lab seminar

Session1

JUN SOUNG KWAK

FASTA files

```
Junsoungs-Mac-mini:~ jun$ head /Volumes/Jun_SSD/genome/Paralichthys_olivaceus/Ensembl_mid/genome/Paralichthys_olivaceus-GCA_024713975.2-softmasked.fa
>1 softmasked:chromosome primary_assembly:ASM2471397v2:1:1:31859138:1
CGAACTAGGTGGATGGGCTTTCTGTAAAGACGGCTCCTGTCGCGGTTTCAAAGCTCTTCG
TGCTTTCTGGCATTTGCTGGAGGCCCGTCAAATTGCAATCTATAATGAAAGGGGAATCAG
AAGTTAGTGTAAAGCCCATTATCAATGCCTAGTATATGCTGGCCCCTCATTTCAGCCAAA
AAAGCCTCAAACTGTTTGAAAGTTACACCTTTGCGATCATTAGGCCCCCCGATCAACATT
TTTTCCCTCTCATTACCCGTATGTACTCTGTATGGCTATGAAAATGGAGCCCTTAGAAAT
GGGCTTAATTGATACCAAAGGTCACAAAGGTCACCTGGGCCCTTCTGCGGTTTTGGTGAA
AATCGGTAATGCAGGCTCAAAGAGCCCCTTGTGACCAGTGGGTCAAATTTGGGGCTCCTA
GGCCCCTGGGAAGGCCCCGAAAGGGATCGTAATTTTCCCAAAAATCTCCAGAACCAAGAG
GCTGCGAGCCCTGTTTAGGCCCCTGACCAGCTAGAGGGCTGAAATCCAAGTGACACCTCA
```

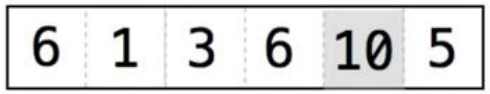
header

sequence

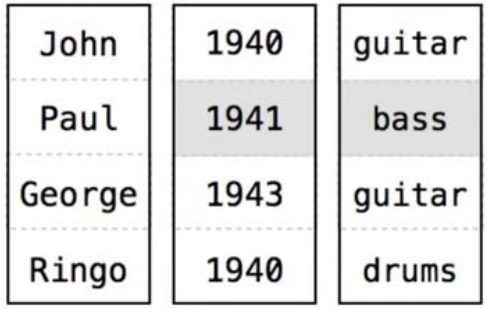
FASTA file is a text file that contains a nucleotide sequence of DNA or RNA or an amino acid sequence of a protein and related information.

Extension	Meaning	Notes
fasta, fa	generic FASTA	Any generic fasta file
fna	FASTA nucleic acid	Used generically to specify nucleic acids
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome
faa	FASTA amino acid	Contains amino acid sequences
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome

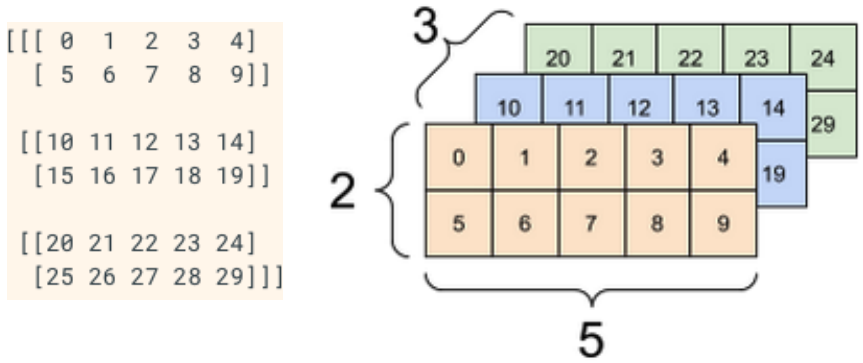
Index (Indexing): 색인, 목록



vec[5]



df[2, c(2,3)]



3D Array (Tensor)

Index notation is a standard method used to specify or reference individual elements of an array, vector, matrix, or tensor in mathematics, computer science, and related fields.
It provides a systematic way to describe specific elements within these data structures.

1D Array (Vector)
2D Array (Matrix, data frame)
3D Array (Tensor)

FASTA Index

: 유전자 데이터를 효율적으로 저장하고 검색하기 위한 기술.

```
Junsoungs-Mac-mini:~ jun$ head -n 50 /Volumes/Jun_SSD/genome/Paralichthys_olivaceus/Ensembl_rapid/genome/Paralichthys_olivaceus-GCA_024713975.2-softmasked.fa.fai
1       31859138        70      60      61
2       29887412        32390264        60      61
3       29078691        62775870        60      61
4       28192728        92339276        60      61
5       28185977        121001953       60      61
6       28085797        149657767       60      61
7       27885095        178211731       60      61
8       27660206        206561648       60      61
9       27364311        234682928       60      61
10      26501863        262503383       60      61
11      26045262        289447016       60      61
12      25908411        315926438       60      61
13      25560671        342266728       60      61
14      24856284        368253483       60      61
15      24354382        393524111       60      61
16      23785052        418284472       60      61
17      21261208        442466014       60      61
18      20994079        464081648       60      61
19      20610995        485425701       60      61
20      20325903        506380285       60      61
21      19848577        527045026       60      61
22      17781304        547224485       60      61
23      17387852        565302217       60      61
24      14749316        582979939       60      61
MT      17090   597975146       60      61
```

( )D array (,)

1st column: 염색체 이름, 또는 시퀀스 이름
2nd column: 염색체 또는 시퀀스의 길이 (염기 수)
3rd column: 해당 염색체 또는 시퀀스가 FASTA file 내에서 시작되는 바이트 위치
4th column: FASTA file에서 한 줄의 염기 서열에 포함된 문자의 개수 (보통 60으로 설정)
5th column: FASTA file에서 한 줄의 염기 서열과 줄 바꿈을 포함한 길이 (보통 61로 설정, 60 (염기서열) + 1 (줄바꿈)

```
Junsoungs-Mac-mini:~ jun$ head /Volumes/Jun_SSD/genome/Paralichthys_olivaceus/Ensembl_rapid/genome/Paralichthys_olivaceus-GCA_024713975.2-softmasked.fa
>1 softmasked:chromosome primary_assembly:ASM2471397v2:1:1:31859138:1
CGAACTAGGTGGATGGGCTTTCTGTAAAGACGGCTCCTGTCGCGGTTTCAAAGCTCTTCG
TGCTTTCTGGCATTTGCTGGAGGCCCGTCAAATTGCAATCTATAATGAAAGGGGAATCAG
AAGTTAGTGTAAAGCCCATTATCAATGCCTAGTATATGCTGGCCCCTCATTTCAGCCAAA
AAAGCCTCAAACTGTTTGAAAGTTACACCTTTGCGATCATTAGGCCCCCCGATCAACATT
TTTTCCCTCTCATTACCCGTATGTACTCTGTATGGCTATGAAAATGGAGCCCTTAGAAAT
GGGCTTAATTGATACCAAAGGTCACAAAGGTCACCTGGGCCCTTCTGCGGTTTTGGTGAA
AATCGGTAATGCAGGCTCAAAGAGCCCCTTGTGACCAGTGGGTCAAATTTGGGGCTCCTA
GGCCCCTGGGAAGGCCCCGAAAGGGATCGTAATTTTCCCAAAAATCTCCAGAACCAAGAG
GCTGCGAGCCCTGTTTAGGCCCCTGACCAGCTAGAGGGCTGAAATCCAAGTGACACCTCA
```

