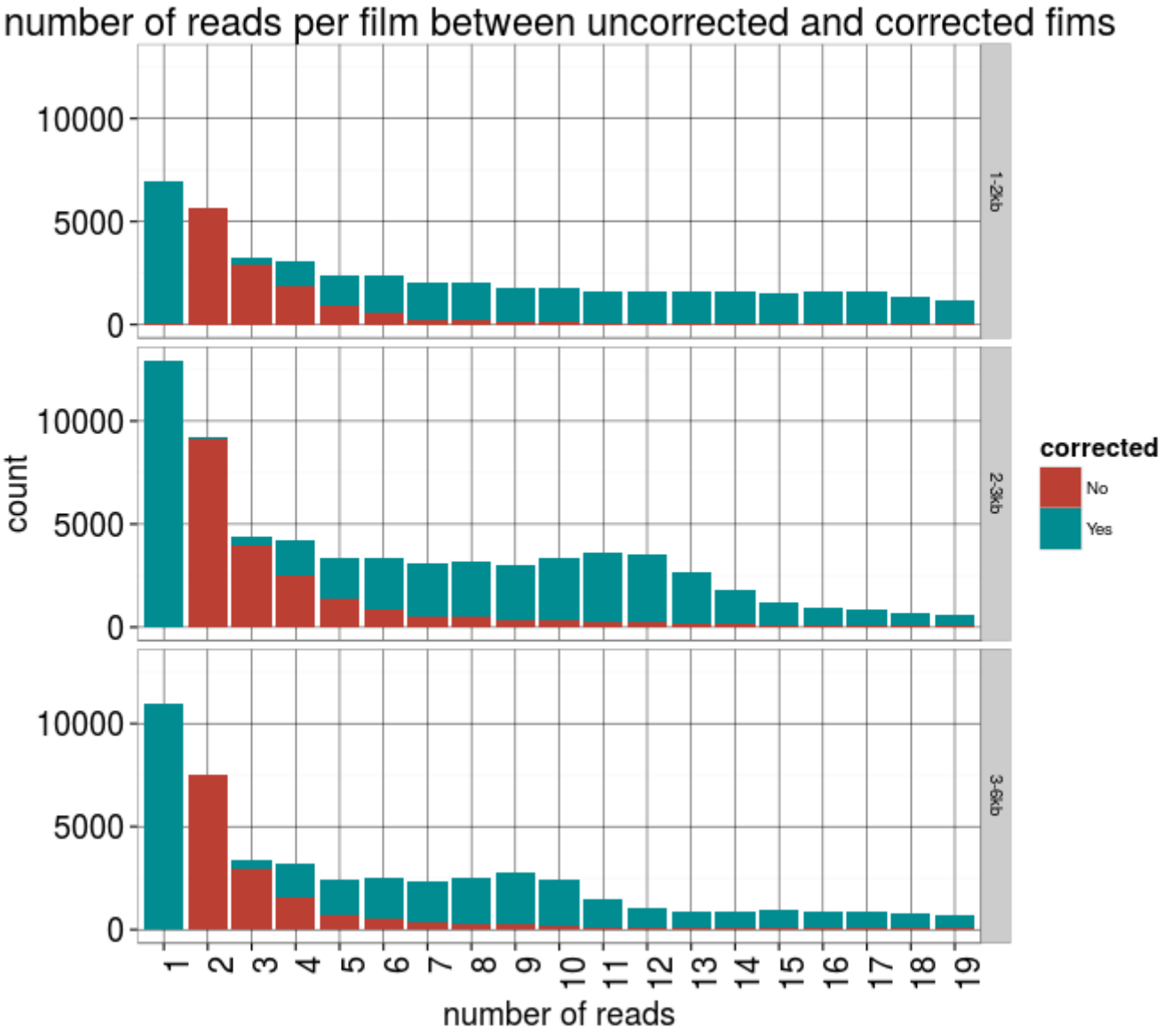


Film splitting and RoI selection





Number of reads per film vs RoI



Conclusions

- The Iso-Seq procedure works.
- It can be improved in different ways :
  - Better fragment sizing (preparation or filtering)
  - More films should produce RoI
  - RoI should be selected differently
- The gene coverage is not bad
- The produced isoforms have still to manually expertized
- We will reprocess the data with the new SMRT software version.