作业5
姓名：许芸阁
学号：2019012302
组别：第4组

1. 请问我们提供的bam文件COAD.ACTB.bam是单端测序分析的结果还是双端测序分析的结果?

`samtools flagstat COAD.ACTB.bam` 结果以及每一行的含义如下：

```
185650 + 0 in total (QC-passed reads + QC-failed reads)  #reads总数
4923 + 0 secondary  #出现比对到参考基因组多个位置的reads数
0 + 0 supplementary  #可能存在嵌合的reads数
0 + 0 duplicates  #属于PCR duplicates的reads数
185650 + 0 mapped (100.00% : N/A)  #比对到参考序列上的reads数
0 + 0 paired in sequencing  #属于PE read (paired-end read)的reads数
0 + 0 read1  #PE read中属于read1的reads数
0 + 0 read2  #PE read中属于read2的reads数
0 + 0 properly paired (N/A : N/A)  #PE read中完美比对的reads数
0 + 0 with itself and mate mapped  #PE read中，两端reads都比对上参考序列的reads数
0 + 0 singletons (N/A : N/A)  #PE read中，其中一端比上，另一端没比上的reads数
0 + 0 with mate mapped to a different chr  #PE read中，两端reads分别比对到不同序列的reads数
0 + 0 with mate mapped to a different chr (mapQ>=5)  #PE read中，两端reads分别比对到不同序列，且mapQ>=5的reads数
```

（注：加号两边的数字的含义——QC pass reads数 + QC fail reads数）

- COAD.ACTB.bam是是单端测序分析的结果，因为 `samtools flagstat COAD.ACTB.bam` 结果的第6行显示属于PE read (paired-end read)的reads数是0，且第7-13行与PE read相关的reads数都是0，因此不是双端测序（paired-end sequencing）分析的结果，是单端测序分析的结果。