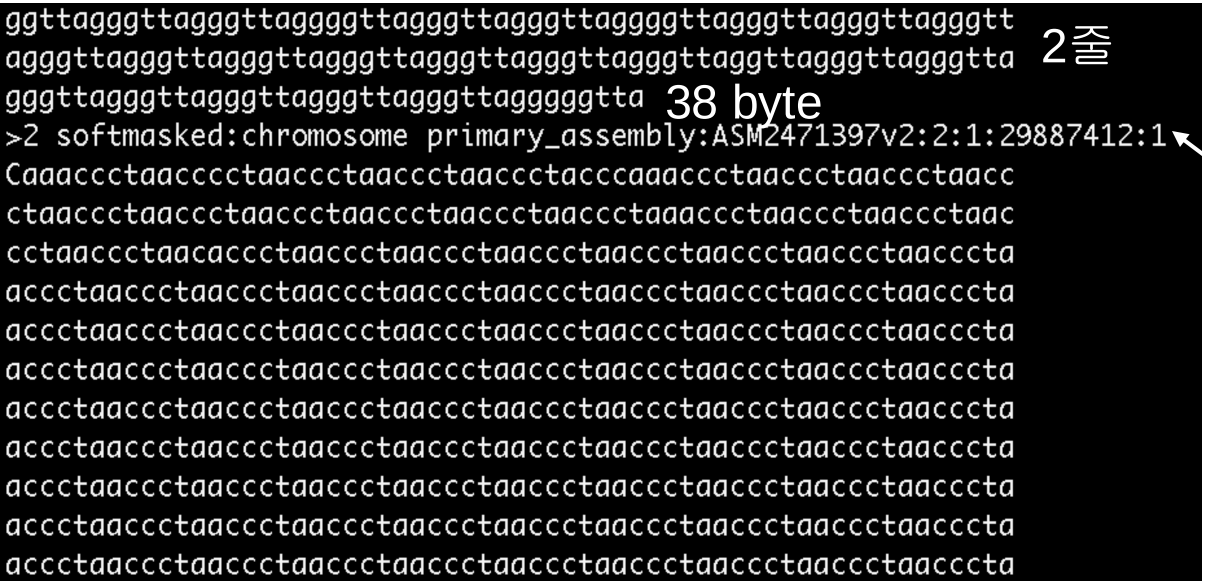
69 byte

(줄 바뀜은 1 byte)

9줄

⋮

몇 줄이 있을까?



```
ggttagggttagggttaggggttagggttagggttaggggttagggttagggttagggtt
agggttagggttagggttagggttagggttagggttagggttaggttagggttagggtta
gggttagggttagggttagggttagggggtta
>2 softmasked:chromosome primary_assembly:ASM2471397v2:2:1:29887412:1
Caaaccctaaccccctaaccctaaccctaaccctacccaaaccctaaccctaaccctaacc
ctaaccctaaccctaaccctaaccctaaccctaaccctaaaccctaaccctaaccctaac
cctaaccctaacaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
```

2줄

38 byte

69 byte

(69 byte + 1 byte (line break)) +
((전체 line #) x 61 byte (60 byte + 1byte(line break) +
(38 byte + 1 byte (line break)) +
(69 byte + 1 byte (line break))
=32390264

Easy indexing without coding → Load FASTA file on IGV
Other ways → Samtools, bowtie ...

GFF / GTF format

GFF3

```
1	ASM2471397v2	region	1	31859138	.	.	.	ID=region:1;Alias=CM045671.1
###
1	ensembl gene	30565	32060	.	-	.	ID=gene:ENSPOLG00010000718;biotype=protein_coding;gene_id=ENSPOLG00010000718;version=1
1	ensembl mRNA	30565	32060	.	-	.	ID=transcript:ENSPOLT00010001876;Parent=gene:ENSPOLG00010000718;biotype=protein_coding;tag=Ensembl_canonical;transcript_id=ENSPOLT00010001876;version=1
1	ensembl exon	30565	30740	.	-	.	Parent=transcript:ENSPOLT00010001876;constitutive=1;exon_id=ENSPOLE00010009024;rank=2;version=1
1	ensembl three_prime_UTR 30565	30740	.	-	.	Parent=transcript:ENSPOLT00010001876
1	ensembl three_prime_UTR 30829	31048	.	-	.	Parent=transcript:ENSPOLT00010001876
1	ensembl exon	30829	32060	.	-	.	Parent=transcript:ENSPOLT00010001876;constitutive=1;exon_id=ENSPOLE00010009023;rank=1;version=1
1	ensembl CDS 31049	31696	.	-	0	ID=CDS:ENSPOLP00010001795;Parent=transcript:ENSPOLT00010001876;protein_id=ENSPOLP00010001795;version=1
1	ensembl five_prime_UTR	31697	32060	.	-	.	Parent=transcript:ENSPOLT00010001876
```

GTF

```
1	ensembl gene	30565	32060	.	-	.	gene_id "ENSPOLG00010000718"; gene_version "1"; gene_source "ensembl"; gene_biotype "protein_coding";
1	ensembl transcript	30565	32060	.	-	.	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_source
"ensembl"; transcript_biotype "protein_coding"; tag "Ensembl_canonical";
1	ensembl exon	30829	32060	.	-	.	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; exon_number "1"; gene_source "ensembl"; gene_biotype "protein_coding";
transcript_source "ensembl"; transcript_biotype "protein_coding"; exon_id "ENSPOLE00010009023"; exon_version "1"; tag "Ensembl_canonical";
1	ensembl CDS 31052	31696	.	-	0	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; exon_number "1"; gene_source "ensembl"; gene_biotype "protein_coding";
transcript_source "ensembl"; transcript_biotype "protein_coding"; protein_id "ENSPOLP00010001795"; protein_version "1"; tag "Ensembl_canonical";
1	ensembl start_codon 31694	31696	.	-	0	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; exon_number "1"; gene_source "ensembl"; gene_biotype "protein_coding";
transcript_source "ensembl"; transcript_biotype "protein_coding"; tag "Ensembl_canonical";
1	ensembl stop_codon	31049	31051	.	-	0	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; exon_number "1"; gene_source "ensembl"; gene_biotype "protein_coding";
transcript_source "ensembl"; transcript_biotype "protein_coding"; tag "Ensembl_canonical";
1	ensembl exon	30565	30740	.	-	.	gene_id "ENSPOLG00010000718"; gene_version "1"; transcript_id "ENSPOLT00010001876"; transcript_version "1"; exon_number "2"; gene_source "ensembl"; gene_biotype "protein_coding";
transcript_source "ensembl"; transcript_biotype "protein_coding"; exon_id "ENSPOLE00010009024"; exon_version "1"; tag "Ensembl_canonical";
```

GFF는 일반적인 유전체 annotation 형식인 반면, GTF는 엄격하게 유전자 annotation 형식임.
GFF는 덜 제어됨으로 더 많은 변수를 필요로 함.
GTF ⟶ GFF 변환은 쉬움 / 필요한 gene/transcript_id 속성을 GFF안의 ID 속성으로 변환하면 됨.
GFF ⟶ GTF 변환은

SAM/BAM file

Sequence alignment and map (SAM) or binary alignment and map (BAM) file
NGS 데이터를 정렬한 결과를 저장하는 파일 형식.
SAM file은 사람이 읽을 수 있지만, BAM은 사람이 이해할 수 없음 - 컴퓨터만 이해함.

