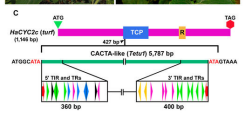
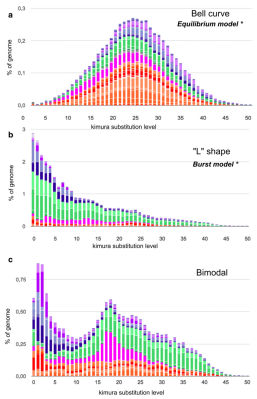
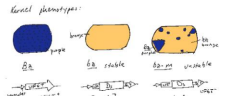



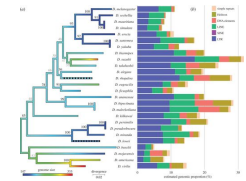




BIOINFORMATIC ANALYSIS OF TRANSPOSABLE ELEMENTS


3rd-7th November 2025

Lecture 3B
TE manual curation






Valentina Peona

NATURHISTORISKA RIKSMUSEET

What is manual curation?

Library 1.0

```
>cons1
ATGAT
>cons2
GATTACA
…
```

The automatic generation of a TE library often comes with issues that must be solved:

- Incomplete consensus sequences (e.g., boundaries are wrong)
- Multiple variants were pooled in the same consensus sequence
- Misclassified sequences

What is manual curation?

Library 1.0

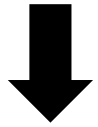
```
>cons1
ATGAT
>cons2
GATTACA
…
```

The automatic generation of a TE library often comes with issues that must be solved:

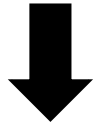
- Incomplete consensus sequences (e.g., boundaries are wrong)
- Multiple variants were pooled in the same consensus sequence
- Misclassified sequences

What can we do?

Collect TE insertions from the genome similar to our consensus sequences



Create a multi-sequence alignment for each consensus sequence and its hits



What is manual curation?

Library 1.0

```
>cons1
ATGAT
>cons2
GATTACA
…
```

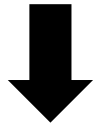
The automatic generation of a TE library often comes with issues that must be solved:

- Incomplete consensus s… …daries are wrong)
- Multiple variants … …us sequence
- Misclassified sequenc…

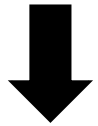
It can be a very long process (and a bit frustrating) but it is worth it!

What can we do?

Collect TE insertions from the genome similar to our consensus sequences



Create a multi-sequence alignment for each consensus sequence and its hits



Why manual curation?

To obtain an as complete and accurate annotation as possible

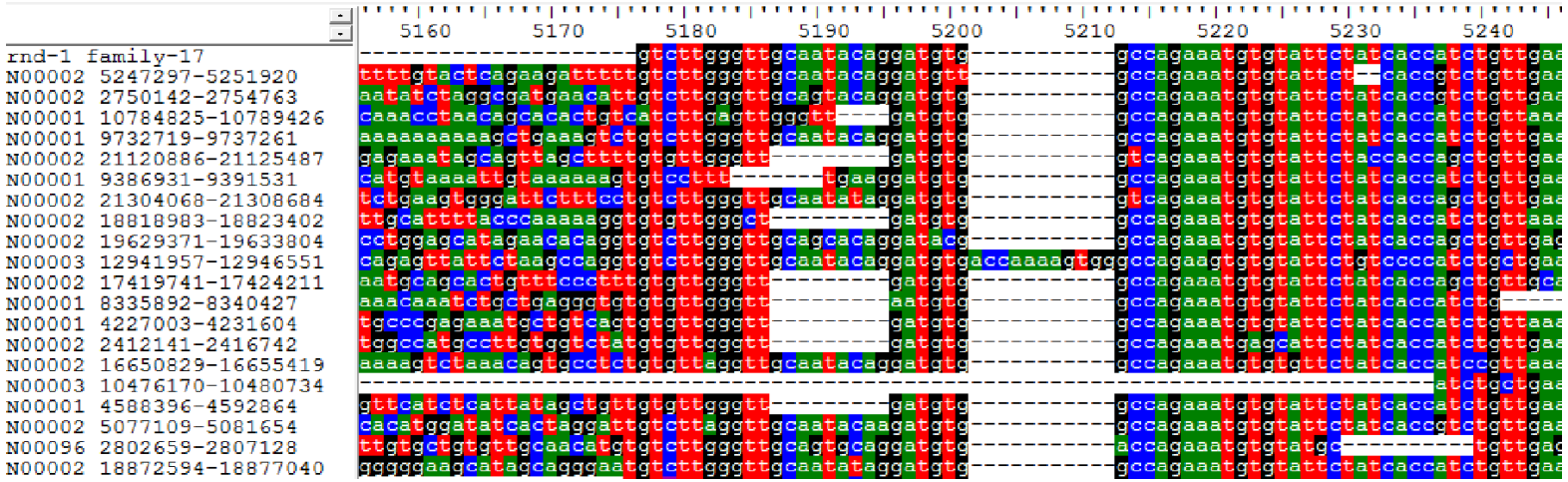
1. Incomplete consensus sequences
2. Misclassified consensus sequences
3. Species-specific libraries are key
4. Complete library equals complete annotation

→ Avoid biological misinterpretations

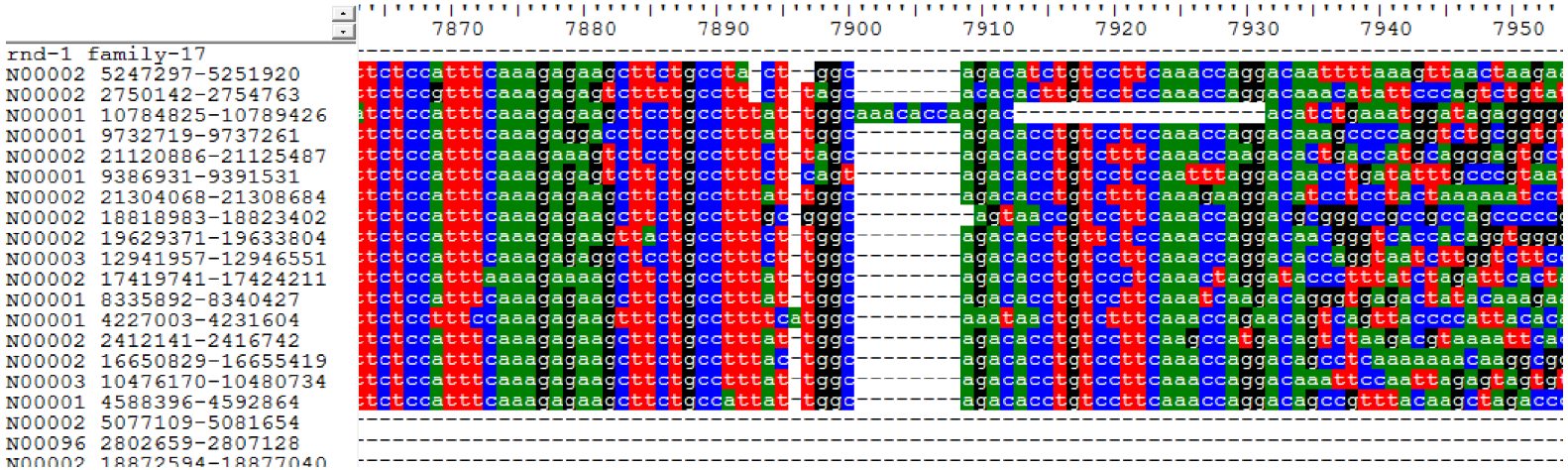
Incomplete consensus sequences

5’

Old consensus sequence



New consensus sequence

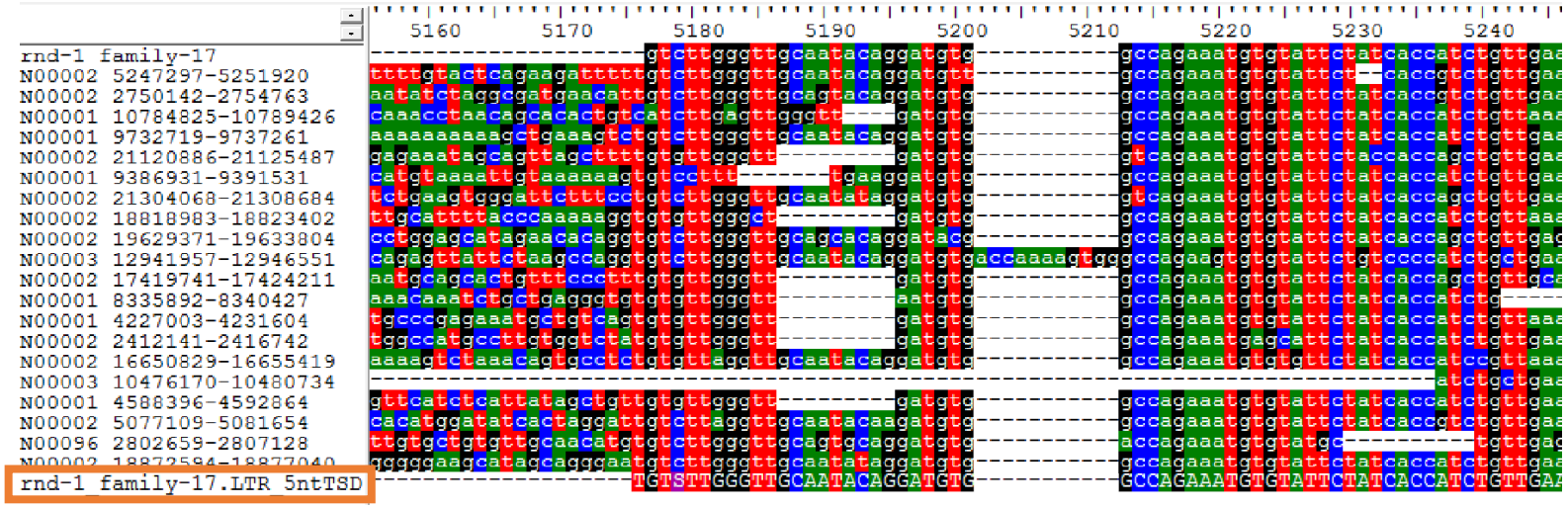


3’

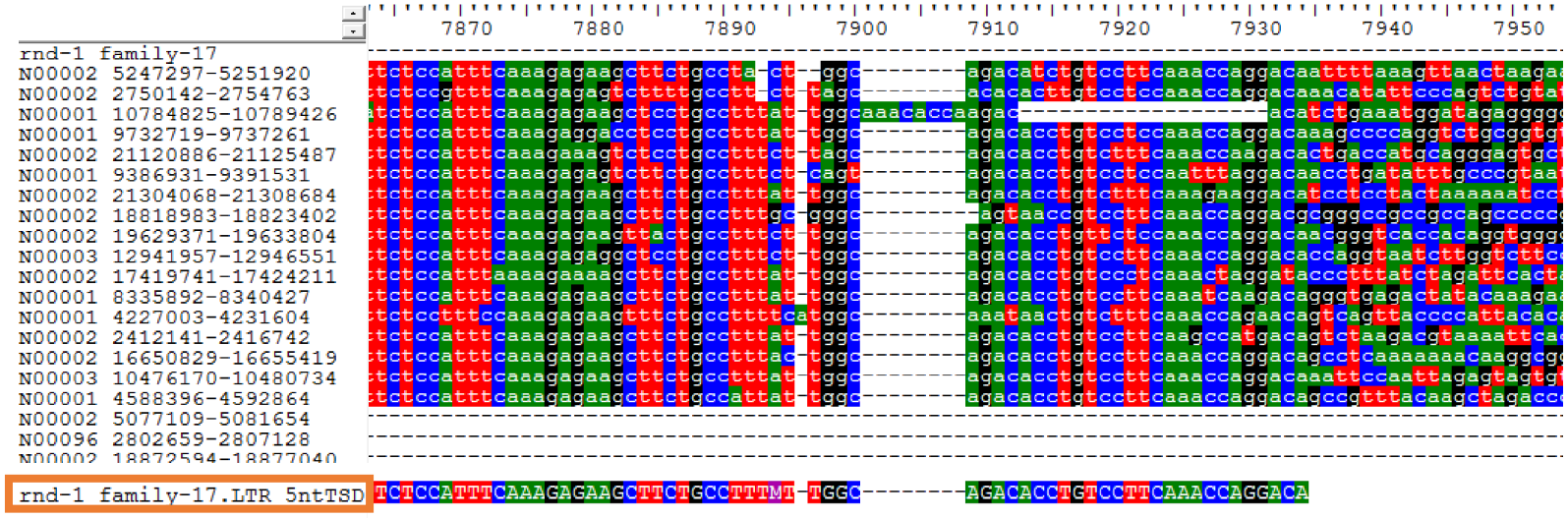
Incomplete consensus sequences

5’