补充——单端测序vs双端测序：

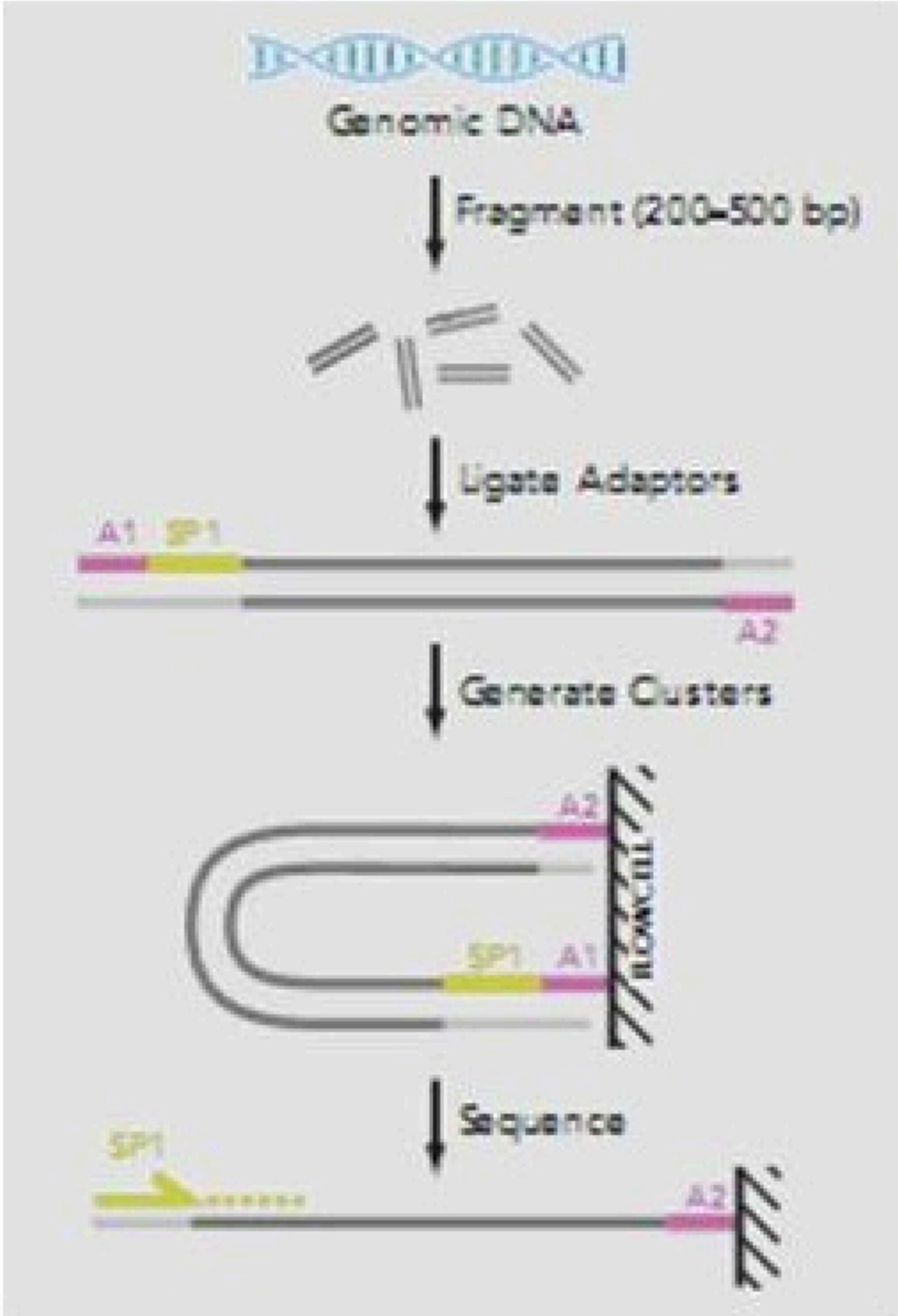


图1 Single-read文库构建方法

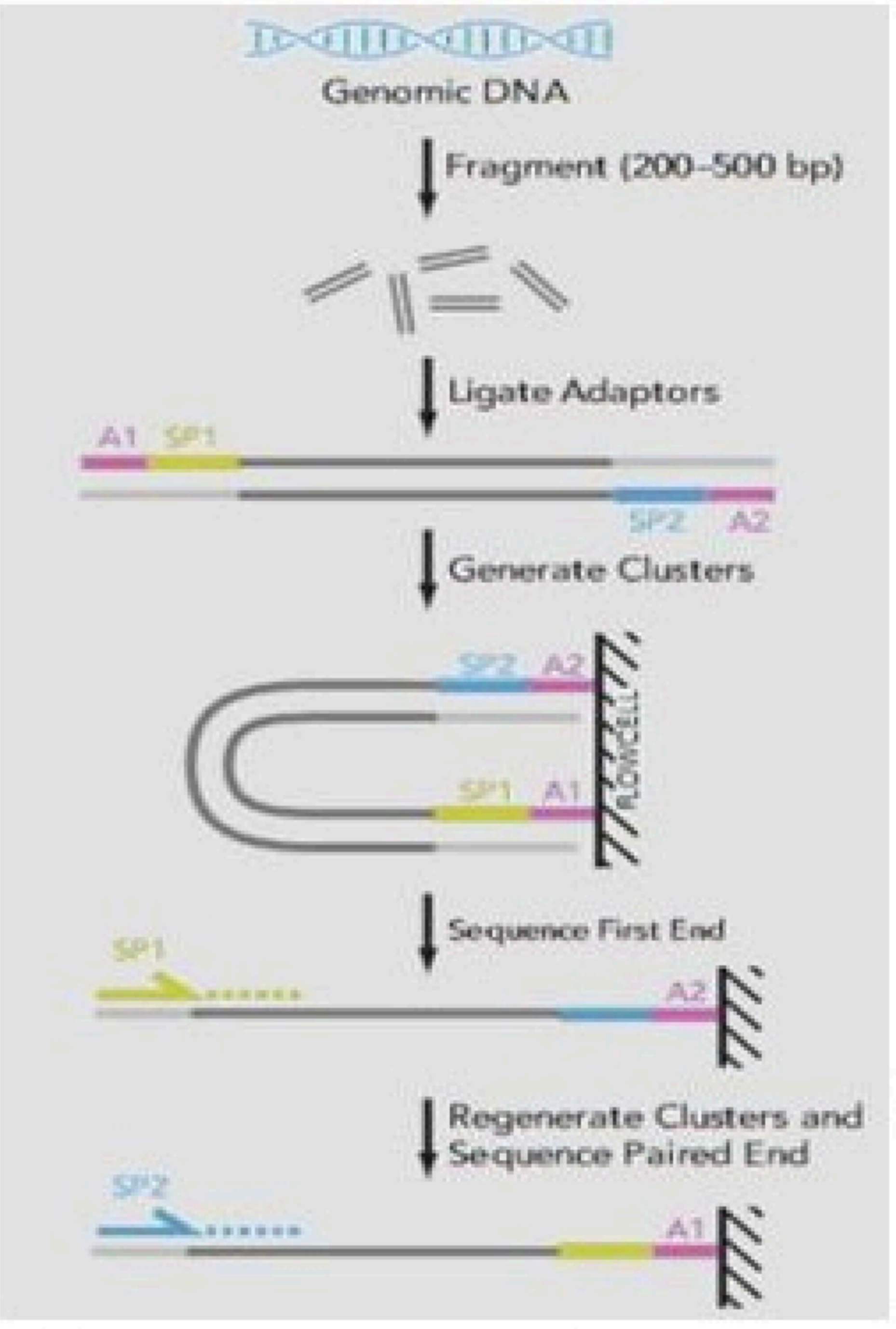


图2 Paired-end文库构建方法

- 单端测序（Single-read）首先将DNA样本进行片段化处理形成200-500bp的片段，引物序列连接到DNA片段的一端，然后末端加上接头，将片段固定在flow cell上生成DNA簇，上机测序单端读取序列。
- 双端测序（Paired-end）方法是指在构建待测DNA文库时在两端的接头上都加上测序引物结合位点，在第一轮测序完成后，去除第一轮测序的模板链，用对读测序模块(Paired-End Module)引导互补链在原位置再生和扩增，以达到第二轮测序所用的模板量，进行第二轮互补链的合成测序。

2.（1）查阅资料回答，什么叫做"secondary alignment"?

- Mapping时会存在multiple mapping的现象，即一条read可能比对到参考序列的多个位置，这种reads被称为multimapper。在一个multimapper能map到的多个位置中，只有一个位置处的记录属于primary alignment，map到其他位置的都属于secondary alignment。

2.（2）我们提供的bam文件中，有多少条记录属于"secondary alignment"?

- `samtools flagstat COAD.ACTB.bam` 结果的第2行显示，COAD.ACTB.bam中有4923条记录(4923 + 0)属于"secondary alignment"。
- 除此之外，还可以用samtools view来计算，代码如下：

```
samtools view -bf 256 COAD.ACTB.bam > COAD.ACTB.f.256.bam #只保留secondary
samtools view COAD.ACTB.f.256.bam > COAD.ACTB.f.256.sam
wc -l COAD.ACTB.f.256.sam  # 4923行