SAM file 구조 –
1. Header (헤더)

@HD VN:1.6 SO:coordinate
@SQ SN:chr1 LN:248956422
@SQ SN:chr2 LN:242193529

2. Alignment section (정렬 데이터)

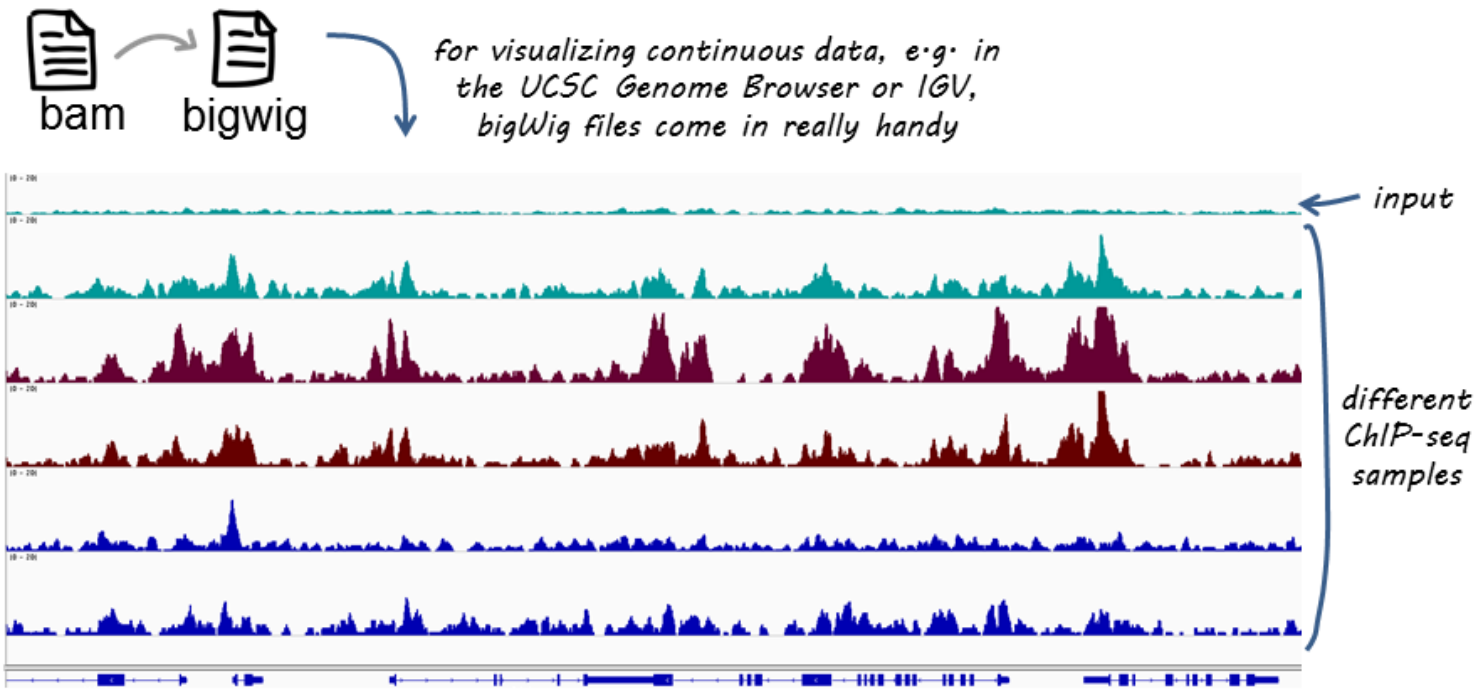
쌍을 이루는 리드

1	2	3	4	5	6	7	8	9	10	11	12
리드 이름	FLAG	참조 서열 이름	정렬 시작 위치	매핑 품질 점수	CIGAR 문자열	정렬된 참조 서열 이름	정렬된 위치	길이	염기서열	품질 기호	사전정의 태그
SRR067577.2766	99	chr1	73240003	60	101M	=	73240004	102	GCTA……..	FHG@…	NM:I:0

jun — -bash — 97×46

```
Junsoungs-Mac-mini:~ jun$ head /Volumes/Scratch/_nf-core/RNA_seq/2024_SMU_KYJ/2_Olive_flounder/Ol
ive_flounder_RNA_seq/star_salmon/C1.markdup.sorted.bam
?BCG)?{?$?}?{?3?H??h&U??ɔ?qm???????9
                                        D?j?ɹw???=?A??ɯ????
                                                                ?c?Eq?@?22?J<%?*ɟŶ?#U??JI???R????k?????ɹ
?}?f{~??;??~?????{u[??{)j???1?b'??
                                    ?????O?o?n??\???2?8p|?ɥ?????rXę?"
                                                                        ,??v!#+?"??U??*'s??E?n*?n?d2?(
?^?lʊ"2??*v?$1#???]?K?z??e?RIdXE?R?.e?????ɟUF|w?k%?#ƀR?Z?"J?"x?v?^xep_?b?
                                                                                u?v?XNb9???UcTU`d?+???H,
??(pv???W7Ï*??l9?n?,+,?x1??R^???}3ÏÂʳ2?-{?f??)?????\??/w????OP)nB?`?{Loɛyn?撰Vo?g}?f????r?j???
?6:??_[=5_e????????SU?*?wmc?㬰~????z?pl?????Fu}?··m??)ve???¦k?ⱴƱ???2q?Cp?u粉???+????<{vu?T??????
??
  ?xn~????????V??       ?6???ʔmm?Nllm?ɬ???(\??}?yl}??)?B????6??O??X?wv??????,?h??????Ꝟ7?-fáo?U??
??.n?Y??j?r?g?93qɹ?ɟ?{???????X?Uoo>?v??U{?????uj?ç??;
                          ?s?ʔ????????\>     _?&??????5ɡ?NYot??ʔu?????IW??V?non??B??N??????
??
t6NF[???|?O?????V?ȩ8wf??????X;??????'?  ?k/?_7;?x???x??,|?ɷs??sẅ?ɬ?c?ğc.
????k??'?&???]??[?czE??x|vJ?L/WO???Z?3????入δ^?>
                                                ;s?+??Ru??9xw?v|V?:?
                                                                  ?+???g?????????X7?6?ep???O?A"
ţsʍ?UP??%?=ɹ?,Mk??????2?-VEĹk?bh??q??ȯ}%~}?b+S
                                              ĬOO?_u???%??Uf??V?|?????\{zi???|A?a
                                                                                1??
                                                                                   ?Z\?3wU?M???{
??q)?g??\917?}???_X?[Ч?%??ĝ?V?O???????]qby?2;?rb~Q?]??ˢ+?SS?g?-D?
                                                                  ⸗??~????w?2?7X?mvX/-wV?W?dyp???
????~a>pS????c??u?K????q??"tM?WgV?}g???>>??bj?jPw?pl??-??????ъ?K
                                                                  3K??>H?jĶ?Kf?N,??ks??k,?Sp???;??-
??????/?\=?}??Zu?t?
                   ?l??Z???Mɥ?s??W?w???`???Uɜpv
M??T;z)?$?m?g?cz???
?"'??????S8[?ʍ???wq??[;s?{?ʔ?[?m~?X?ujm#???급??ă??
           .l????t?rl'[??P$e?1?"?p??O?i}?;
m5?xqLRh?!Gb??@?qD=??O??????????z???????/?n??S??:????ʔH?????D???/XG2u?ä???=c????ώ??;3?????!K???k?
!G}???!O?\??}(Pz?a?P?????}(Q???C????u?P?n?s?>T??>θ9?:t?d?T?{?!G)K??C?:~??P??????s??m??a????ɔf??I?
]D?     ???Pè??YM:??9?/?ɑîdSN?$
??p??????F????e????z?G???rϑ???R?^*?{?f??(???~]??5z<N*??C˜~??i?L-???????S??2N^ṃ??;9?-O?t?v4j???s%"
ɹ??O\????A???^????A???σ?w??+?U~D?`l???.Rk?ɜ????Y}?r????JQ?k?ɬ{f?\ojZ??F??+5C?5ɜ"ɜЧi?}?6{f?_??jZ=
G"K?51?X
?1??????zjЧ?ʢ??zd8X???{/
?W?籾??????ʔ???{Y?      X{??:wlHXt?ek_XwL
??f????t????J
            8Vb*öxA?y??Z?ʔ?Fb?z?^??????ч??.^?{?ZM3{4],?z%?V3?i?"iF?I?????~Bo??(???Nɬ???%?XAn?"?
?*?c?l@_?[?].?e?[+?p??f?YF?z?T6k4]3?5?X,?t?W,=?V??B?ʌ$BW9?( ?B??}?<E??  ?%?G?y??^??+?????~V?K????
???%tU"???~?GP6G?C"ꝑTUR?g??J%r?J:?1'YAB?BB???1?o?????DH[?r-Q?????B*逮??F ???
```

Bigwig file



remember that there are 2 deepTools for bam → bigWig conversion:

- ***bamCoverage:*** for individual files (like those shown here)
- ***bamCompare:*** to normalize two files to each other