Old consensus sequence



New consensus sequence



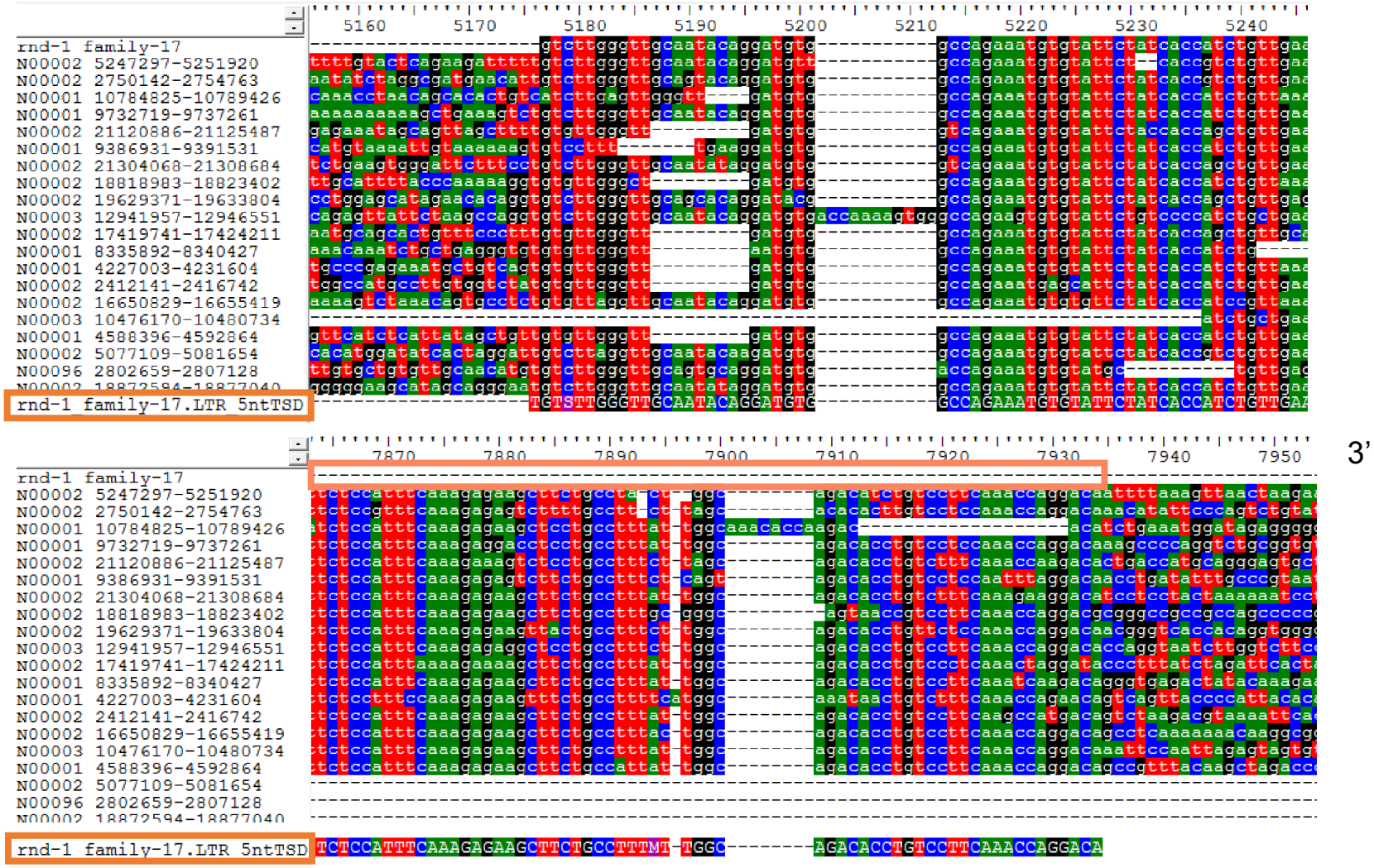
3’

Incomplete consensus sequences

5’

Old consensus sequence

New consensus sequence



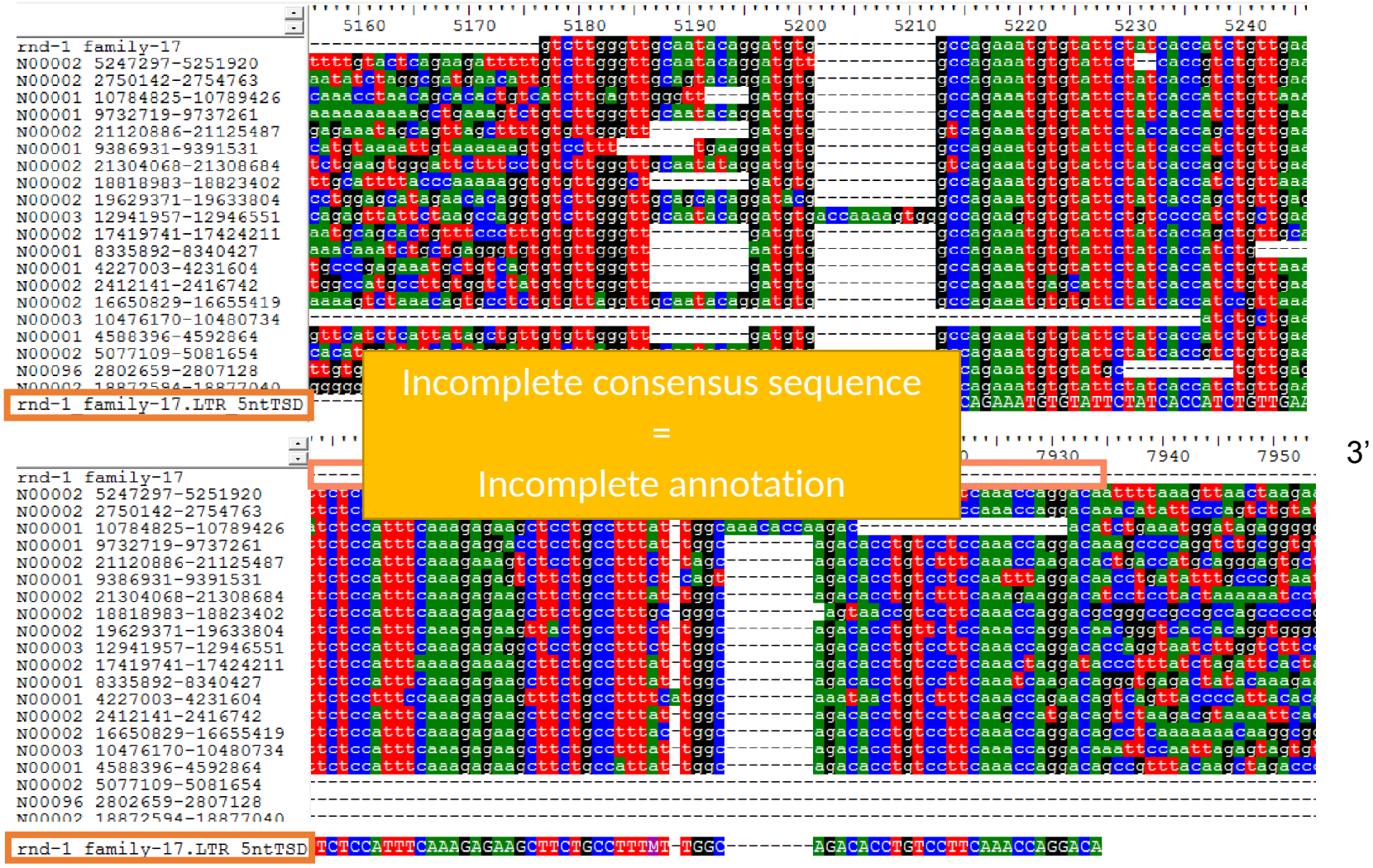
3’

Incomplete consensus sequences

5’

Old consensus sequence

New consensus sequence



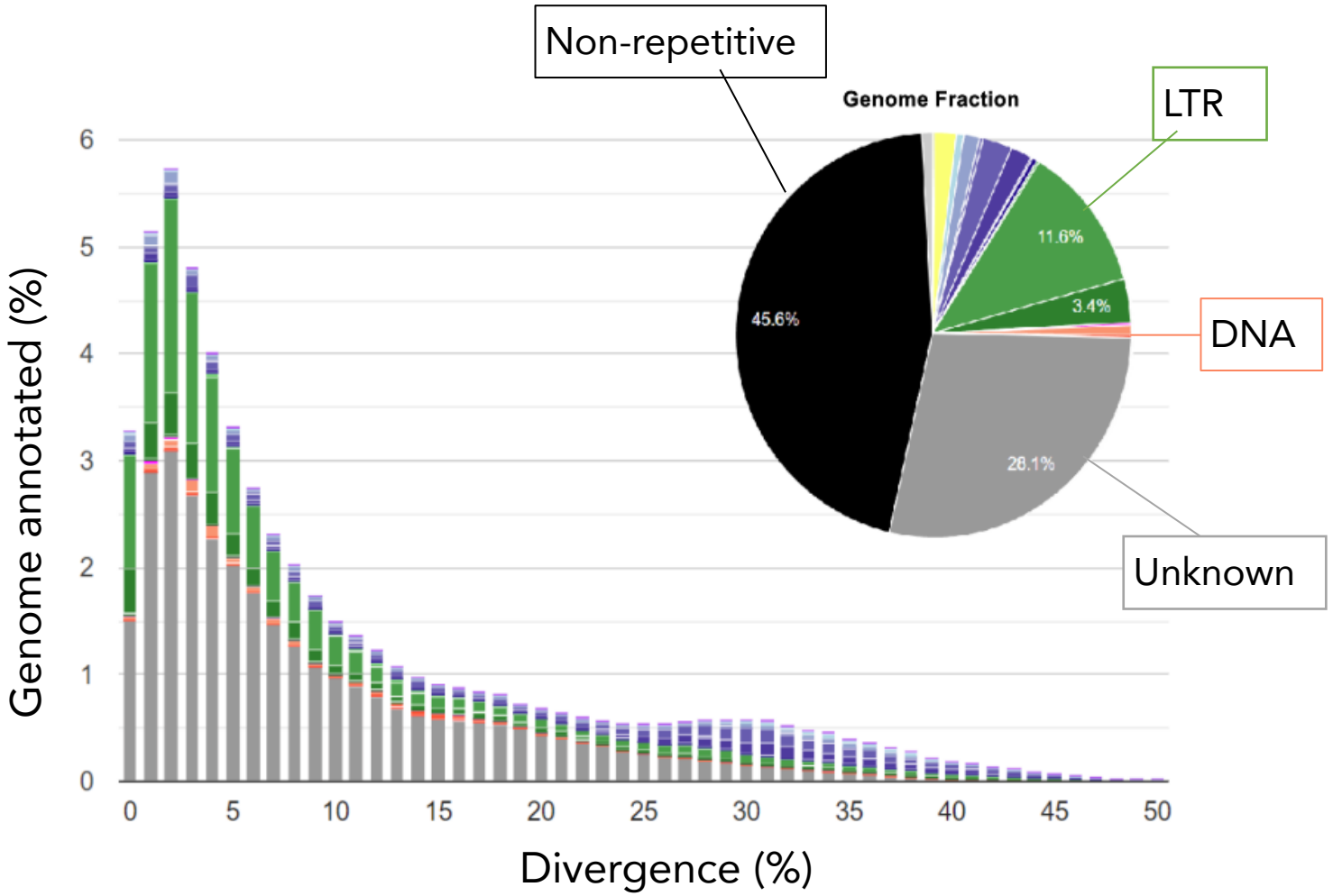
3’

Misclassification

The case of a crustacean

Repeat library of raw consensus sequences from RepeatModeler2

TE category	# consensus sequences
DNA	114
LTR	1909
LINE	378
SINE	3
RC	10
Unknown	3774



Demirci, Peona, Suh et al. unpublished

Misclassification

The case of a crustacean

Original (RM1) \ Curated (RM2)	Unknown	LTR	LINE	*Mariner*	*hAT*	*Merlin*	*Transib*	*P*	*PiggyBac*	*CACTA*	*Sola2*	Total	Total (%)
Unknown	14	1	0	14	11	1	1	1	8	4	1	56	62.22%
LTR	6	1	0	7	1	0	0	0	17	0	0	32	35.56%
LINE	0	0	2	0	0	0	0	0	0	0	0	2	2.22%
Mariner	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
hAT	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
Merlin	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
Transib	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
P	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
PiggyBac	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
CACTA	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
Sola2	0	0	0	0	0	0	0	0	0	0	0	0	0.00%
Total	20	2	2	21	12	1	1	1	25	4	1	90	
Total (%)	22.22%	2.22%	2.22%	23.33%	13.33%	1.11%	1.11%	1.11%	27.78%	4.44%	1.11%		

Key:	Unknown	LTR	LINE	DNA

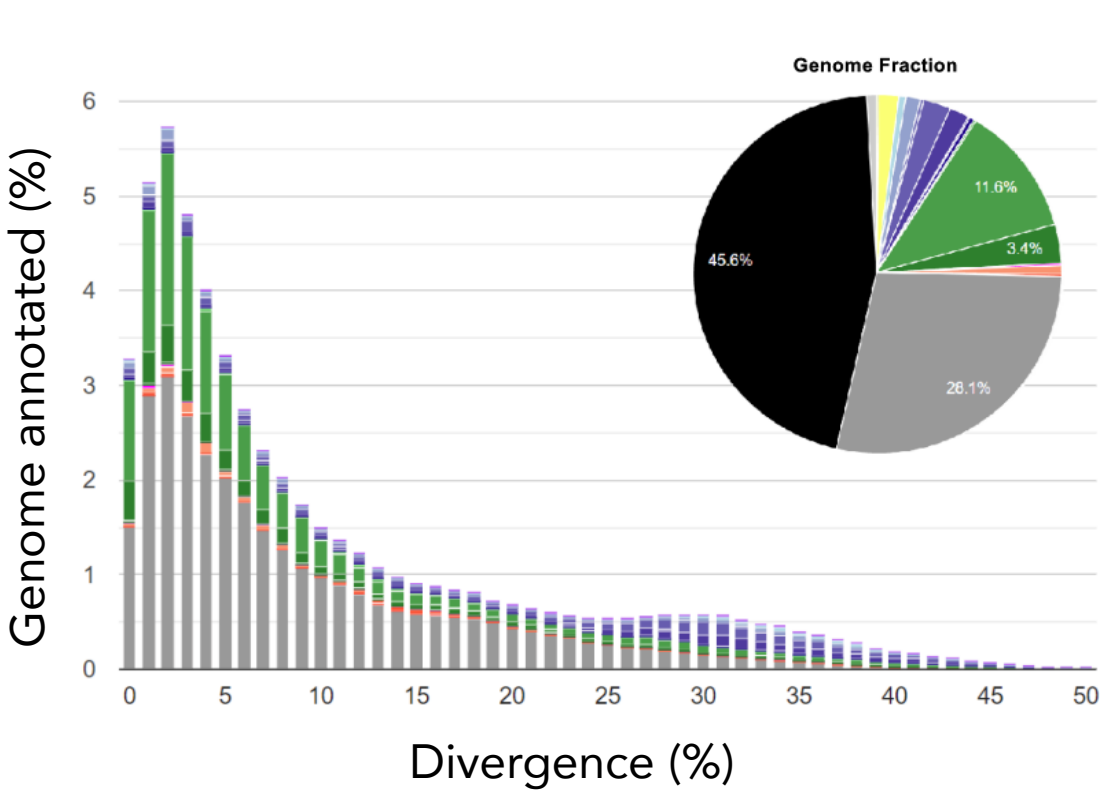
- 32 consensus sequences preliminary classified as LTR by RepeatModeler were curated
- Only one of them was confirmed to be an LTR
- Most of them were DNA transposons!
- Many more remain in the library to be curated…!
- My solution: label all LTRs as unknown unless there are protein-based evidence that indicate LTR