```

3.（1）根据hg38.ACTB.gff计算出在ACTB基因的每一条转录本中都被注释成intron的区域，以bed格式输出

- 使用的命令如下：

```
cat hg38.ACTB.gff | cut -f 1,3,4,5,6,7 | awk '{print $1, $3 - 1, $4, $2, $5, $6}' | sed 's/ /\t/g' > hg38.ACTB.bed
cat hg38.ACTB.bed | grep -w 'gene' > hg38.ACTB.gene.bed
cat hg38.ACTB.bed | grep -w 'exon' > hg38.ACTB.exon.bed
bedtools subtract -a hg38.ACTB.gene.bed -b hg38.ACTB.exon.bed > hg38.ACTB.intron.bed
```

- 得到的文件：hg38.ACTB.intron.bed
- 内容：

```
chr7    5528185 5528280 gene    .   -
chr7    5529982 5530523 gene    .   -
chr7    5530627 5540675 gene    .   -
chr7    5540771 5561851 gene    .   -
chr7    5561949 5562389 gene    .   -
chr7    5562828 5563713 gene    .   -
```

3.（2）利用COAD.ACTB.bam计算出reads在ACTB基因对应的genomic interval上的coverage，以bedgraph格式输出

- 使用的命令如下：

```
samtools sort -@ 7 COAD.ACTB.bam > COAD.ACTB.sorted.bam
samtools index COAD.ACTB.sorted.bam
bedtools genomecov -ibam COAD.ACTB.sorted.bam -bg -split > COAD.ACTB.coverage.bedgraph
```