BAM file 을 시각화 하기 위해 변환한 **data** 형식으로 마찬가지로 **binary** 형태 **–** 인간이 이해 불가**.**

JBrowse 2

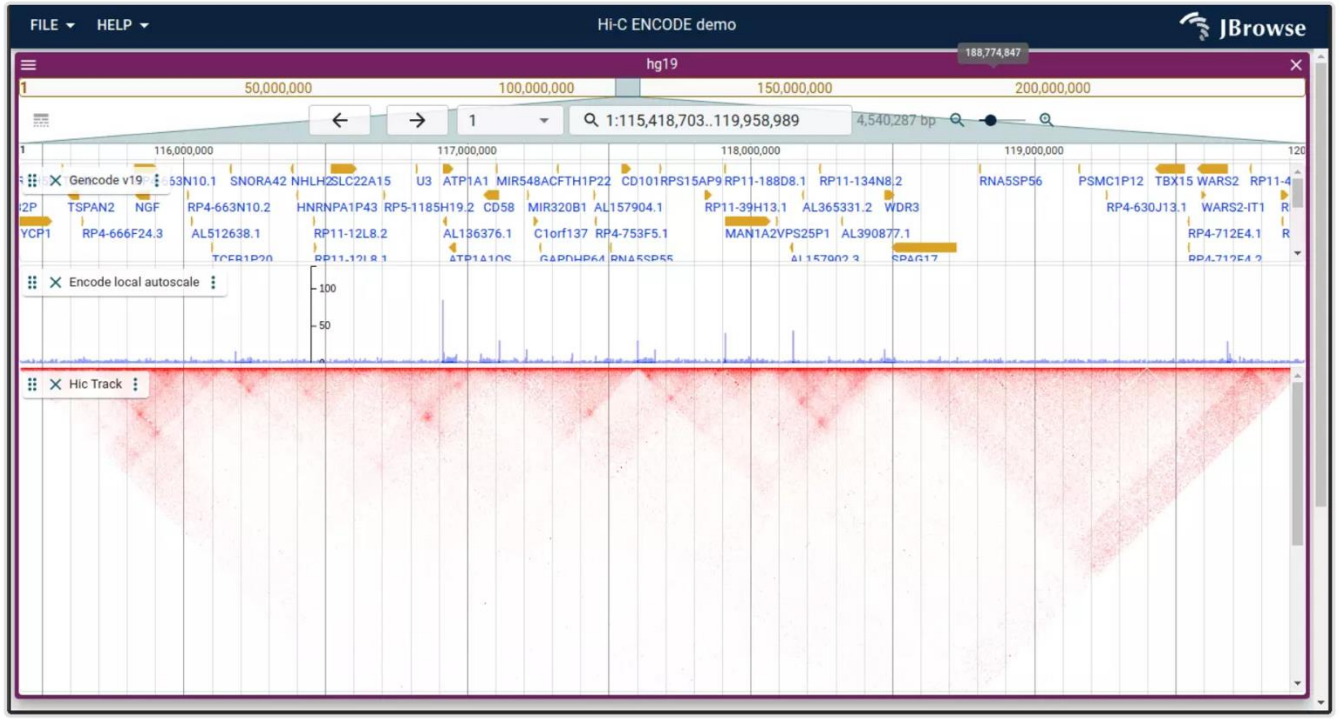
JBrowse 2: a modular genome browser with views of synteny and structural variation

Colin Diesh, Garrett J Stevens, Peter Xie, Teresa De Jesus Martinez, Elliot A. Hershberg, Angel Leung, Emma Guo, Shihab Dider, Junjun Zhang, Caroline Bridge, Gregory Hogue, Andrew Duncan, Matthew Morgan, Tia Flores, Benjamin N. Bimber, Robin Haw, Scott Cain, Robert M. Buels, Lincoln D. Stein & Ian H. Holmes

Genome Biology **24**, Article number: 74 (2023) | Cite this article

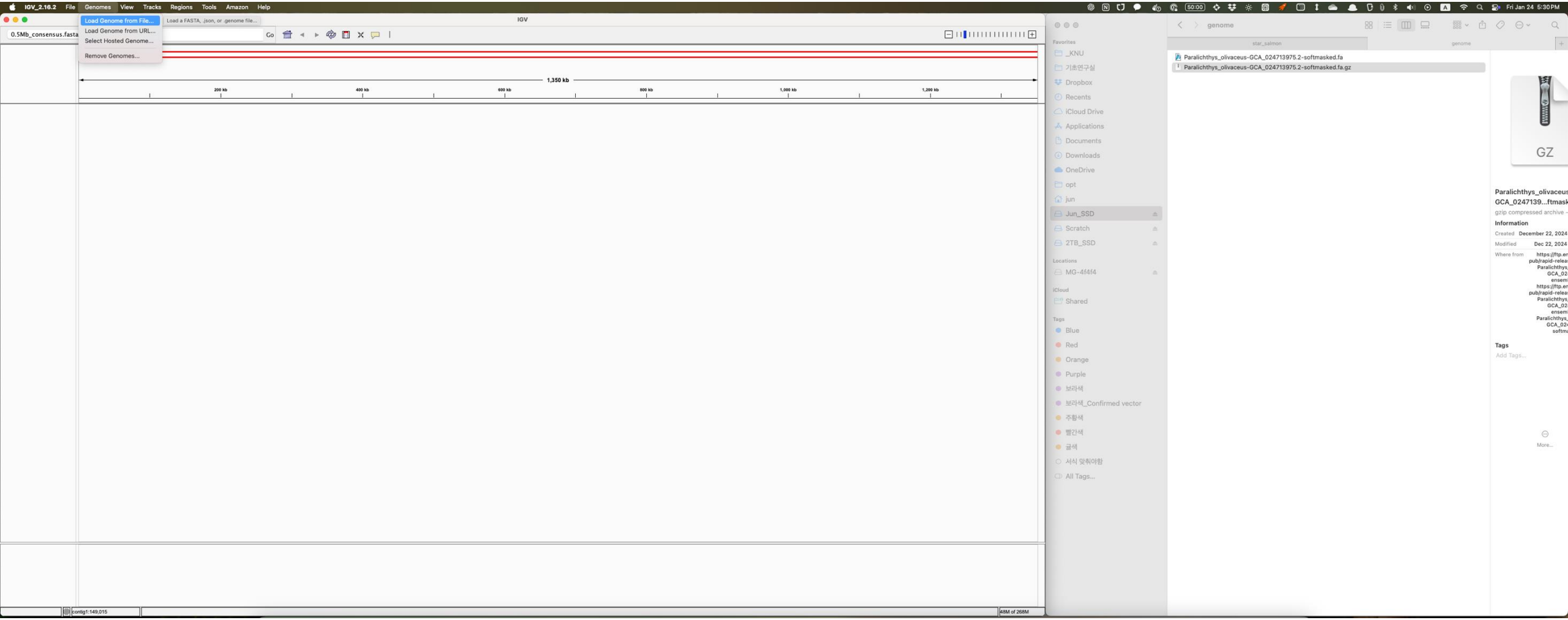
Abstract

We present JBrowse 2, a general-purpose genome annotation browser offering enhanced visualization of complex structural variation and evolutionary relationships. It retains core features of JBrowse while adding new views for synteny, dotplots, breakpoints, gene fusions, and whole-genome overviews. It allows users to share sessions, open multiple genomes, and navigate between views. It can be embedded in a web page, used as a standalone application, or run from Jupyter notebooks or R sessions. These improvements are enabled by a ground-up redesign using modern web technology. We describe application functionality, use cases, performance benchmarks, and implementation notes for web administrators and developers.