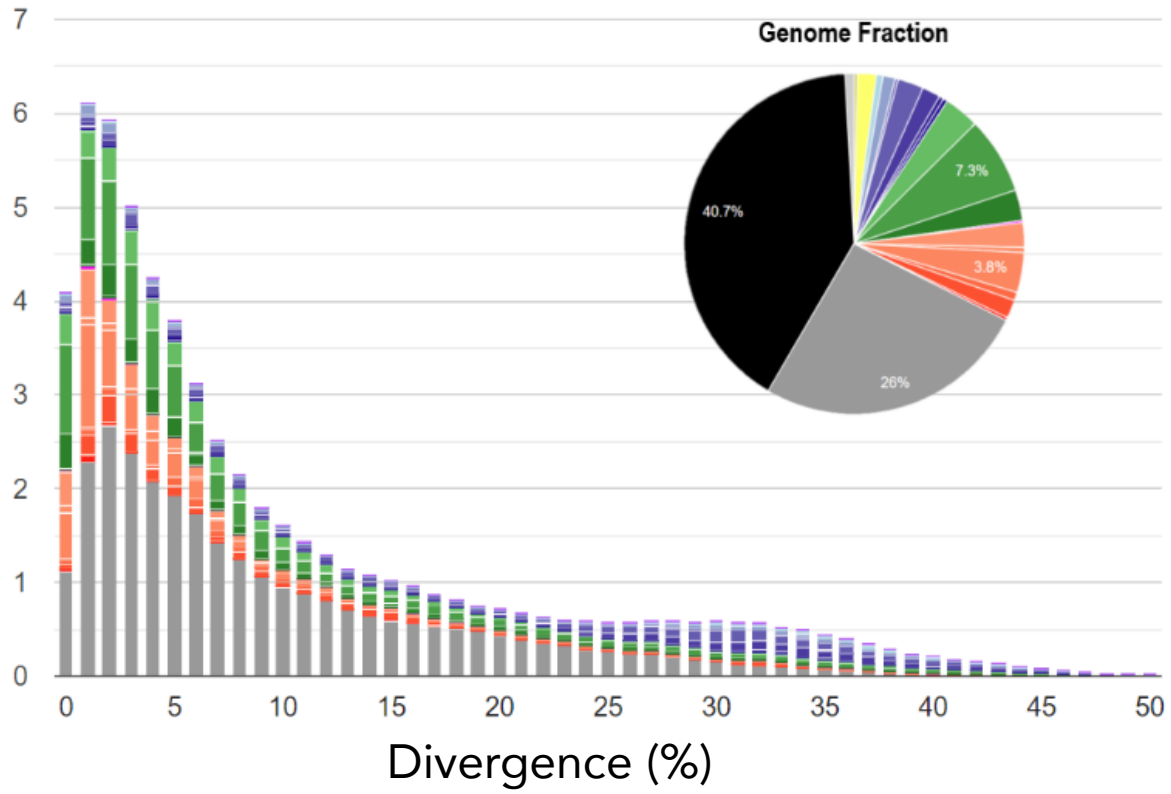
Demirci, Peona, Sun et al. unpublished

Annotation improvements

The case of a crustacean

After curating 100 consensus sequences



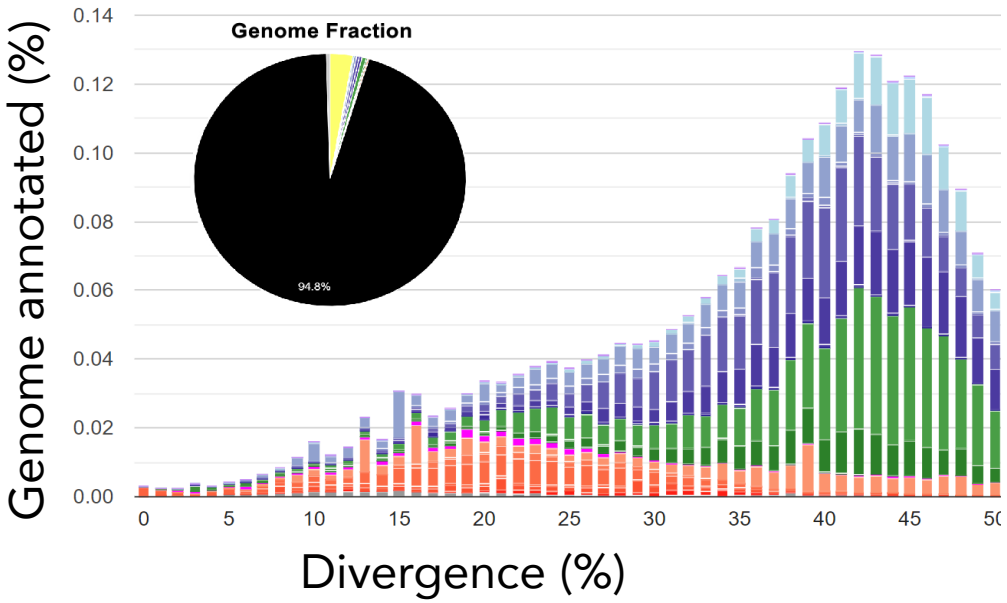


LTR DNA

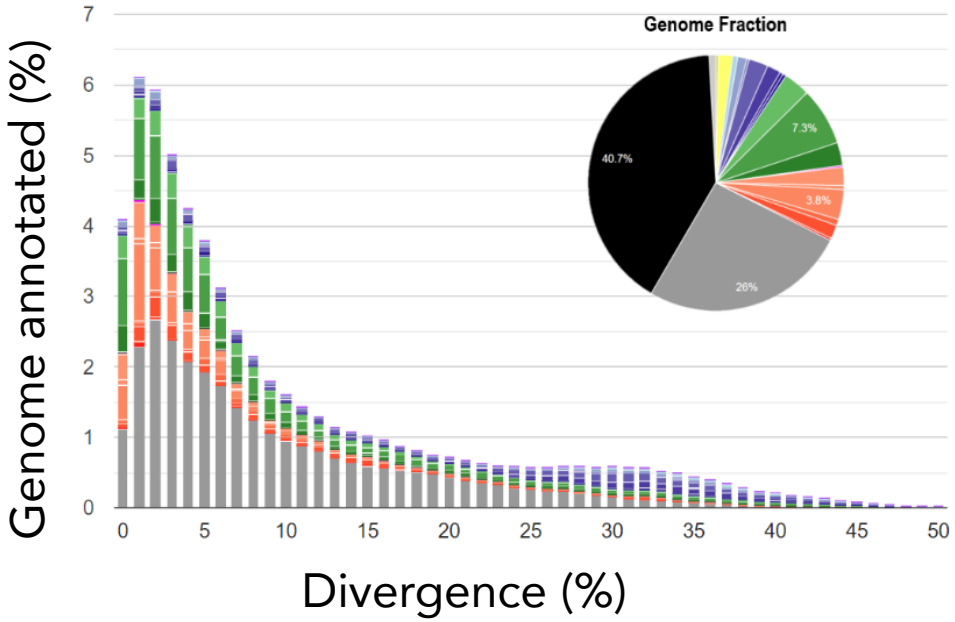
Demirci, Peona, Suh et al. unpublished

Species-specific libraries

Library from database



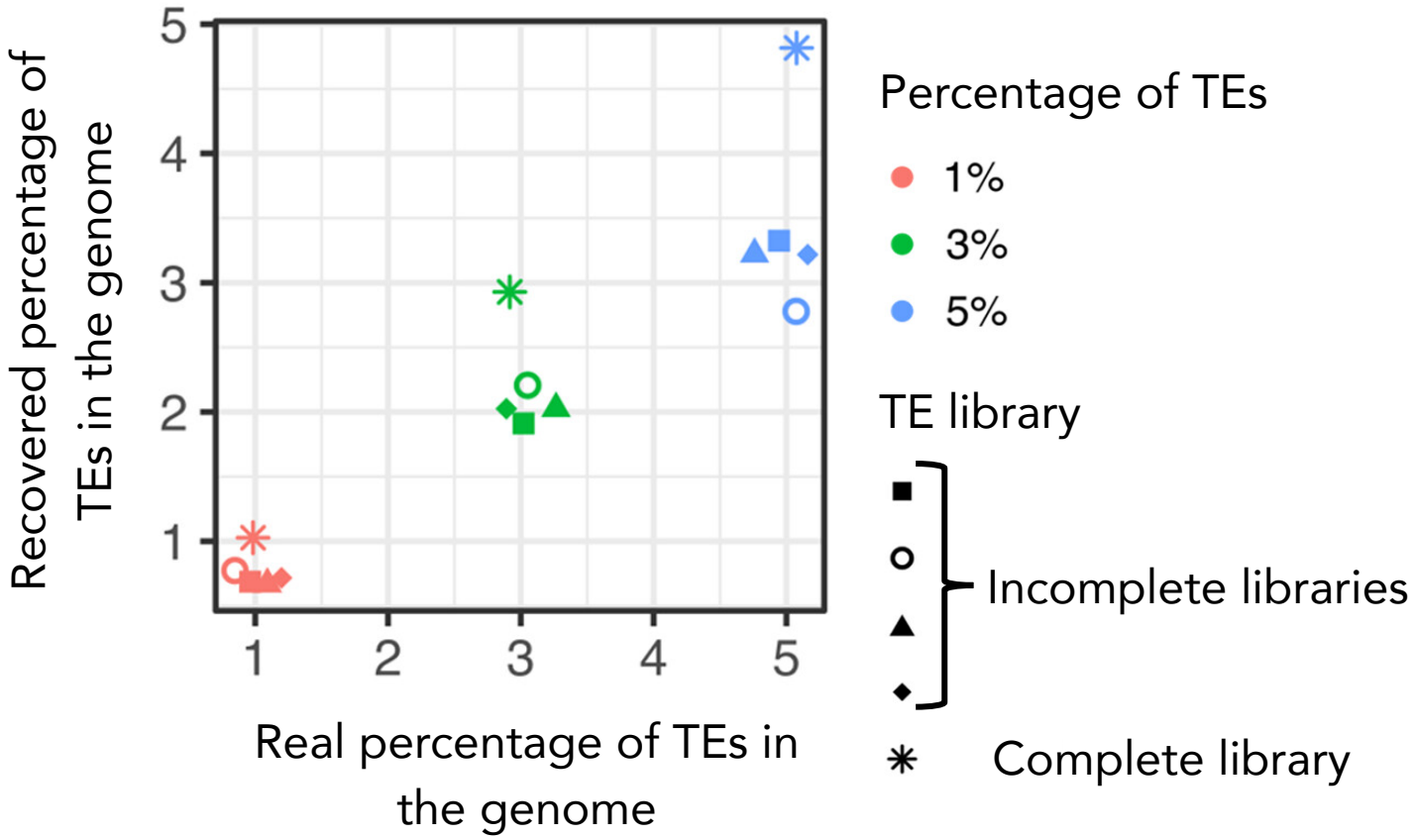
Species-specific library



- Overall percentage
- Different accumulation distribution

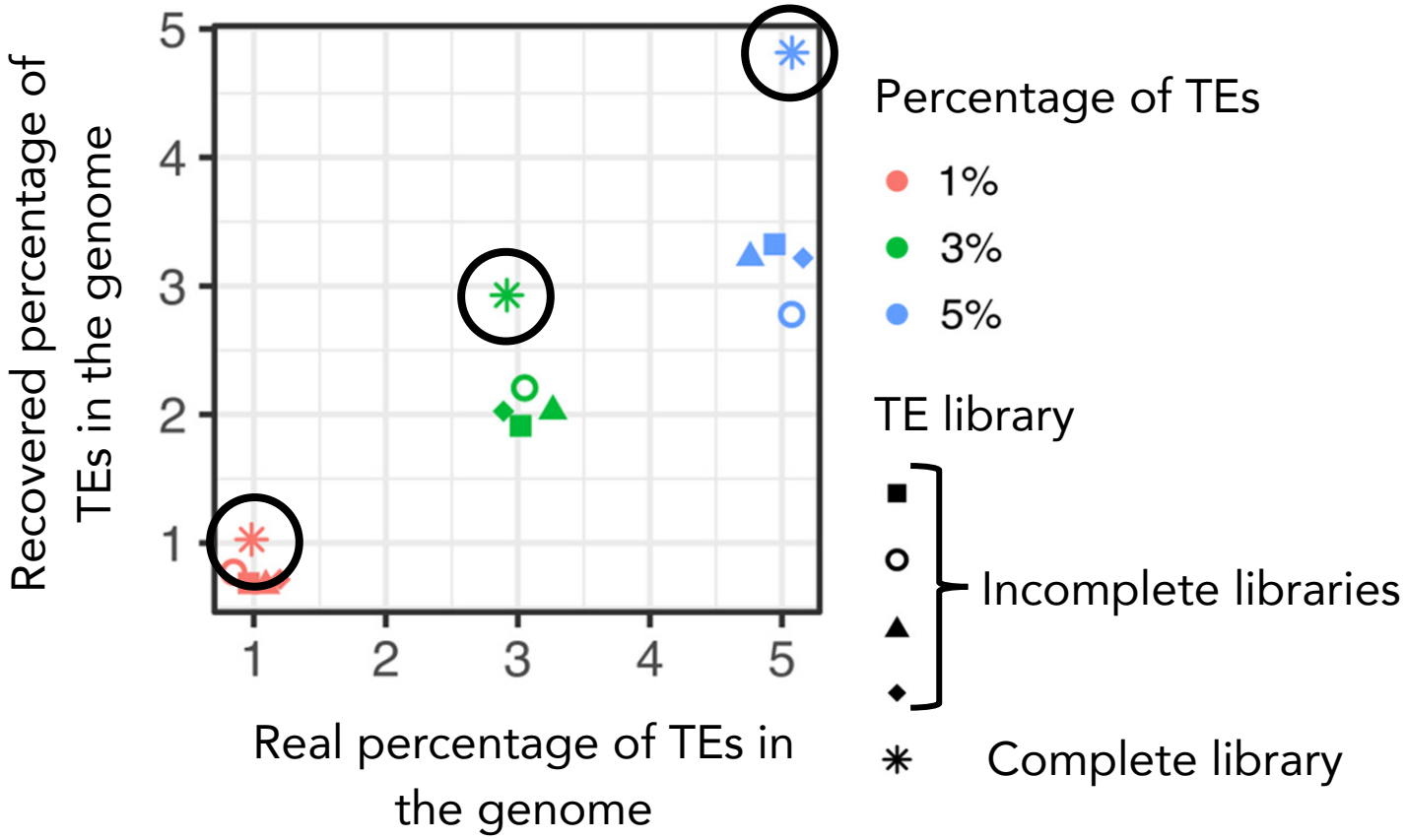
Demirci, Peona, Suh et al. unpublished

You have it, you find it!