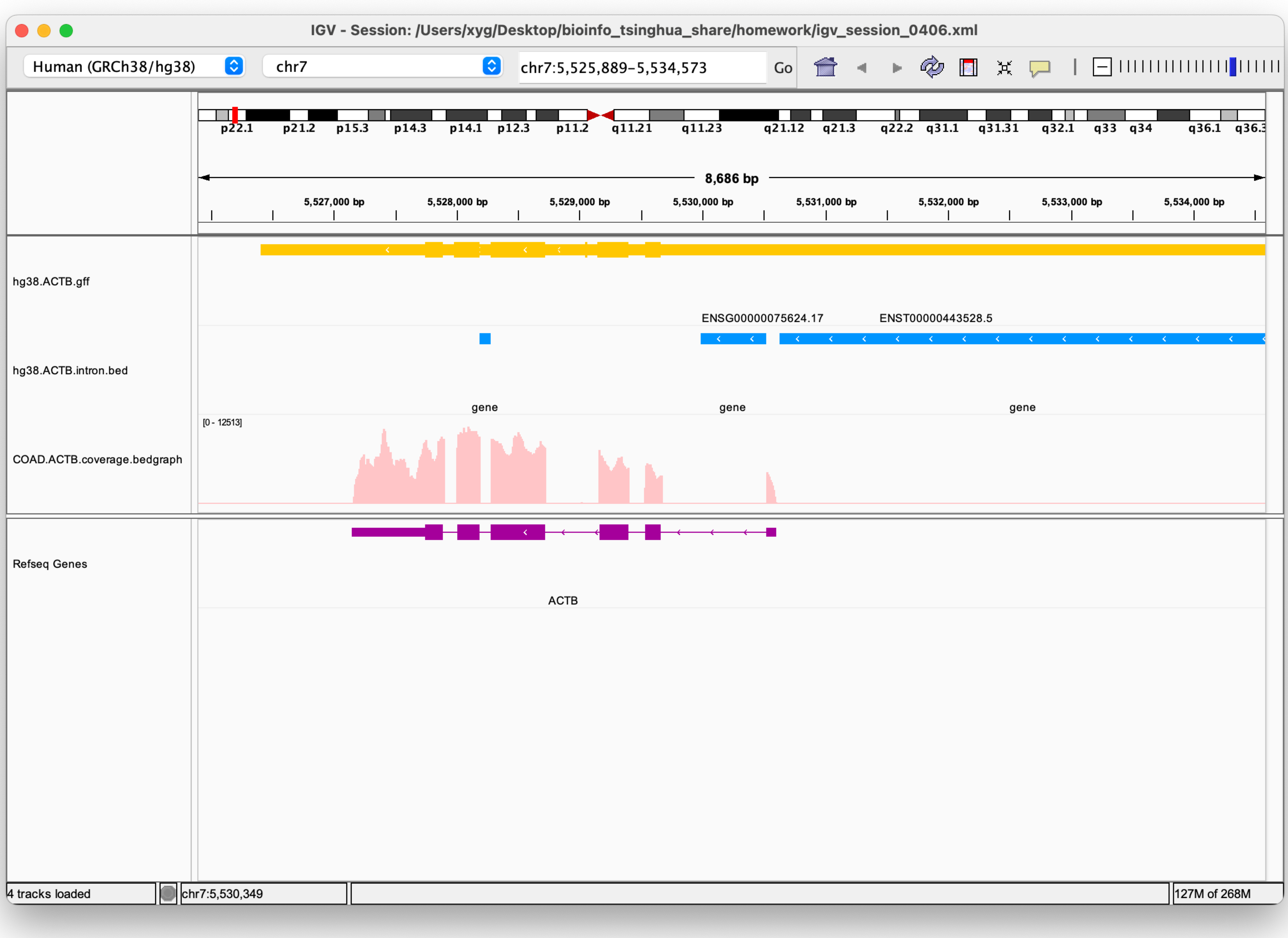
- 得到的文件：COAD.ACTB.coverage.bedgraph
- 前10行：

```
chr7    5045717 5045731 1
chr7    5058689 5058695 1
chr7    5072542 5072543 2
chr7    5072543 5072554 5
chr7    5073147 5073157 1
chr7    5077437 5077447 1
chr7    5080560 5080572 1
chr7    5118106 5118117 1
chr7    5121776 5121782 7
chr7    5121782 5121784 6
```