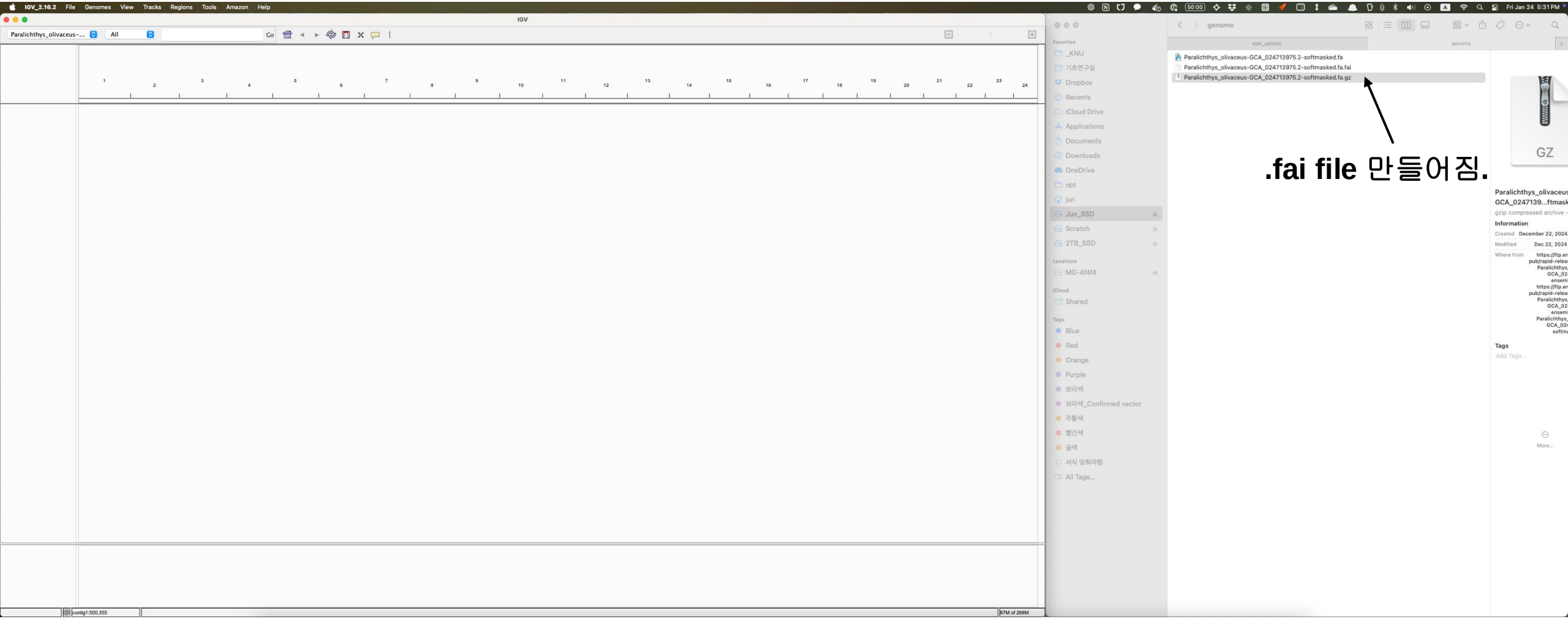


Download:
https://jbrowse.org/jb2/

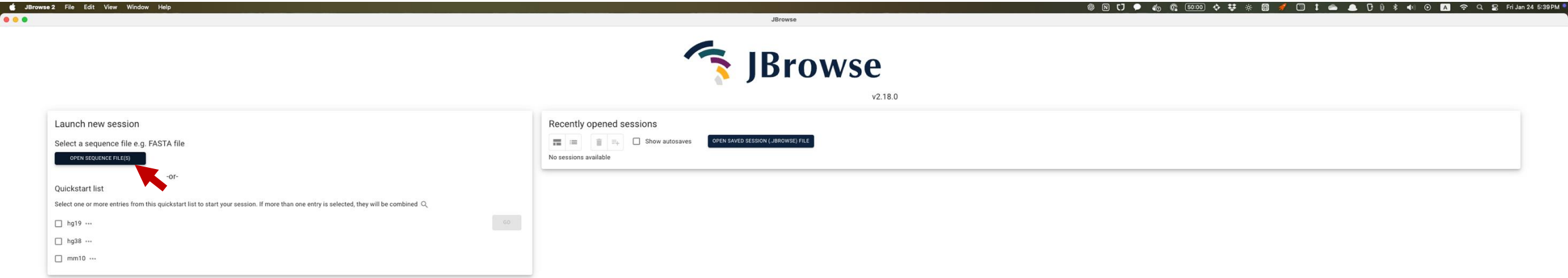
Load FASTA file on IGV



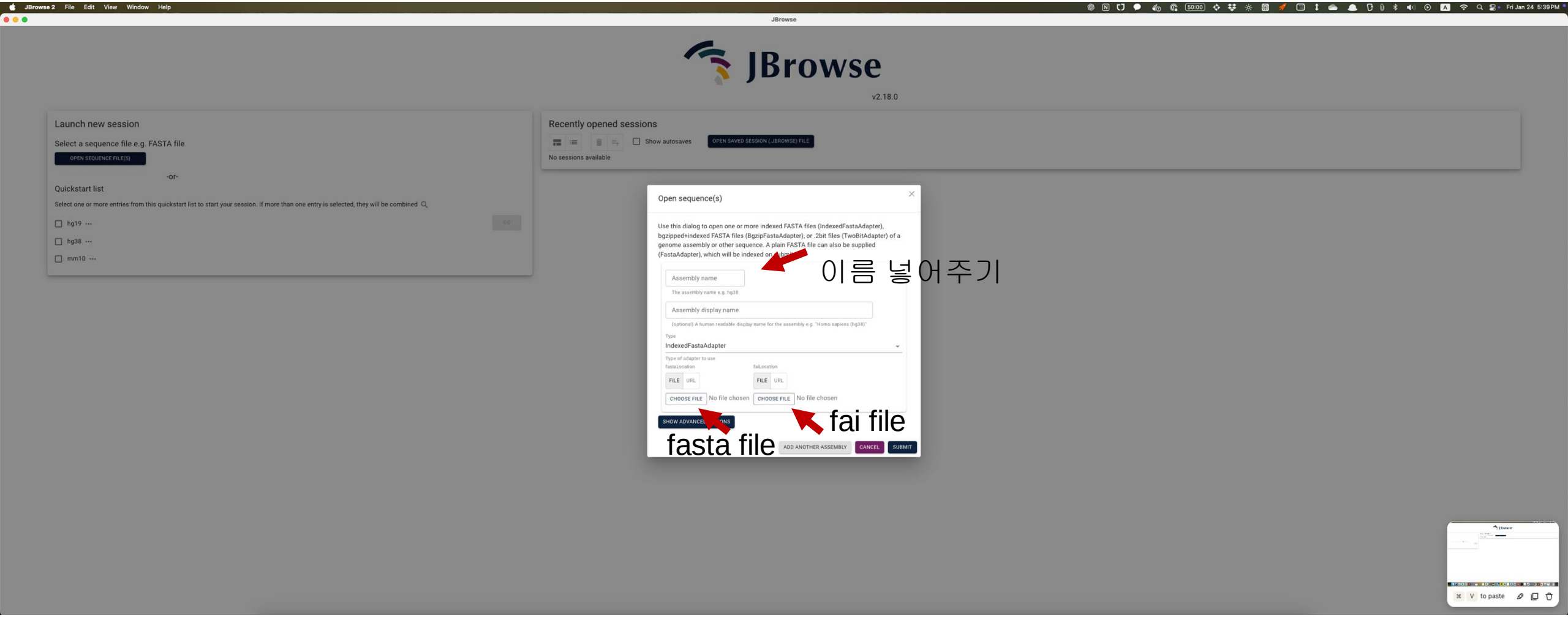
Load FASTA file on IGV



Load FASTA File and fai file on JBrowse -1



Load FASTA File and fai file on JBrowse -2



JBrowse 2 File Edit View Window Help

JBrowse

JBrowse

v2.18.0

Launch new session

Select a sequence file e.g. FASTA file

OPEN SEQUENCE FILE(S)

-or-

Quickstart list

Select one or more entries from this quickstart list to start your session. If more than one entry is selected, they will be combined

- hg19 ...
- hg38 ...
- mm10 ...