Peona et al. 2023, MolEcol

You have it, you find it!



Peona et al. 2023, MolEcol

Some problems of manual curation

1. Time consuming
2. Difficult to scale up to the number of genomes we have now available
3. Always a bit of subjectivity that hinders complete reproducibility

Time consuming: prioritise!

.divsum file can be very useful!

```
Jukes/Cantor and Kimura subsitution levels adjusted for CpG sites
=================================================================
File: haploid_assembly_BUSCOmancur.fasta.align
Weighted average Kimura divergence for each repeat family
Class    Repeat   absLen   wellCharLen      Kimura%
-----    ------   ------   -----------      -------
LTR/Gypsy        ltr-1-1295       3061696 3041108 3.78
LTR/Gypsy        ltr-1-1269       3033238 3016481 4.16
LTR/Gypsy        ltr-1-1589       2885608 2854985 5.77
Unknown balImp1-86       2847824 2819143 8.79
LTR/Gypsy        ltr-1-205        2710211 2684973 3.89
LTR/Unknown      ltr-1-1379       2496705 2463106 4.88
LTR/Gypsy        ltr-1-1115       2478128 2460958 3.11
Unknown balImp1-19       2439731 2433087 2.07
Unknown balImp3-158      2358241 2334267 5.35
LTR/Gypsy        ltr-1-1358       2123961 2083228 6.42
```

1

Time consuming: prioritise!

.divsum file can be very useful!

```
Jukes/Cantor and Kimura subsitution levels adjusted for CpG sites
=================================================================
File: haploid_assembly_BUSCOmancur.fasta.align
Weighted aver                                              peat family
Class   Repea                                              %
-----   -----                                              -
LTR/Gypsy                                                  3 3.78
LTR/Gypsy                                                  1 4.16
LTR/Gypsy                                                  5 5.77
Unknown balIm
LTR/Gypsy                                                  3 3.89
LTR/Unknown      ltr-1-1379       2496705 2463106 4.88
LTR/Gypsy        ltr-1-1115       2478128 2460958 3.11
Unknown balImp1-19       2439731 2433087 2.07
Unknown balImp3-158      2358241 2334267 5.35
LTR/Gypsy        ltr-1-1358       2123961 2083228 6.42
```

Look for abundant repeats and/or young repeats

Or focus on categories you are already interested in

Scalability

SPECIAL SECTION ZOONOMIA

RESEARCH ARTICLE

ZOONOMIA

Insights into mammalian TE diversity through the curation of 248 genome assemblies

Austin B. Osmanski[1], Nicole S. Paulat[1], Jenny Korstian[1], Jenna R. Grimshaw[1], Michaela Halsey[1], Kevin A. M. Sullivan[1], Diana D. Moreno-Santillán[1], Claudia Crookshanks[1], Jacquelyn Roberts[1], Carlos Garcia[1], Matthew G. Johnson[1], Llewellyn D. Densmore[1], Richard D. Stevens[2], Zoonomia Consortium†, Jeb Rosen[3], Jessica M. Storer[3], Robert Hubley[3], Arian F. A. Smit[3], Liliana M. Dávalos[4,5], Elinor K. Karlsson[6,7], Kerstin Lindblad-Toh[7,8,9], David A. Ray[1]*

Employed hundreds of students

Curated two repeat libraries
~40 participants (two editions)

Peona et al. Mobile DNA (2024) 15:10
https://doi.org/10.1186/s13100-024-00319-8

Mobile DNA

RESEARCH Open Access

Teaching transposon classification as a means to crowd source the curation of repeat annotation – a tardigrade perspective

Valentina Peona[1,2,3*†], Jacopo Martelossi[4*†], Dareen Almojil[5], Julia Bocharkina[6], Ioana Brännström[7,8], Max Brown[9], Alice Cang[10], Tomàs Carrasco-Valenzuela[11,12], Jon DeVries[13], Meredith Doellman[14,15], Daniel Elsner[16], Pamela Espíndola-Hernández[17], Guillermo Friis Montoya[18], Bence Gaspar[19], Danijela Zagorski[20], Paweł Hałakuc[21], Beti Ivanovska[22], Christopher Laumer[23], Robert Lehmann[24], Ljudevit Luka Boštjančić[25], Rahia Mashoodh[26], Sofia Mazzoleni[27], Alice Mouton[28], Maria Anna Nilsson[25], Yifan Pei[1,29], Giacomo Potente[30], Panagiotis Provataris[31], José Ramón Pardos-Blas[32], Ravindra Raut[33], Tomasa Sbaffi[34], Florian Schwarz[35], Jessica Stapley[36], Lewis Stevens[37], Nusrat Sultana[38], Radka Symonova[39], Mohadeseh S. Tahami[40], Alice Urzi[41], Heidi Yang[42], Abdullah Yusuf[43], Carlo Pecoraro[44] and Alexander Suh[1,45,46*]

How to curate a repeat library

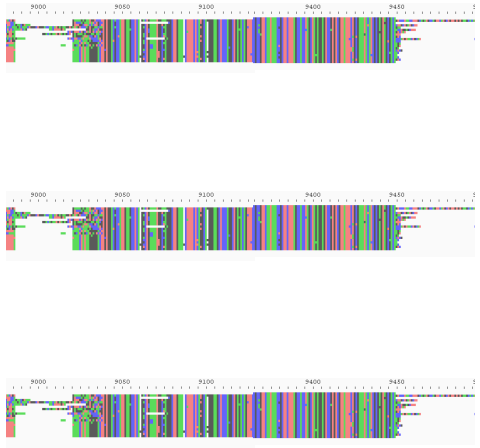
Overview

Library 1.0

```
>cons1
ATGAT
>cons2
GGATC
...
```

Align to the reference genome assembly

Collect the best hits + 2kb flanks and get a multisequence alignment for each consensus sequence



Get refined consensus sequences

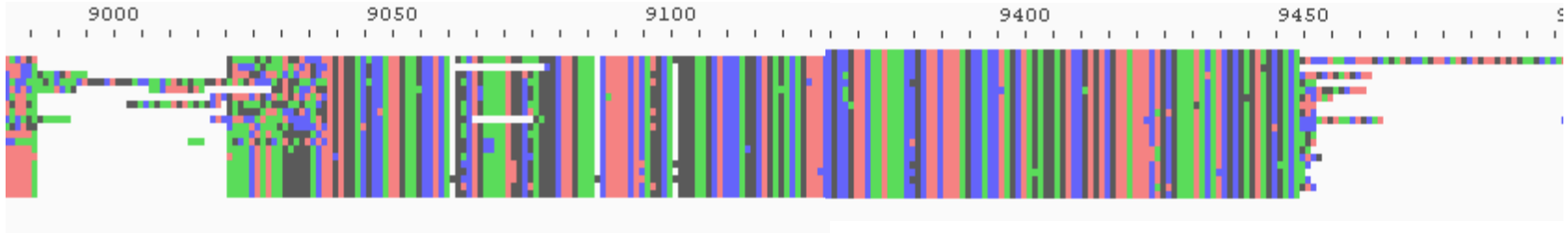
DOCUMENT THE PROCESS!!

Library 2.0

```
>cons1
ATGAC
>cons2
GCATC
...
```

What to look for and fix during curation

Part 1



0. Prepare the alignments and the Excel file to document every step of the curation
1. Create a new consensus sequence out of the alignment, e.g., Advanced Consensus Maker, built-in options in your alignment tool, CIAlign
2. Establish (temporary) boundaries of the new consensus sequence
3. Go through the alignment and fix ambigous bases and gaps by applying a majority rule
4. Pay attention to CpG sites
5. Find the correct boundaries + TSDs of the new consensus sequence – remove TSDs from consensus!!

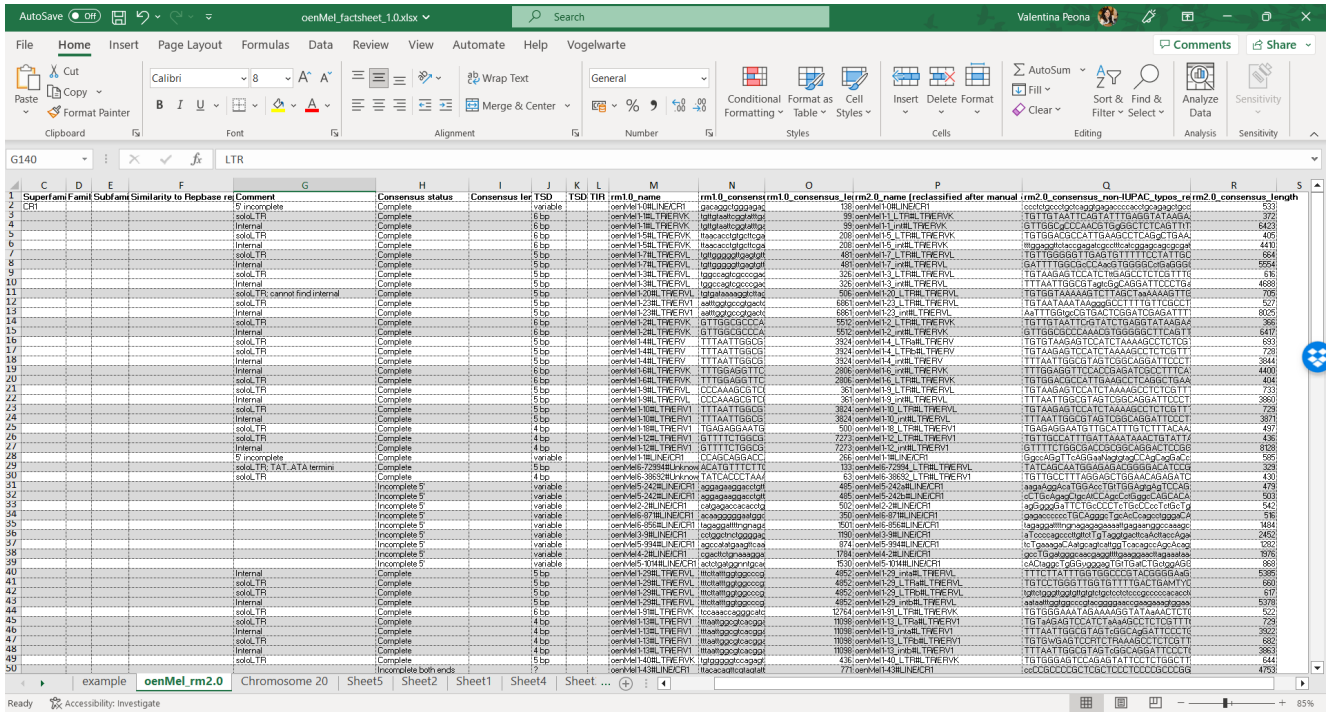
Alignments and documentation of the library

Align and extend process

RMDL_pipeline.pl + run_RMDL_pipeline.sh

- BLAST to align consensus sequence to reference genome
- Collect up to best 20 hits + flanking regions
- Align the best hits to the consensus sequence with MAFF

The code and citations can be found in the Github repository and on Google Drive (Practical2_part2)

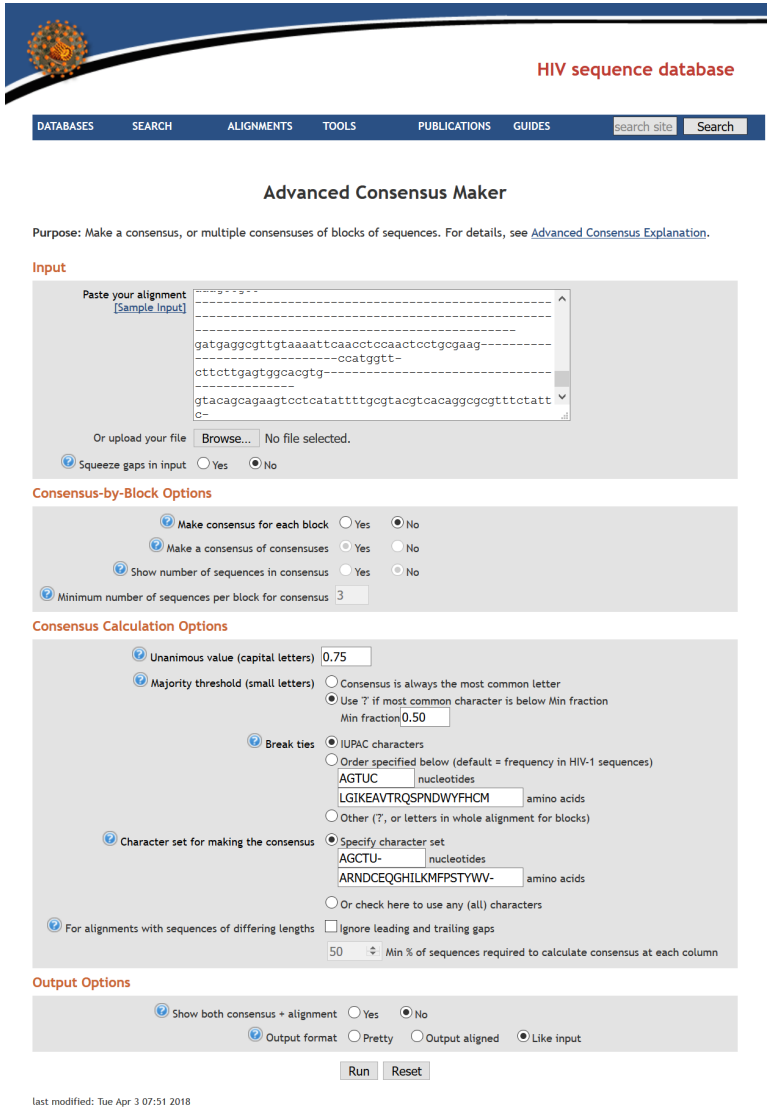


Empty template in the Google Drive folder: https://docs.google.com/spreadsheets/d/1FGPWUsdIibTqAK3GVIt3A0ZSXXzJp0XO/edit?usp=drive_link&ouid=116433763462028904068&rtpof=true&sd=true