3.（3）将上述两个文件以及hg38.ACTB.gff文件load到IGV上进行可视化

- 得到的session：igv_session_0406.xml（用igv打开session时，session和它包含的三个文件需在同一路径下）
- 图片：

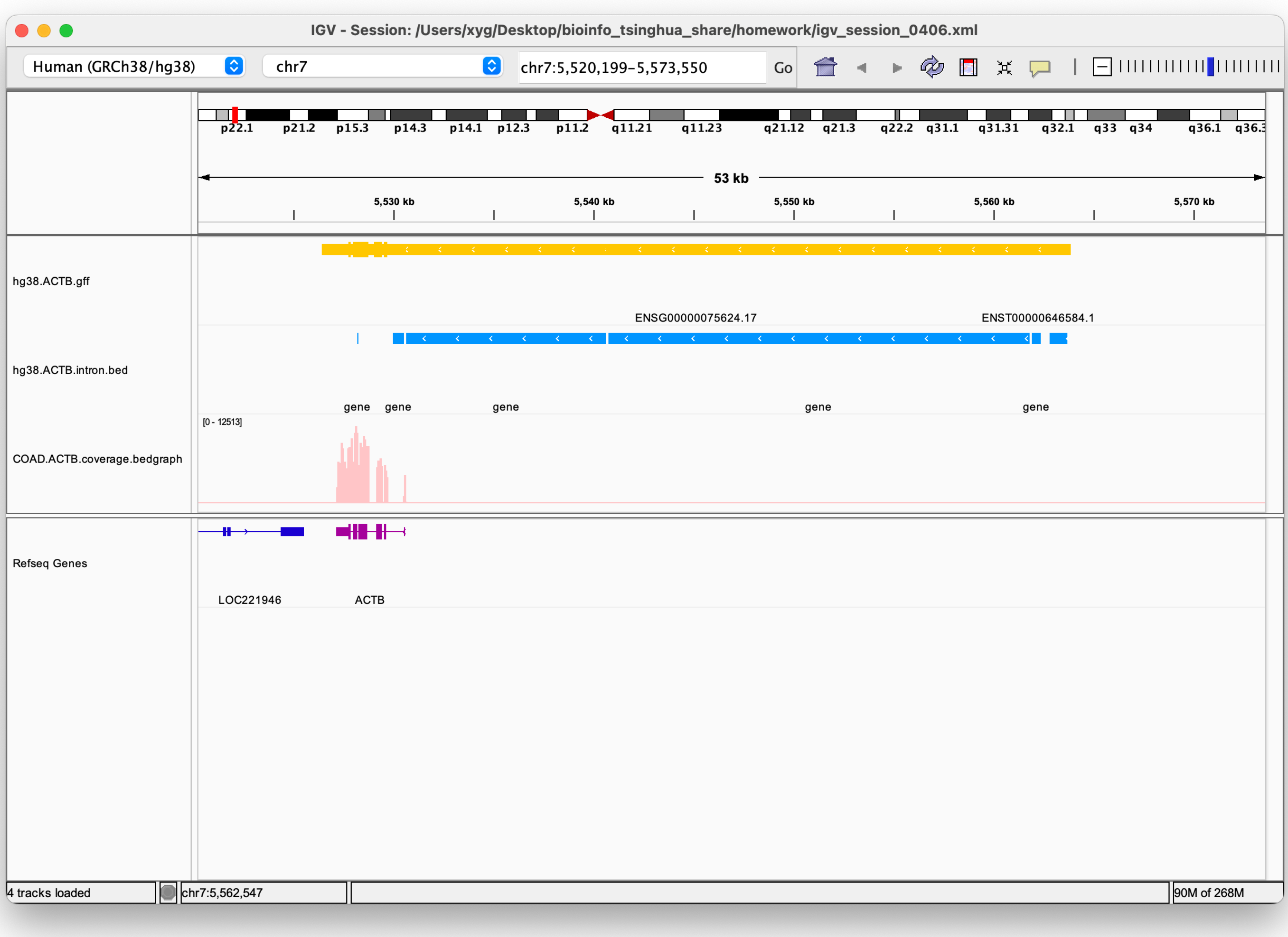
(1)局部放大


IGV - Session: /Users/xyg/Desktop/bioinfo_tsinghua_share/homework/igv_session_0406.xml