GO

Recently opened sessions

Show autosaves

OPEN SAVED SESSION (.JBROWSE) FILE

No sessions available

Open sequence(s)

Use this dialog to open one or more indexed FASTA files (IndexedFastaAdapter), bgzipped+indexed FASTA files (BgzipFastaAdapter), or .2bit files (TwoBitAdapter) of a genome assembly or other sequence. A plain FASTA file can also be supplied (FastaAdapter), which will be indexed on submit.

Assembly name

flounder

The assembly name e.g. hg38

Assembly display name

flounder

(optional) A human readable display name for the assembly e.g. "Homo sapiens (hg38)"

Type

IndexedFastaAdapter

Type of adapter to use

fastaLocation

FILE URL

CHOOSE FILE /Volumes/Jun_SSD/genome/Paralichthys_olivaceus/Ensembl_rapid/genome/P
GCA_024713975.2-softmasked.fa

faiLocation

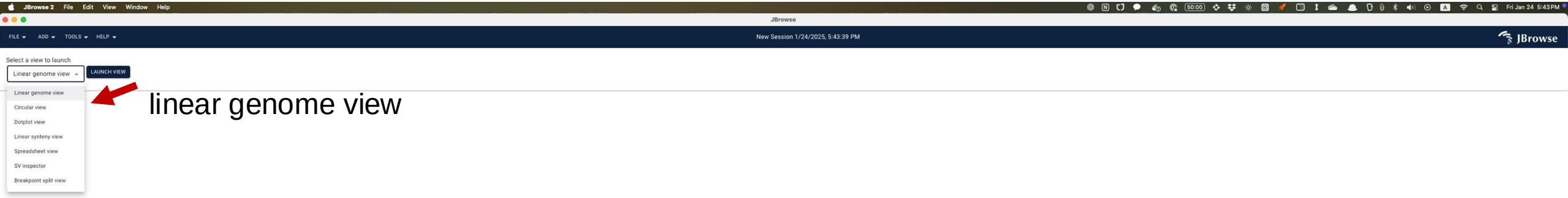
FILE URL

CHOOSE FILE /Volumes/Jun_SSD/genome/Paralichthys_olivaceus/Ensembl_rapid/genome/P
GCA_024713975.2-softmasked.fa.fai

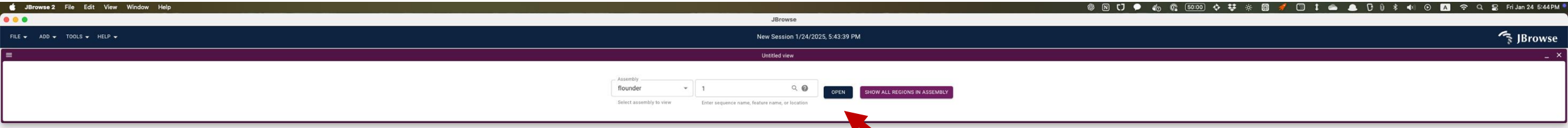
SHOW ADVANCED OPTIONS

ADD ANOTHER ASSEMBLY CANCEL SUBMIT

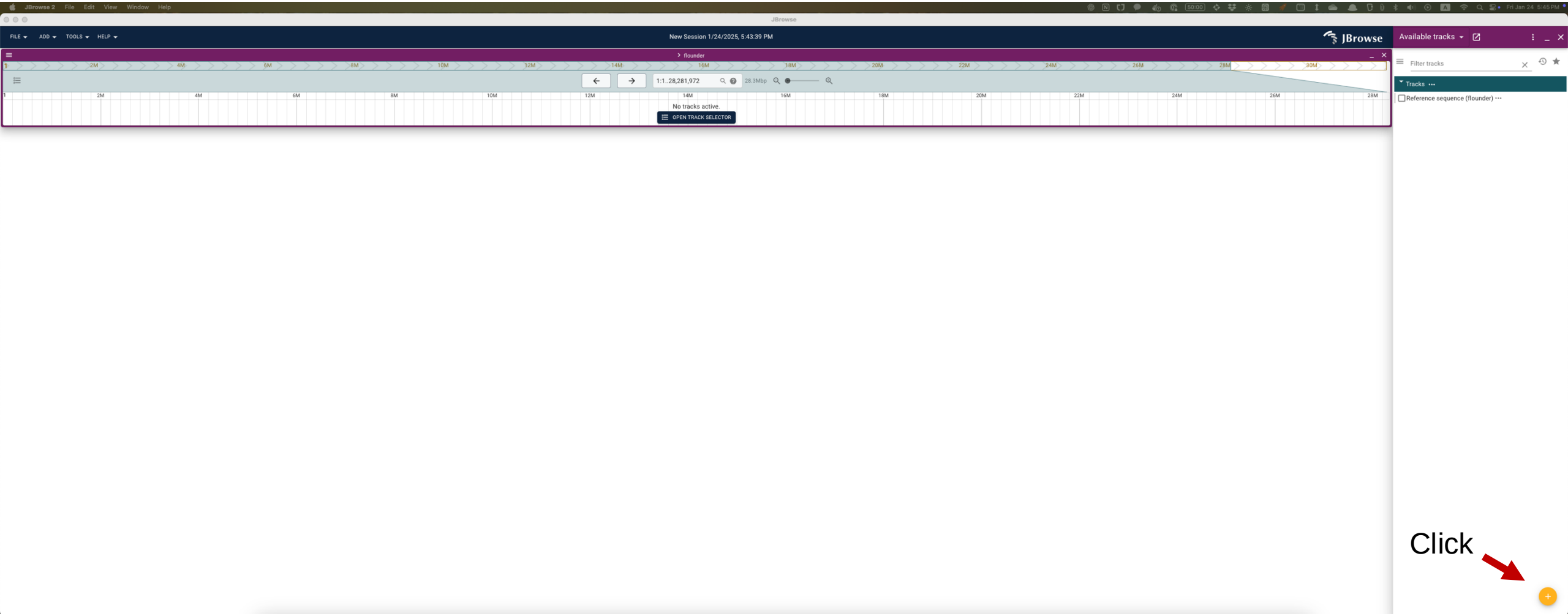
Open sequence(s)