1

Create a new consensus sequence

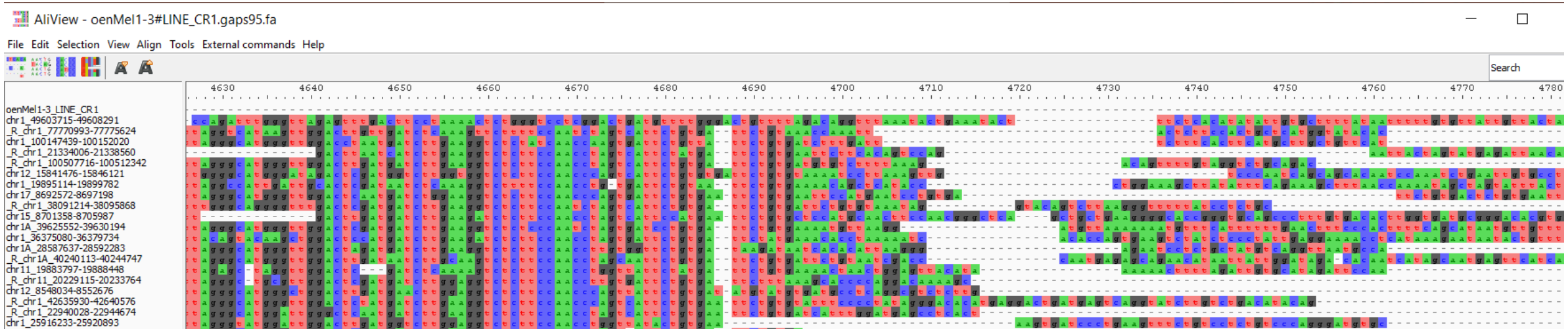
Example: Advanced Consensus Maker



- Go to the website: https://www.hiv.lanl.gov/content/sequence/CONSENSUS/AdvCon.html
- Paste your alignment
- Set parameters (example in figure)
- Get new consensus sequence and copy-paste it back into your alignment

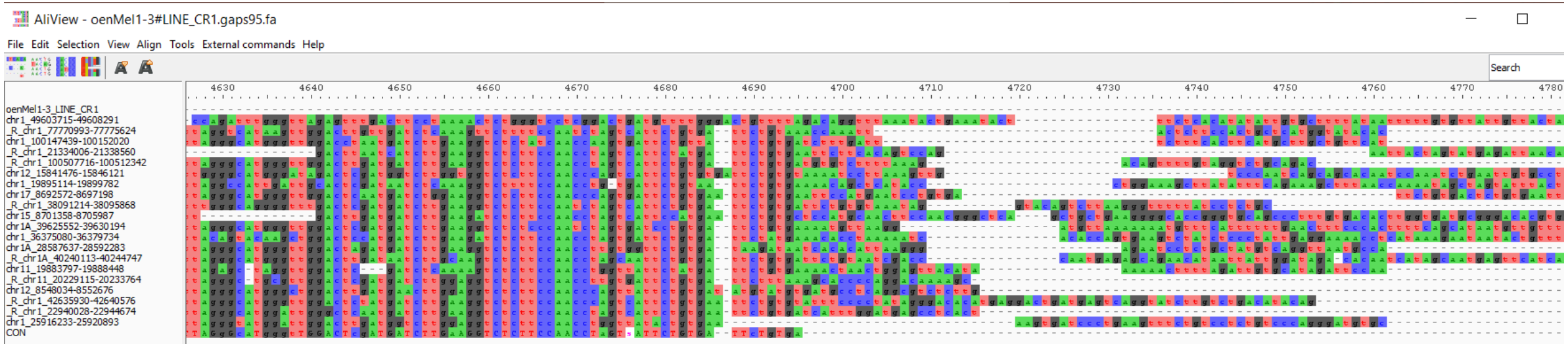
1

Create a new consensus sequence



1

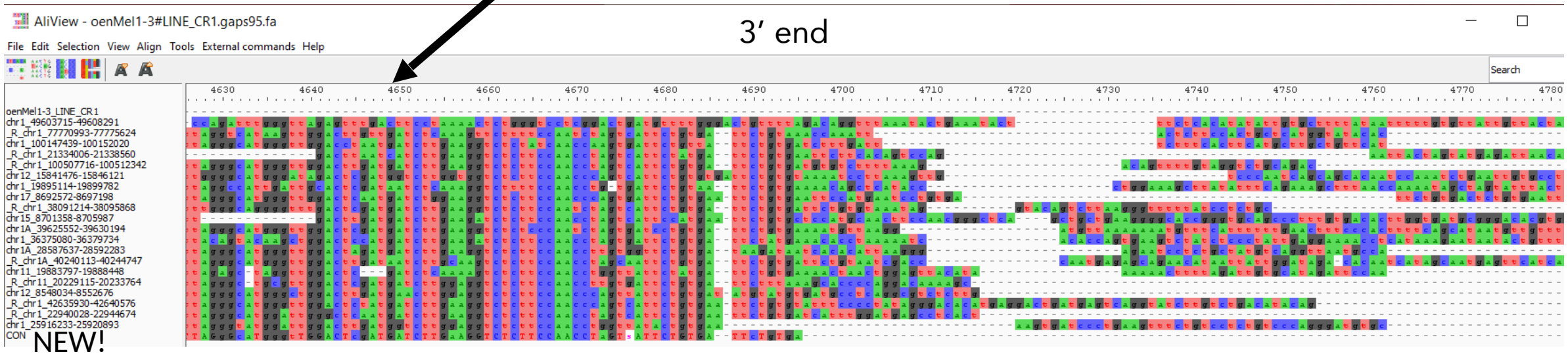
Create a new consensus sequence



2

Find the termini of the consensus sequence

Original consensus is incomplete



NEW!

2

Find the termini of the consensus sequence

Original consensus is incomplete

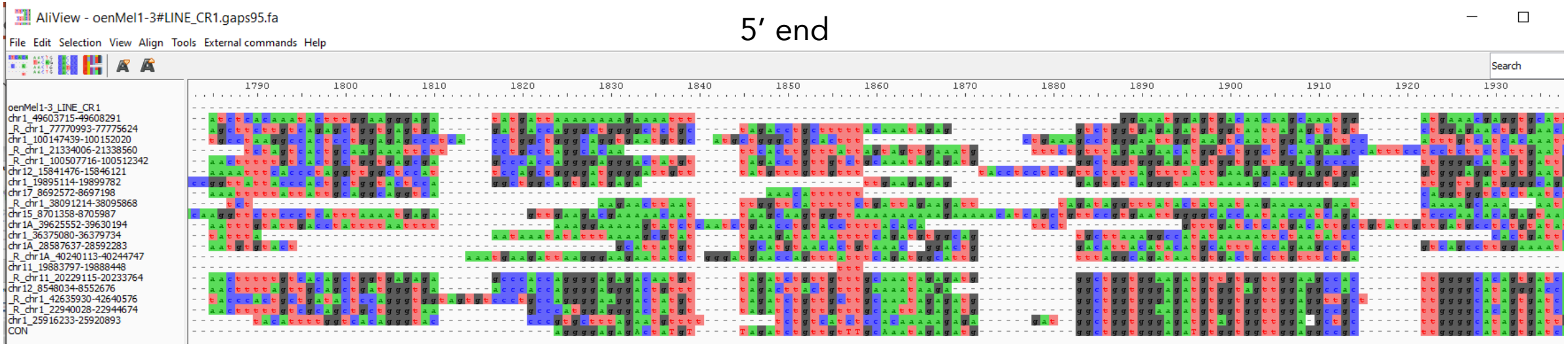
NEW!

2

Find the termini of the consensus sequence

Original consensus is incomplete

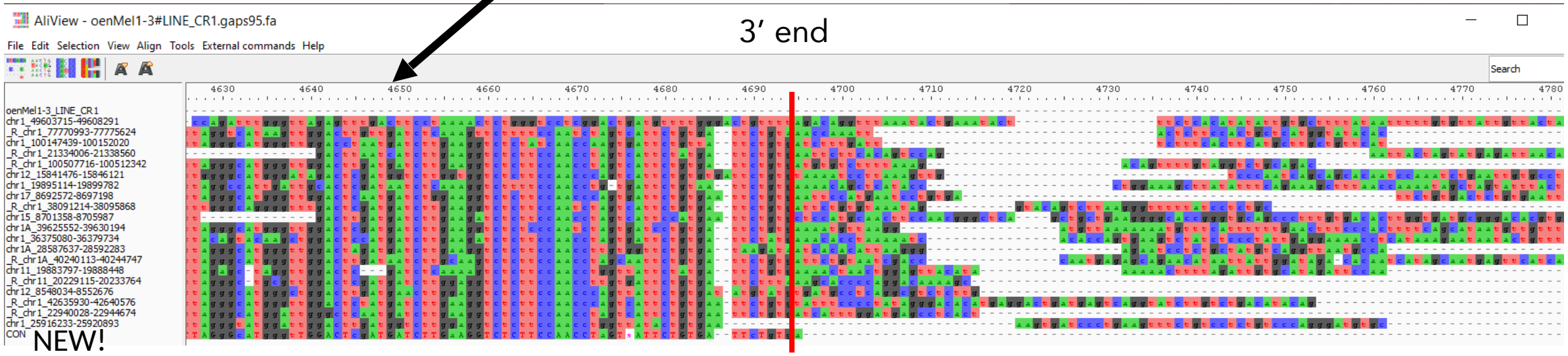
NEW!