Human (GRCh38/hg38)　chr7　chr7:5,525,889-5,534,573　Go

4 tracks loaded　chr7:5,530,349　127M of 268M

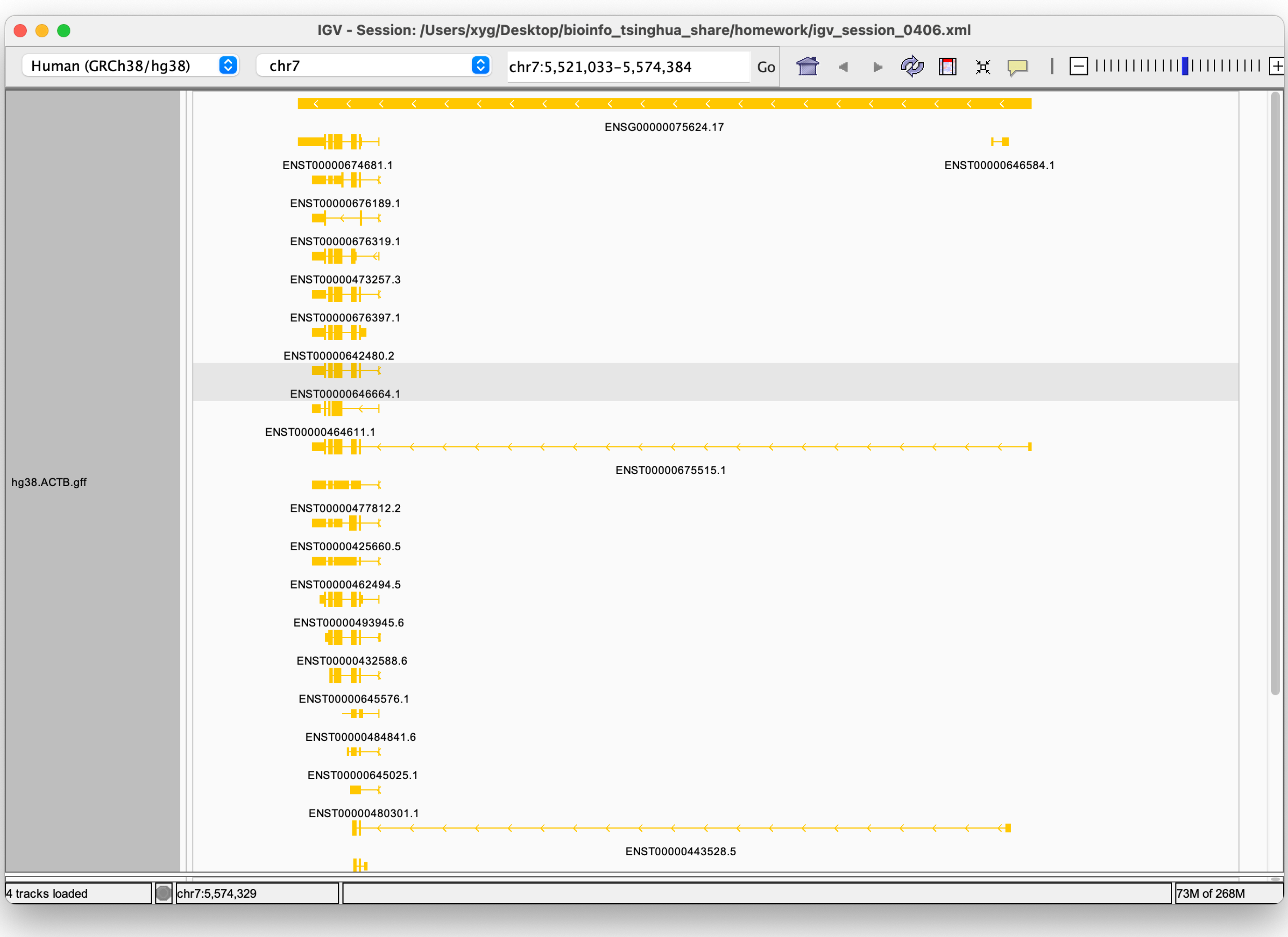
(2)整体


IGV - Session: /Users/xyg/Desktop/bioinfo_tsinghua_share/homework/igv_session_0406.xml

Human (GRCh38/hg38)　chr7　chr7:5,520,199-5,573,550　Go

4 tracks loaded　chr7:5,562,547　90M of 268M

(3)hg38.ACTB.gff collapsed -> expanded


IGV - Session: /Users/xyg/Desktop/bioinfo_tsinghua_share/homework/igv_session_0406.xml

Human (GRCh38/hg38)　chr7　chr7:5,521,033-5,574,384　Go

4 tracks loaded　chr7:5,574,329　73M of 268M