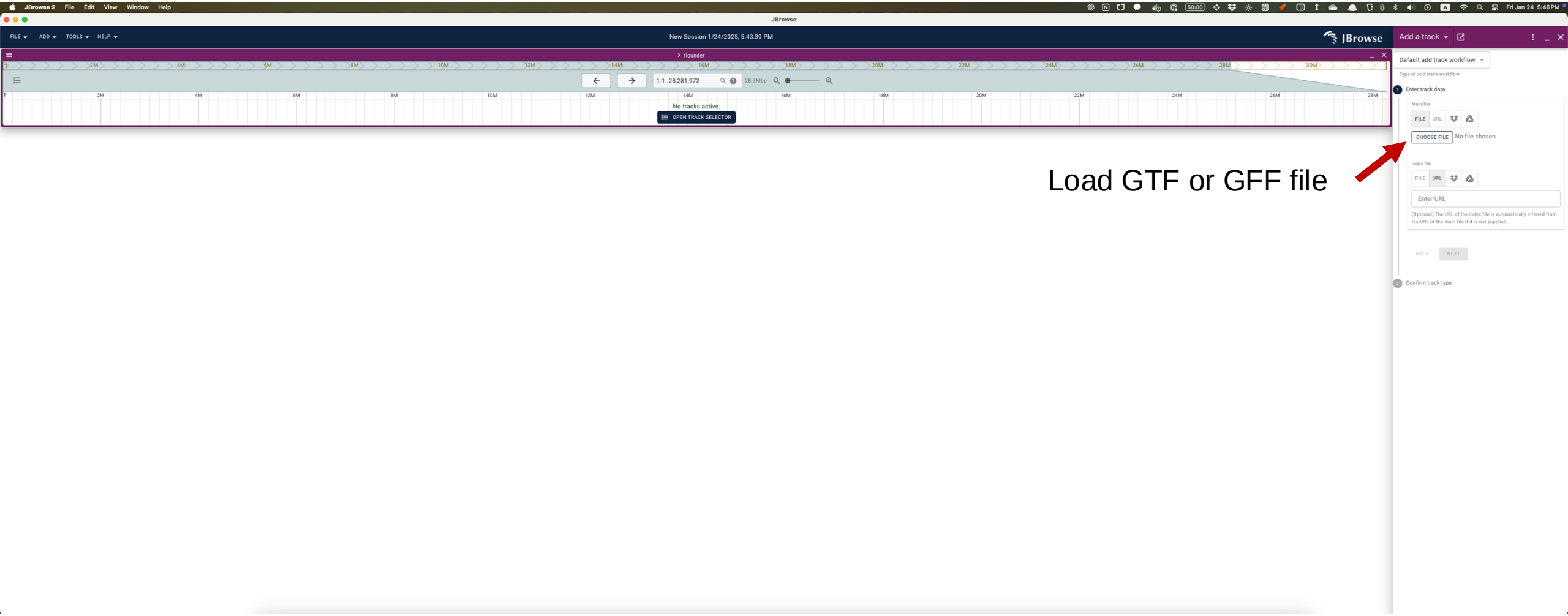


linear genome view

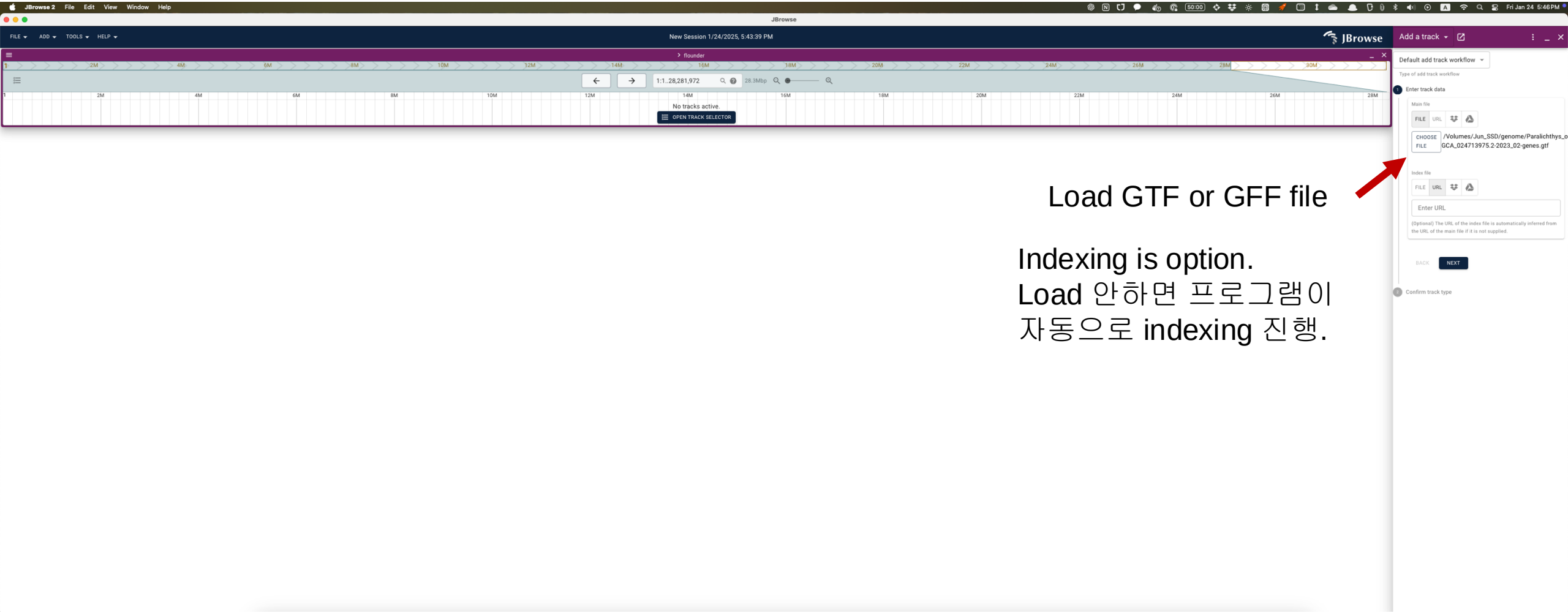

JBrowse 2
File
Edit
View
Window
Help
Fri Jan 24 5:44PM
JBrowse
FILE
ADD
TOOLS
HELP
New Session 1/24/2025, 5:43:39 PM
JBrowse
Untitled view
Assembly
flounder
Select assembly to view
1
Enter sequence name, feature name, or location
OPEN
SHOW ALL REGIONS IN ASSEMBLY

Open


JBrowse 2
File
Edit
View
Window
Help
JBrowse
FILE
ADD
TOOLS
HELP
New Session 1/24/2025, 5:43:39 PM
JBrowse
Available tracks
flounder
1:1..28,281,972
28.3Mbp
No tracks active.
OPEN TRACK SELECTOR
Filter tracks
Tracks
Reference sequence (flounder)
Click