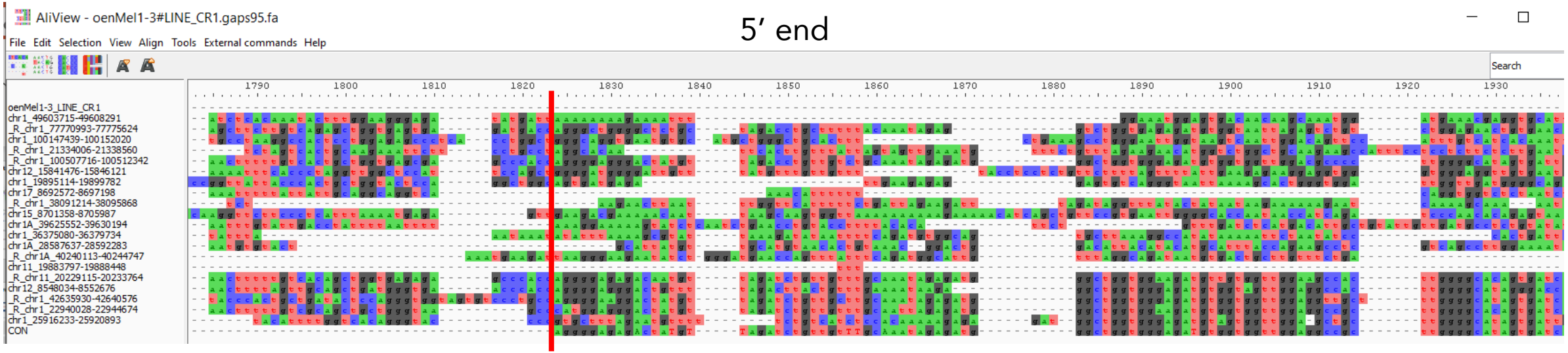


2

Find the termini of the consensus sequence

Original consensus is incomplete

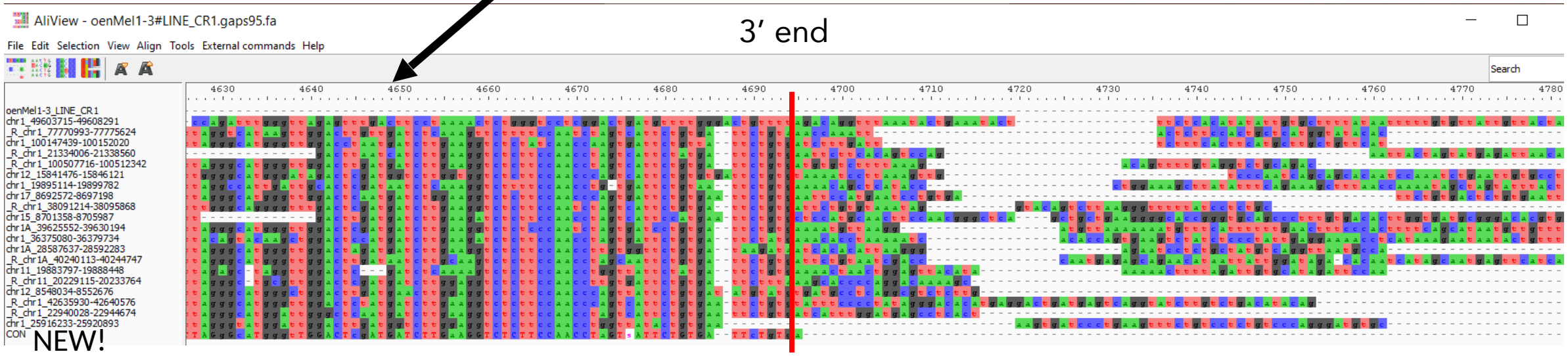


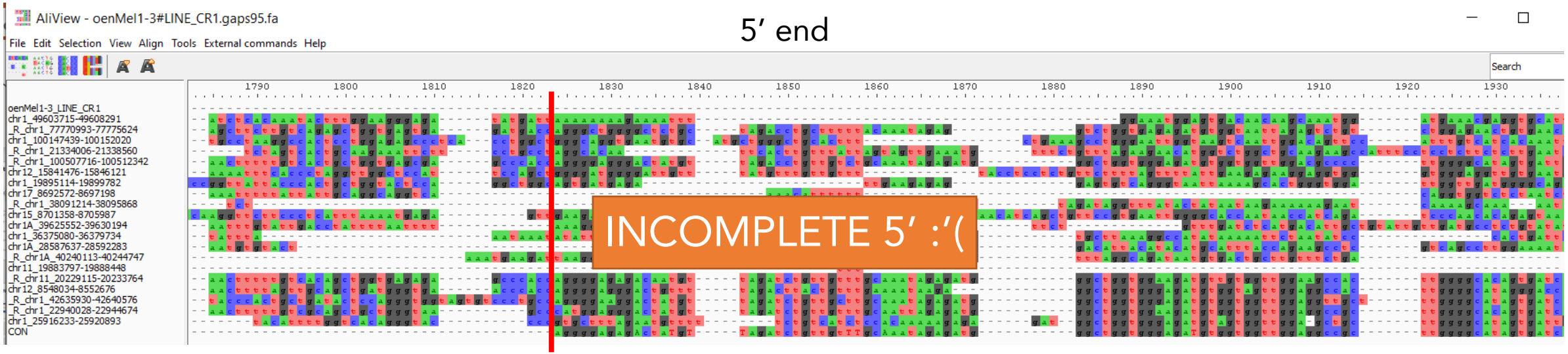


2

Find the termini of the consensus sequence

Original consensus is incomplete

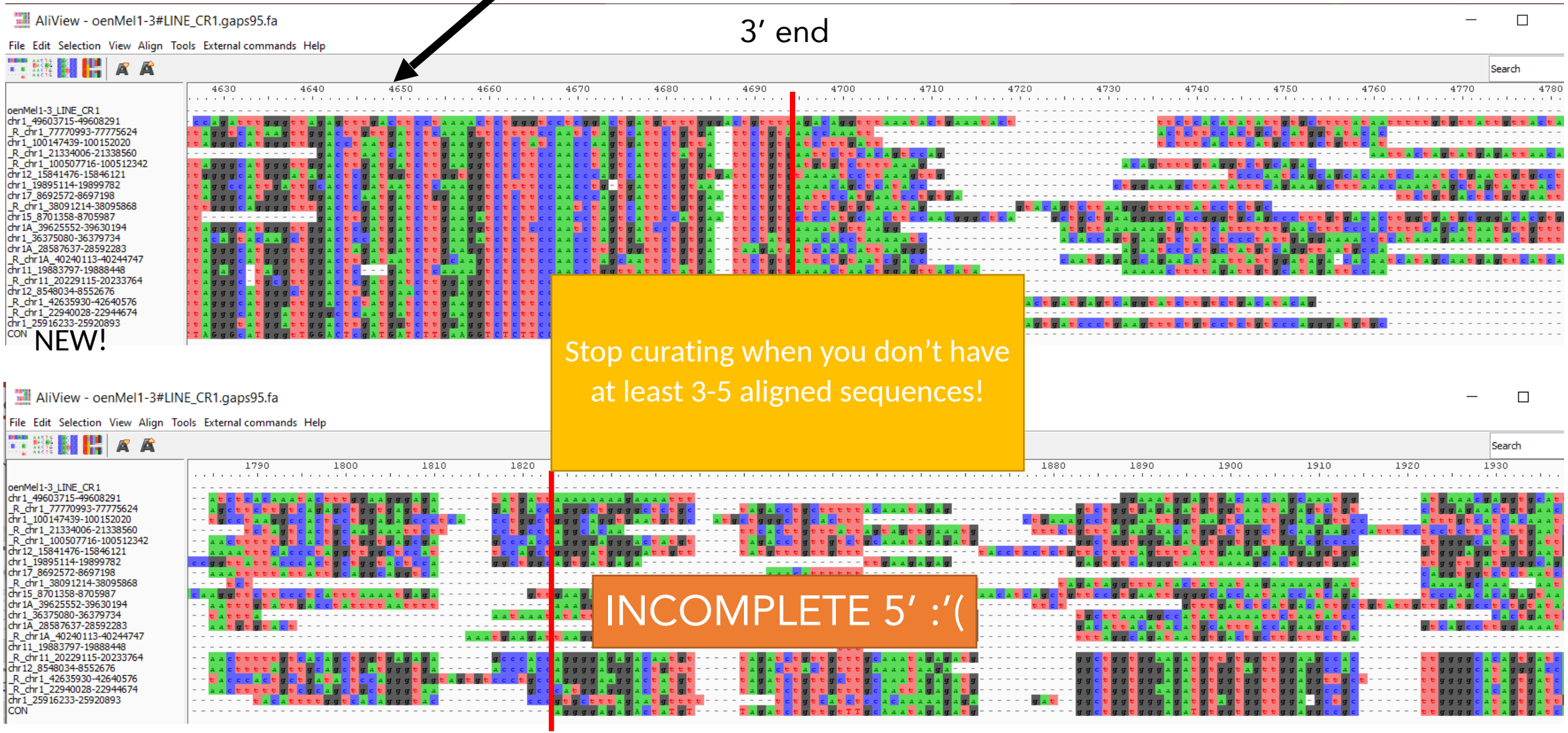




2

Find the termini of the consensus sequence

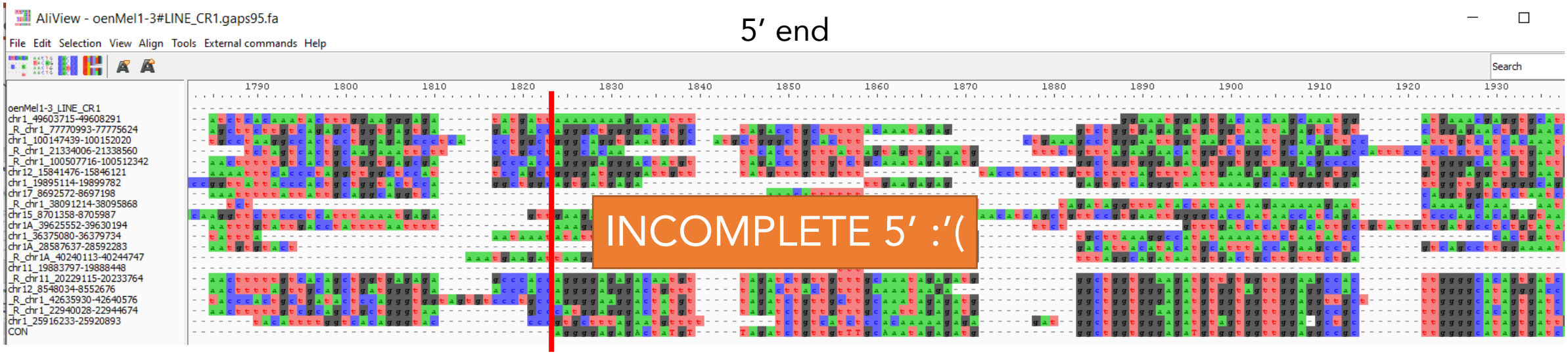
Original consensus is incomplete



2

Find the termini of the consensus sequence

Example: 5’ truncation



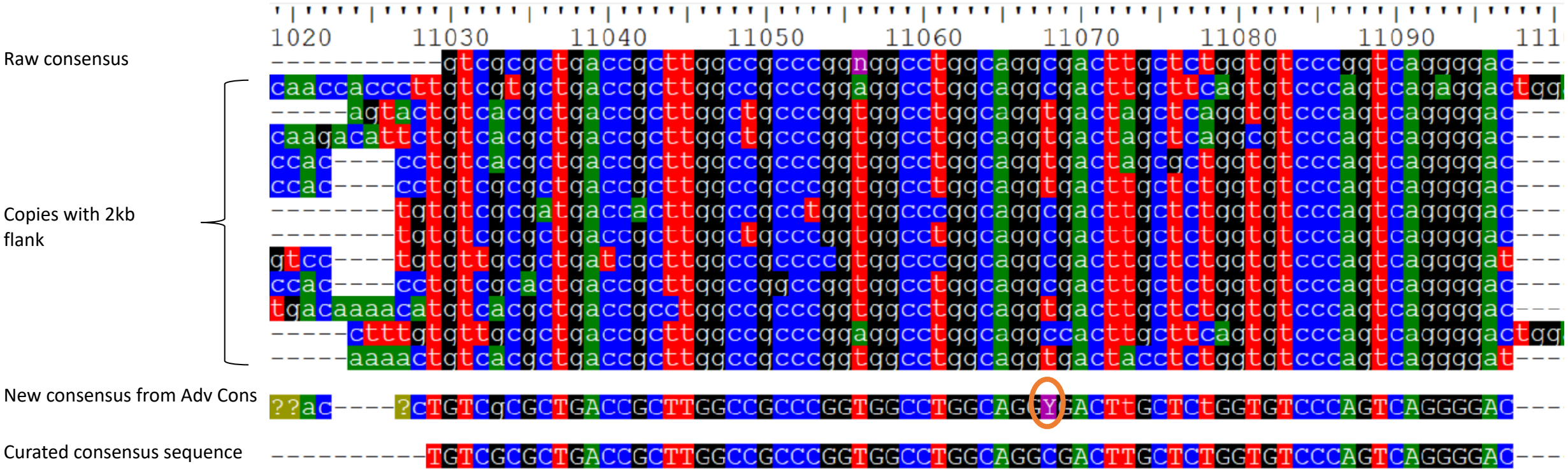
Run another round of alignments by chopping ~500 bp at the 5’ end and extend the consensus

Annotate that the consensus sequence is 5’ incomplete (colunm G) and you can add ‘.inc’ to the name before classification tag

Note: it is very difficult to find full-length LINE elements to build a complete consensus

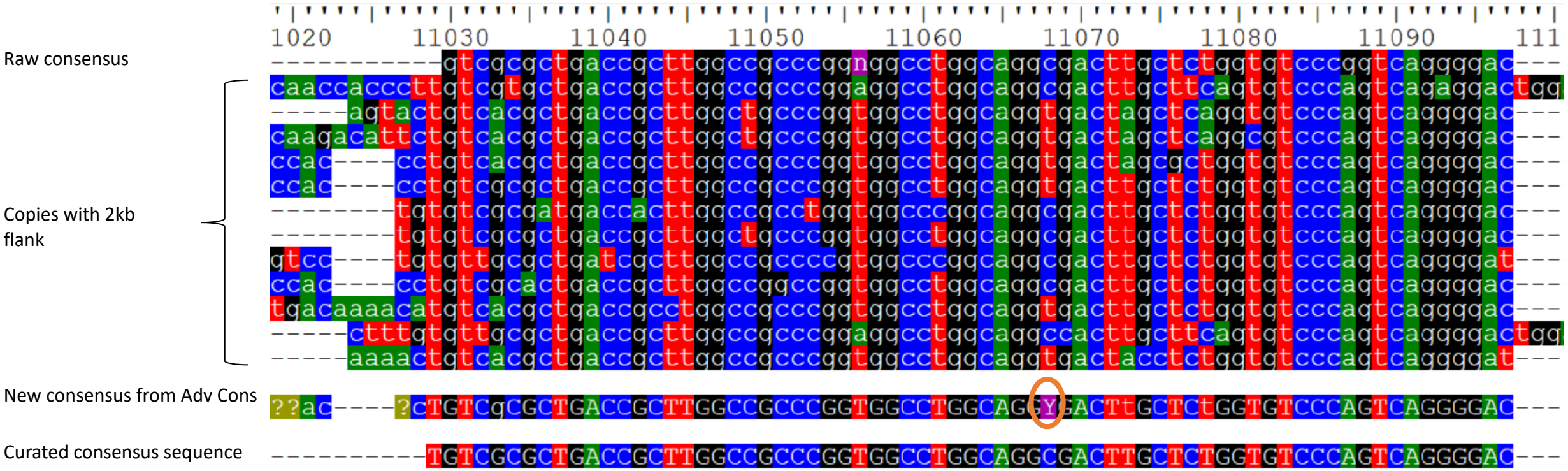
3

Majoritiy-rule consensus on SNPs



3

Majoritiy-rule consensus on SNPs

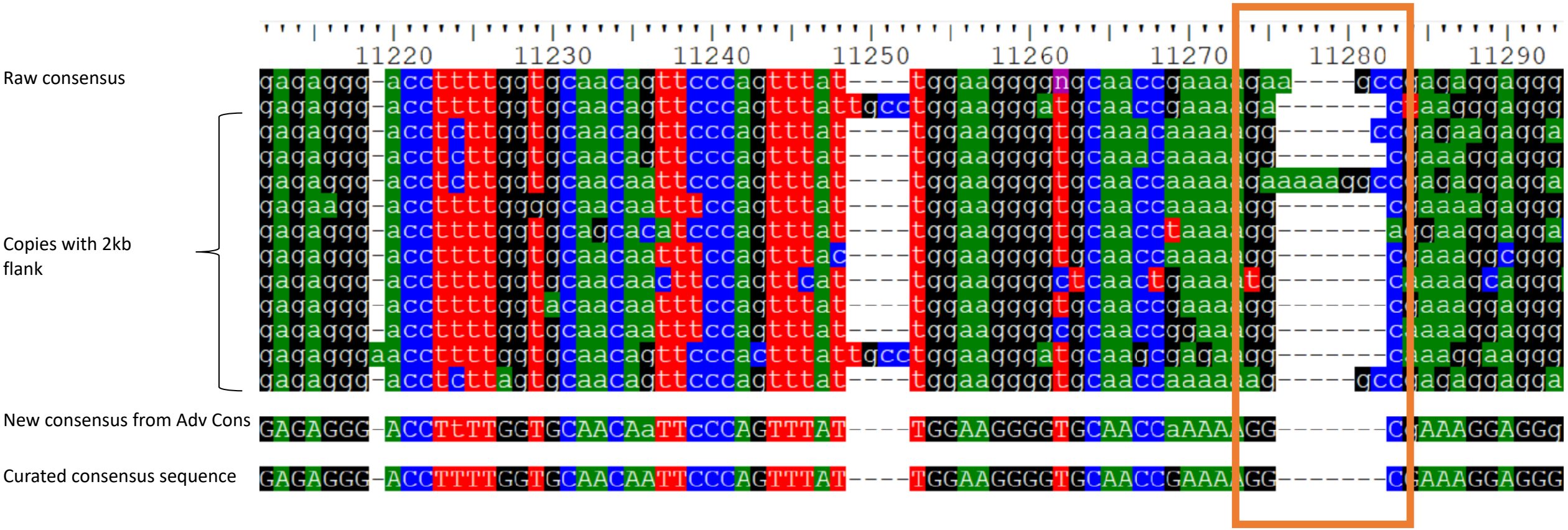


Get to the most accurate IUPAC character

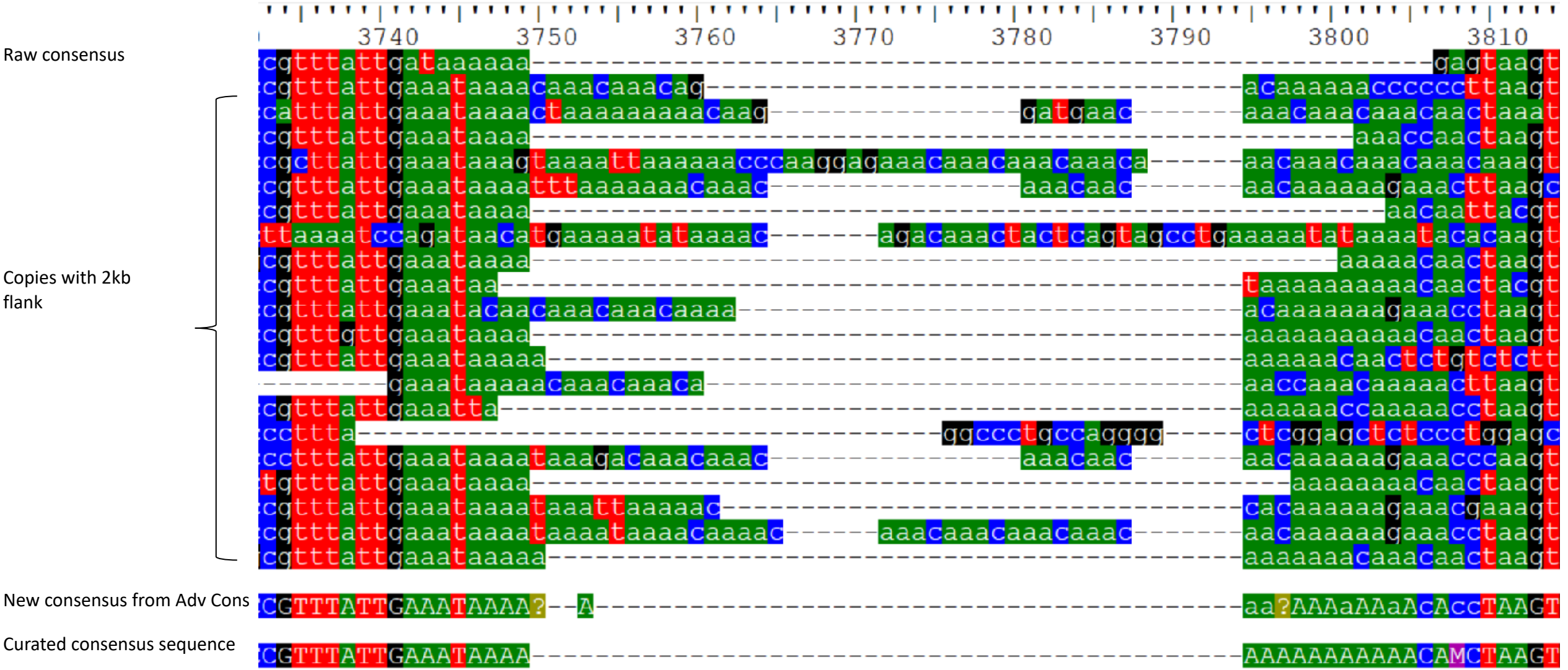
IUPAC Code	Meaning
M	A/C
R	A/G
W	A/T
S	C/G
Y	C/T
K	G/T
V	A/C/G
H	A/C/T
D	A/G/T

3

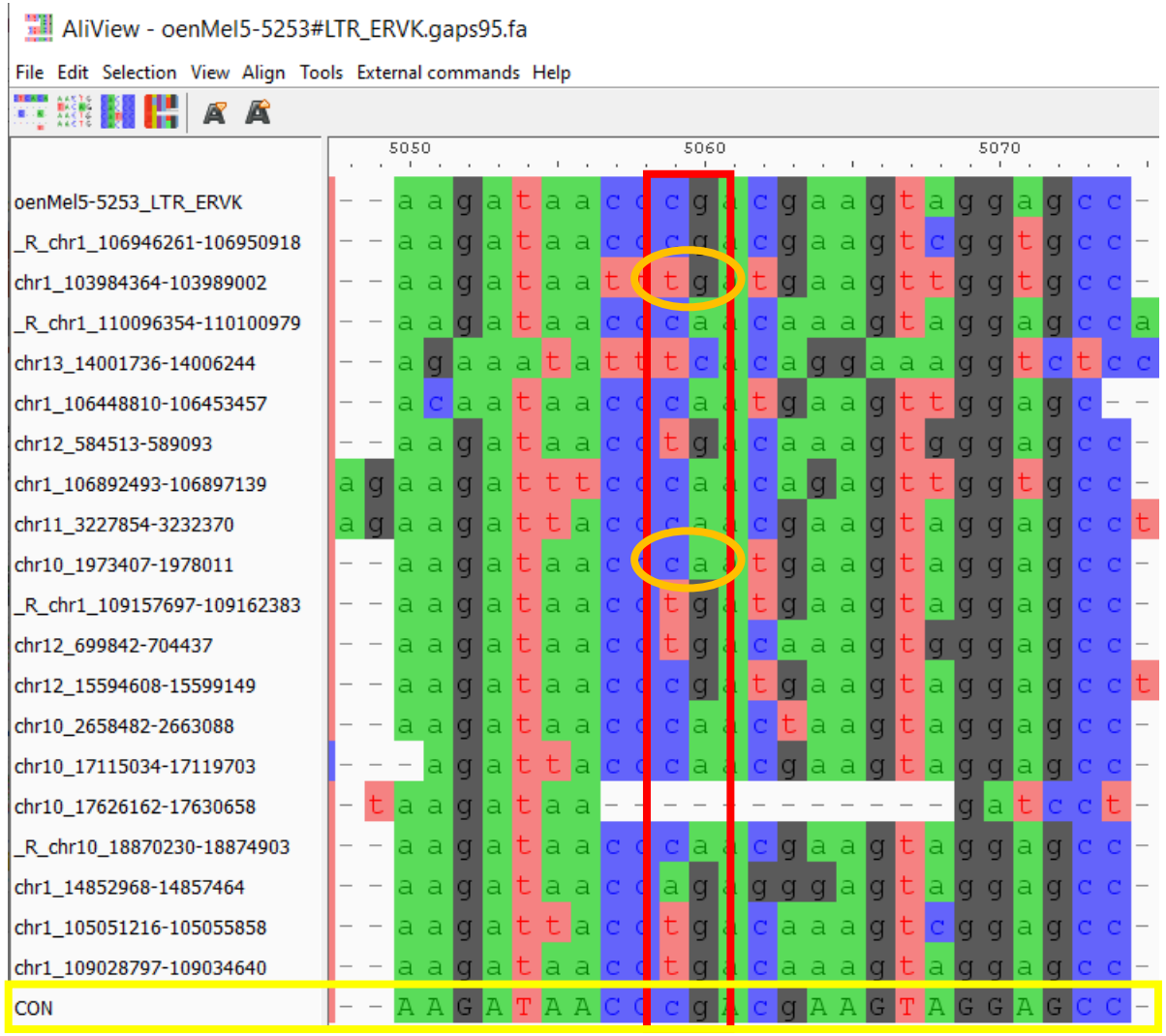
Majority-rule consensus on indels



Tricky regions



Attention to CpG sites



- You can choose to reconstruct the ancestral state CG
- I suggest to do it so to not inflate the divergence from consensus

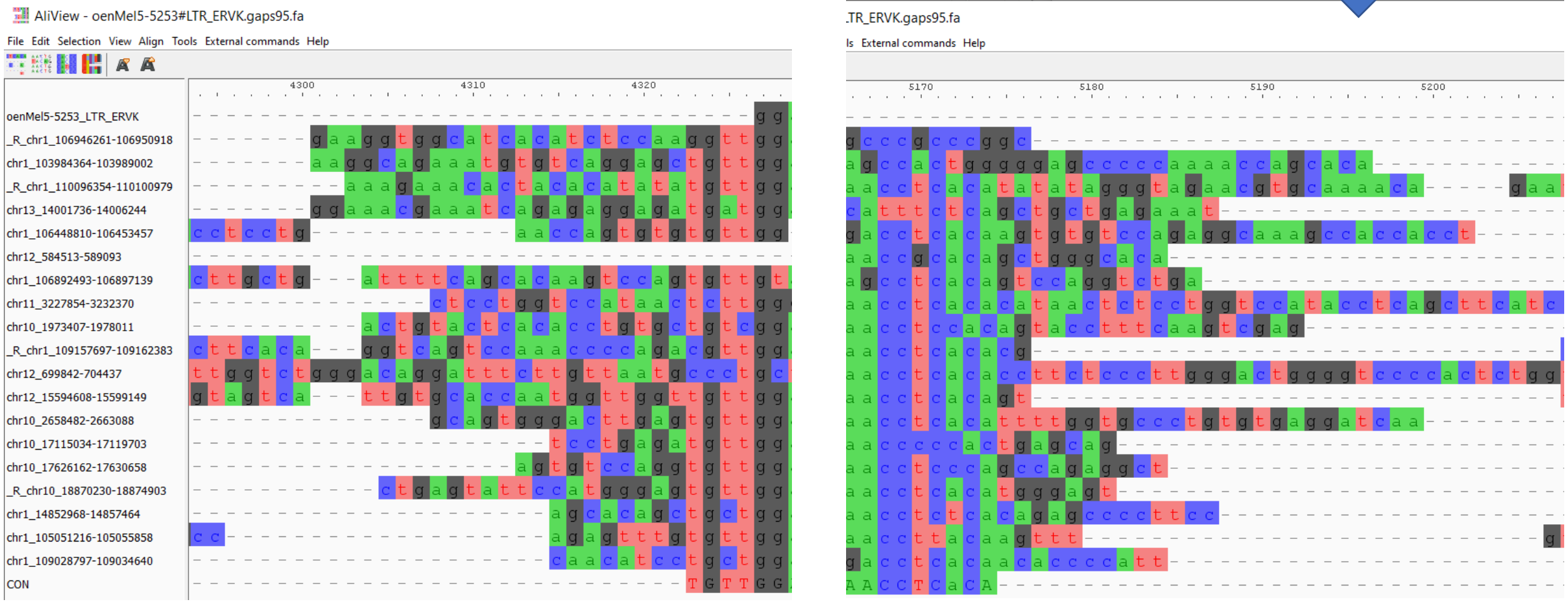
Find the TSDs

(if any!)