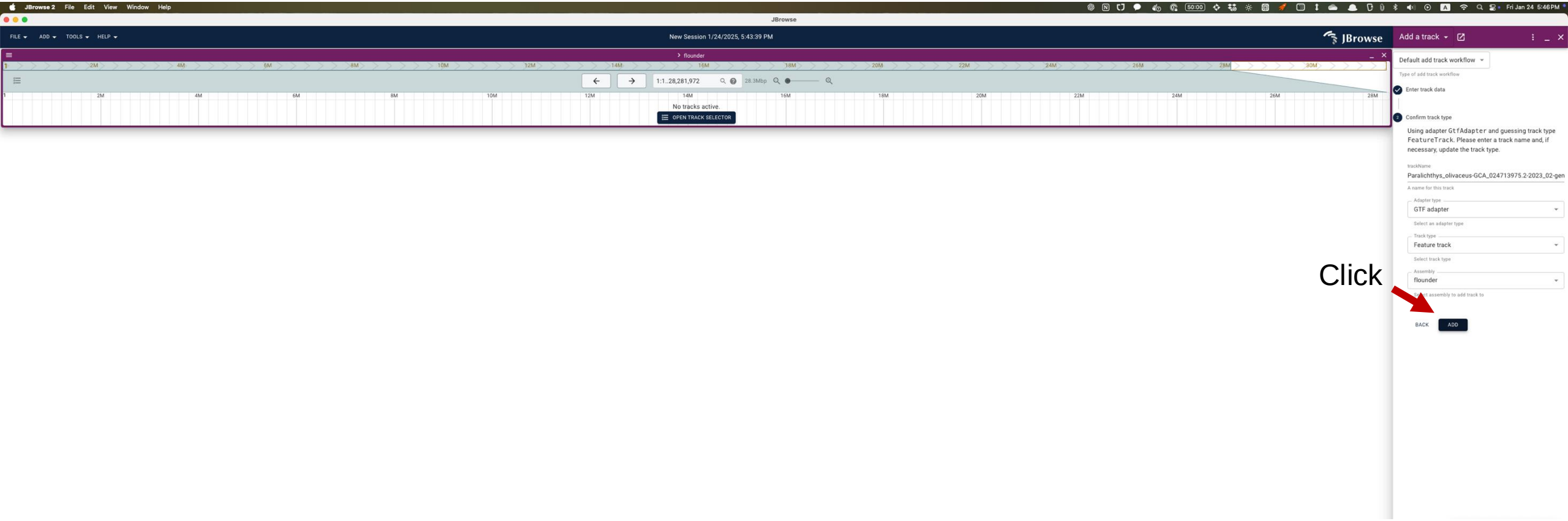


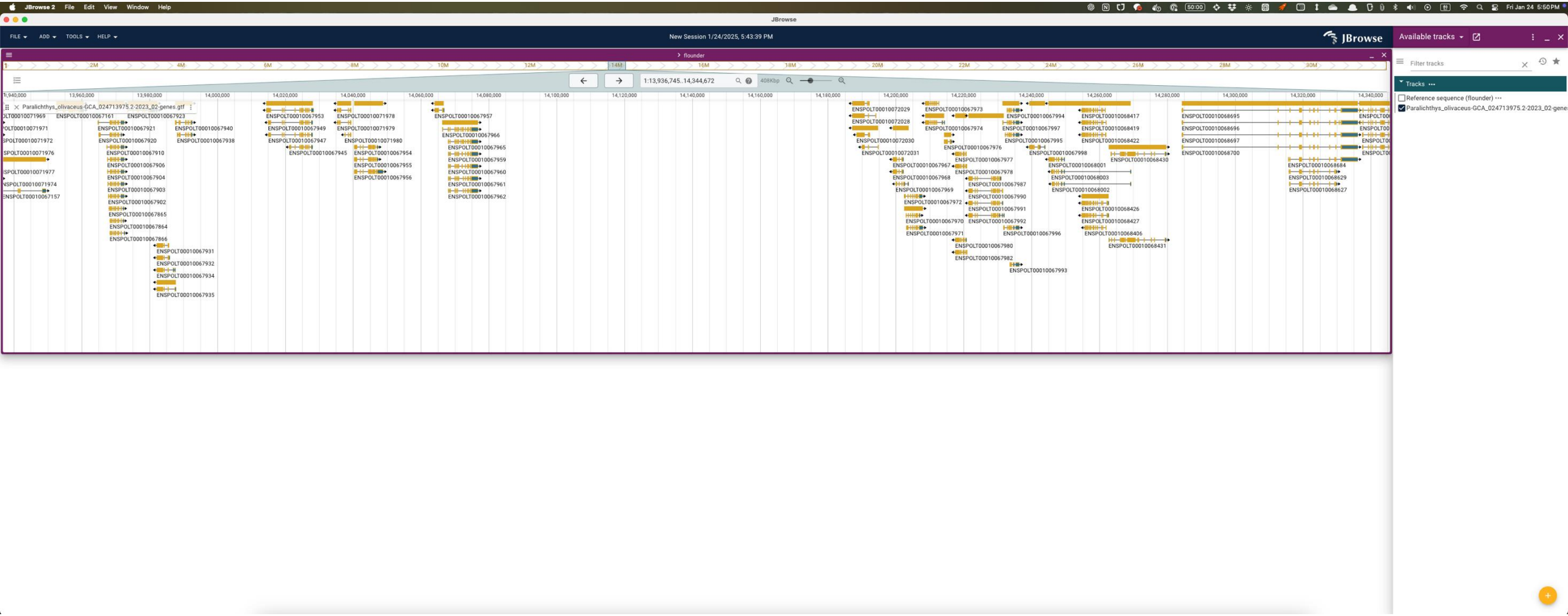
Load GTF or GFF file

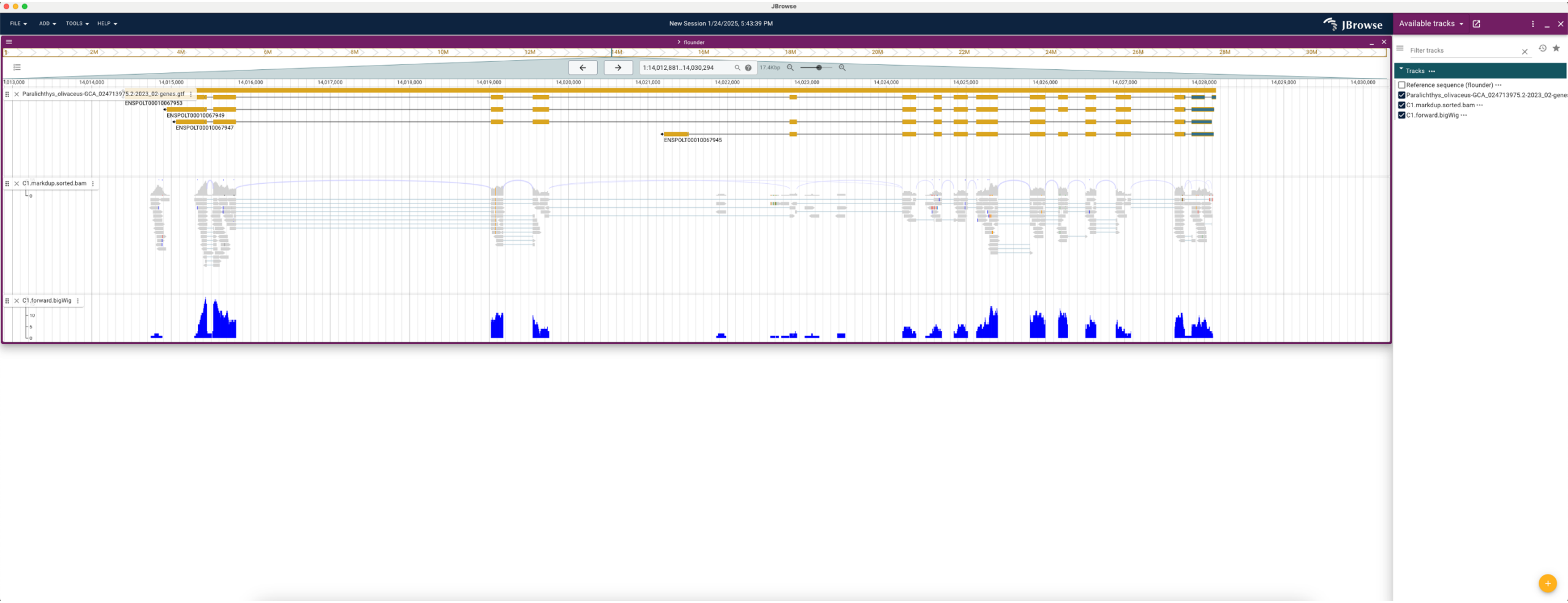


Load GTF or GFF file

Indexing is option.
Load 안하면 프로그램이
자동으로 indexing 진행.


JBrowse 2 File Edit View Window Help
Fri Jan 24 5:46 PM
JBrowse
FILE ADD TOOLS HELP
New Session 1/24/2025, 5:43:39 PM
JBrowse
Add a track
flounder
1:1..28,281,972
28.3Mbp
No tracks active.
OPEN TRACK SELECTOR
Default add track workflow
Type of add track workflow
Enter track data
Confirm track type
Using adapter GtfAdapter and guessing track type FeatureTrack. Please enter a track name and, if necessary, update the track type.
trackName
Paralichthys_olivaceus-GCA_024713975.2-2023_02-gen
A name for this track
Adapter type
GTF adapter
Select an adapter type
Track type
Feature track
Select track type
Assembly
flounder
BACK
ADD
Click


JBrowse 2
File
Edit
View
Window
Help
Fri Jan 24 5:50 PM
JBrowse
FILE
ADD
TOOLS
HELP
New Session 1/24/2025, 5:43:39 PM
JBrowse
Available tracks
flounder
1:13,936,745..14,344,672
408Kbp
Paralichthys_olivaceus-GCA_024713975.2-2023_02-genes.gtf
Filter tracks
Tracks
Reference sequence (flounder)
Paralichthys_olivaceus-GCA_024713975.2-2023_02-genes


JBrowse
FILE
ADD
TOOLS
HELP
New Session 1/24/2025, 5:43:39 PM
JBrowse
Available tracks
flounder
1:14,012,881..14,030,294
17.4Kbp
Paralichthys_olivaceus-GCA_024713975.2-2023_02-genes.gtf
ENSPOLT00010067953
ENSPOLT00010067949
ENSPOLT00010067947
ENSPOLT00010067945
C1.markdup.sorted.bam
C1.forward.bigWig
Filter tracks
Tracks
Reference sequence (flounder)
Paralichthys_olivaceus-GCA_024713975.2-2023_02-gene
C1.markdup.sorted.bam
C1.forward.bigWig