Start looking from the 3’ terminus

5’ end

3’ end



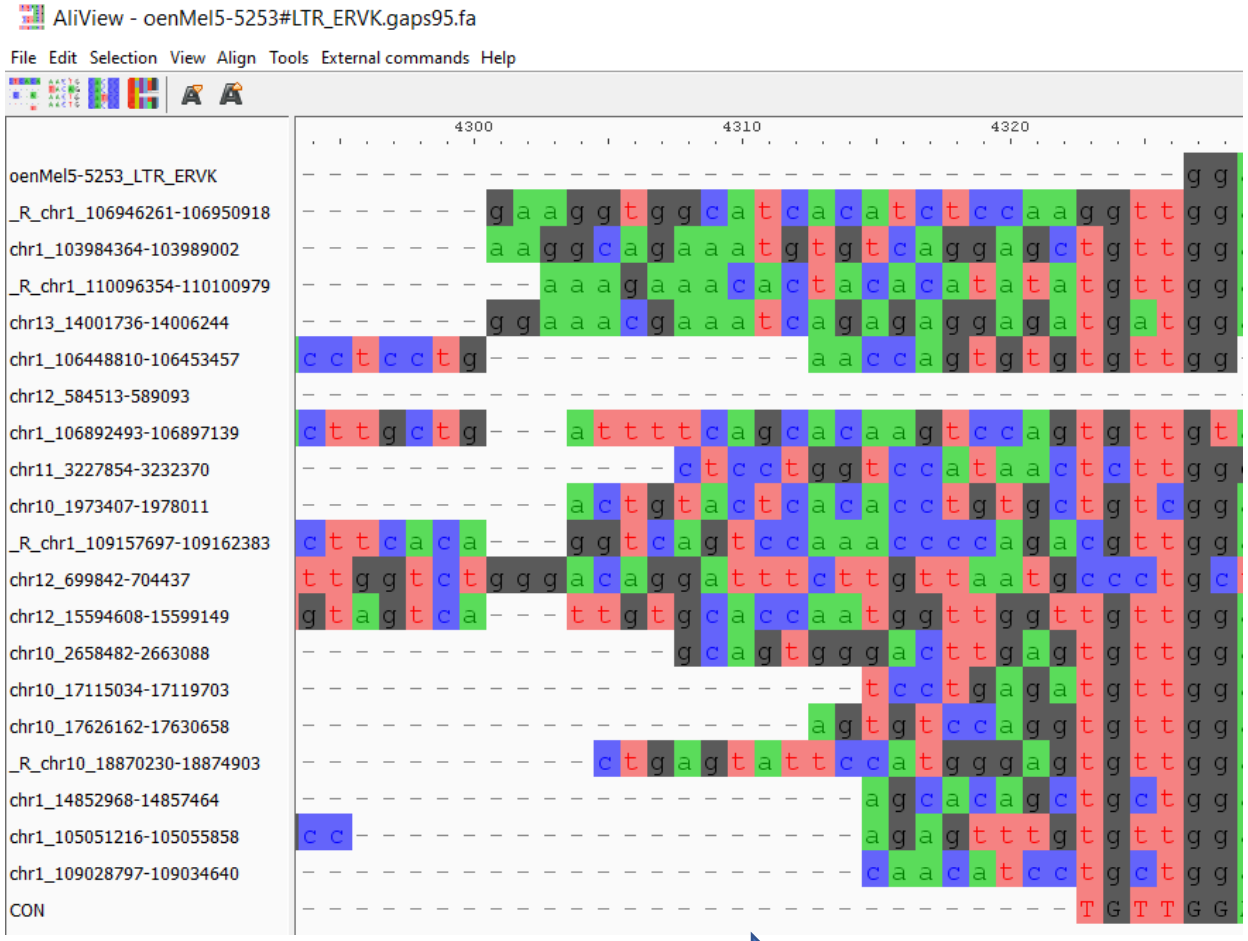
Find the TSDs

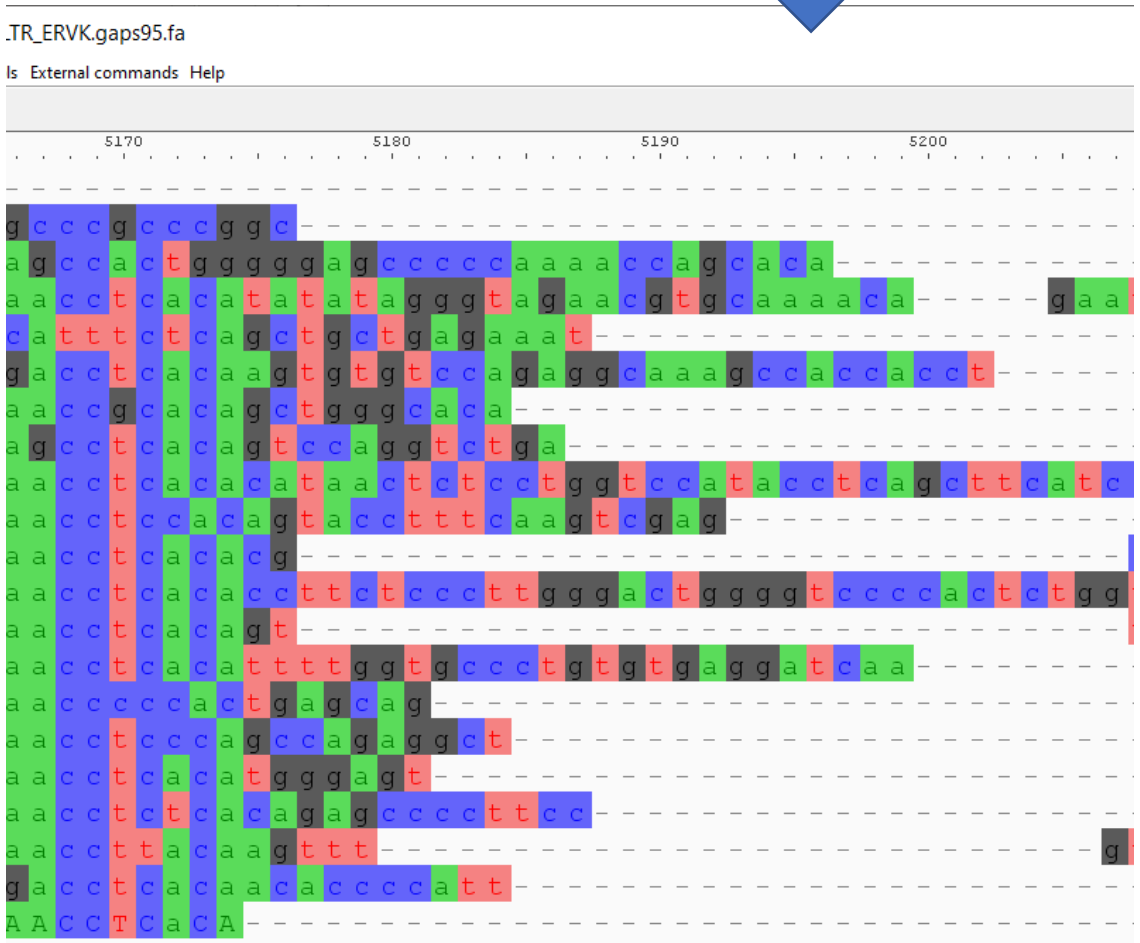
(if any!)

Start looking from the 3’ terminus

5’ end

3’ end





LTRs and LTR elements usually start with 5’-TG …. CA-3’

5

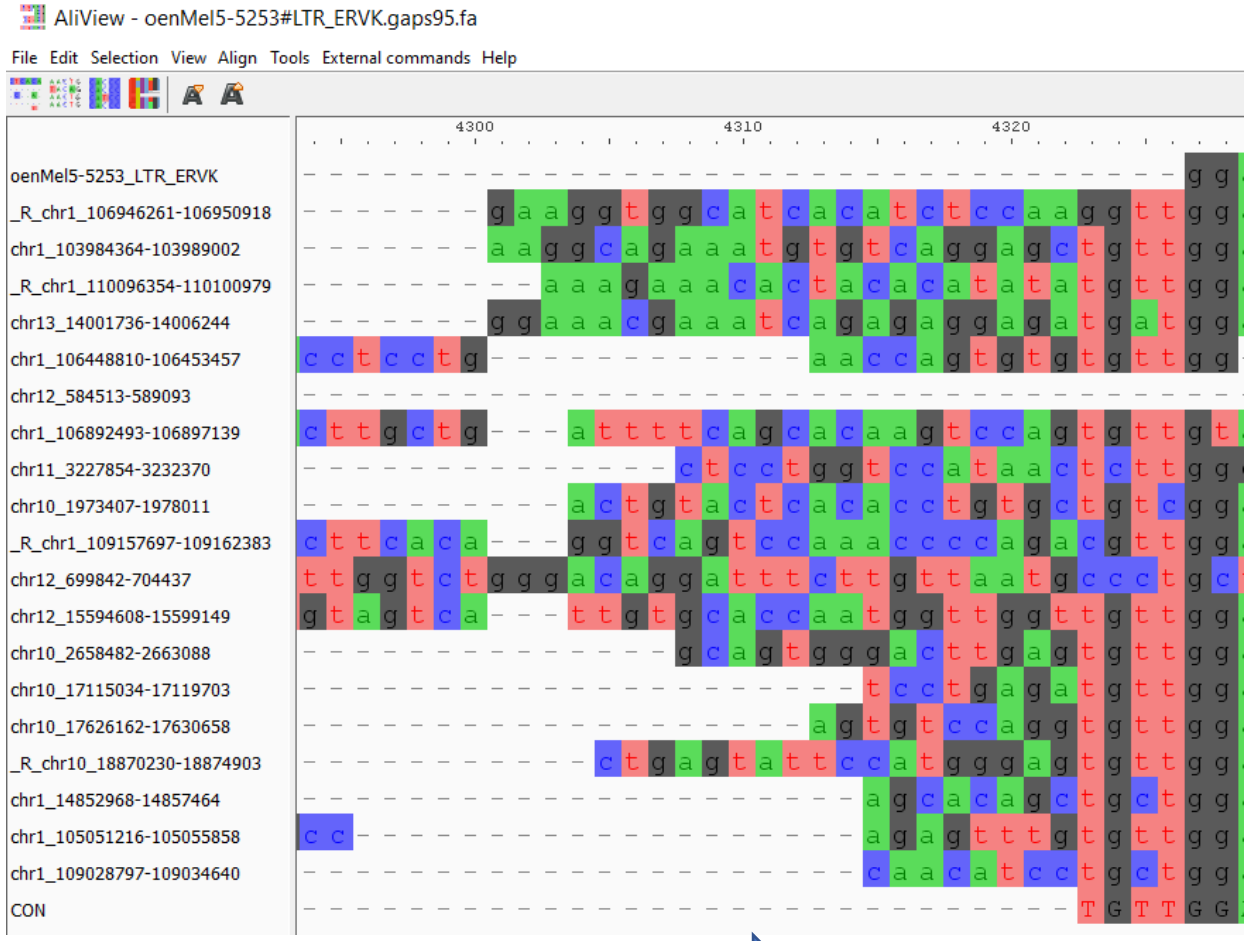
Find the TSDs

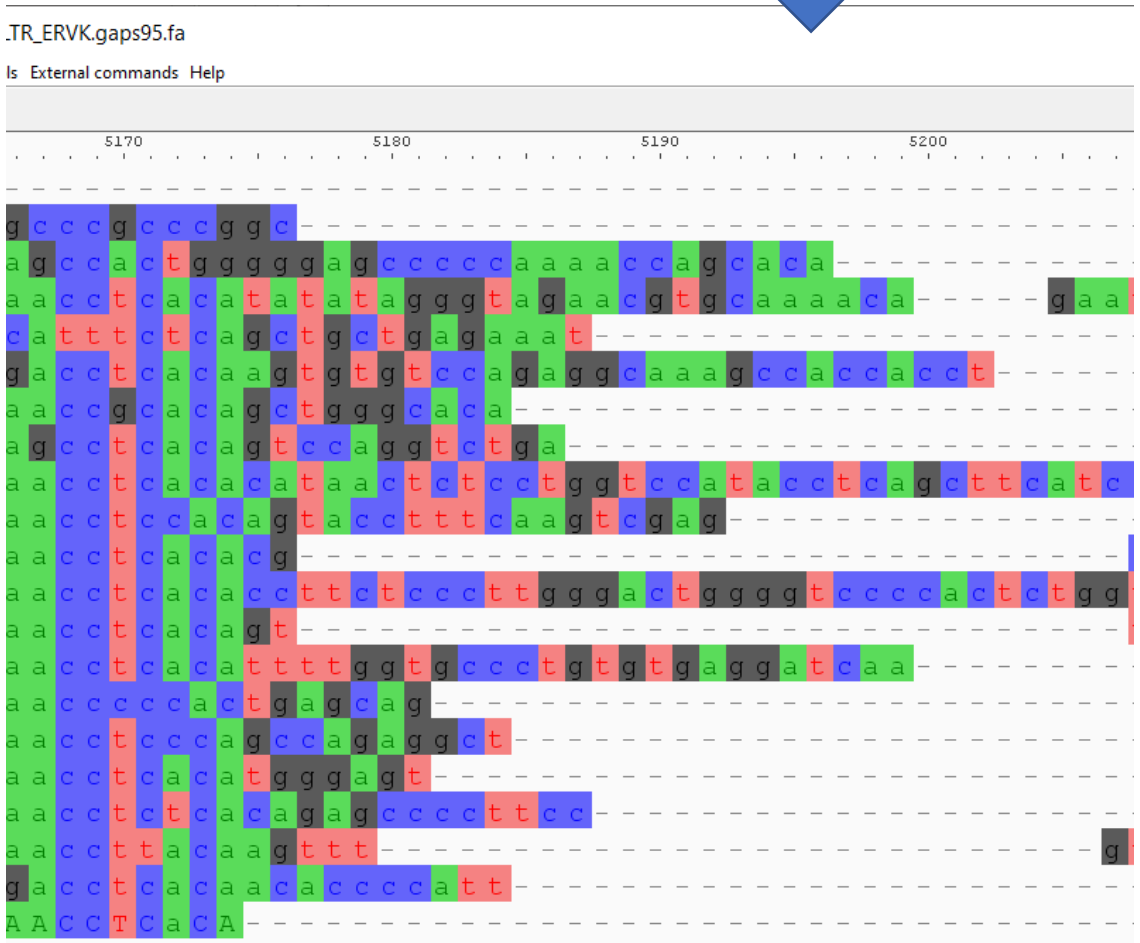
(if any!)

Start looking from the 3’ terminus

5’ end

3’ end

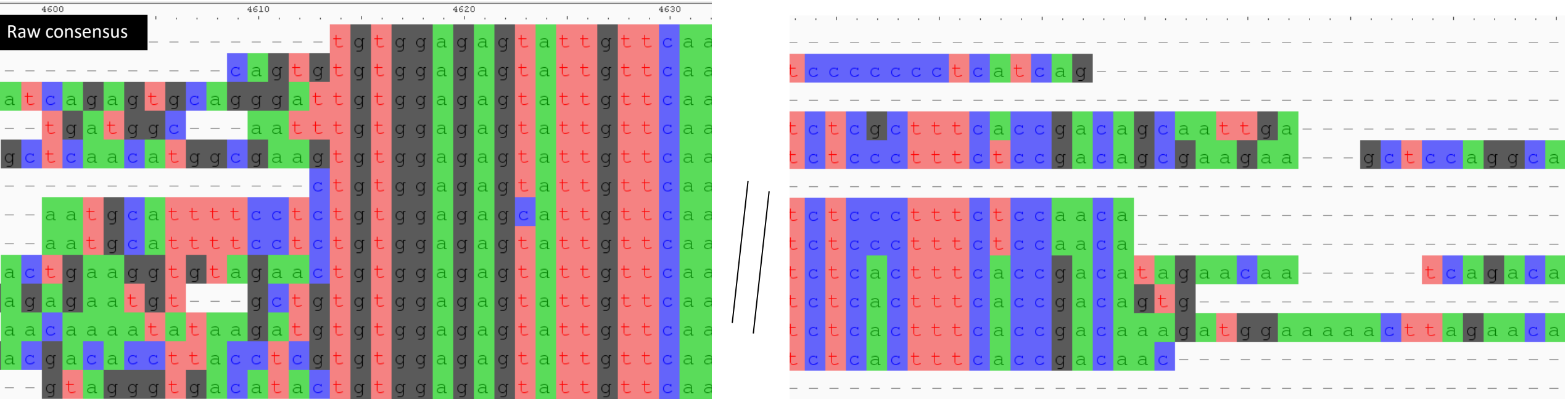




LTRs and LTR elements usually start with 5’-TG …. CA-3’

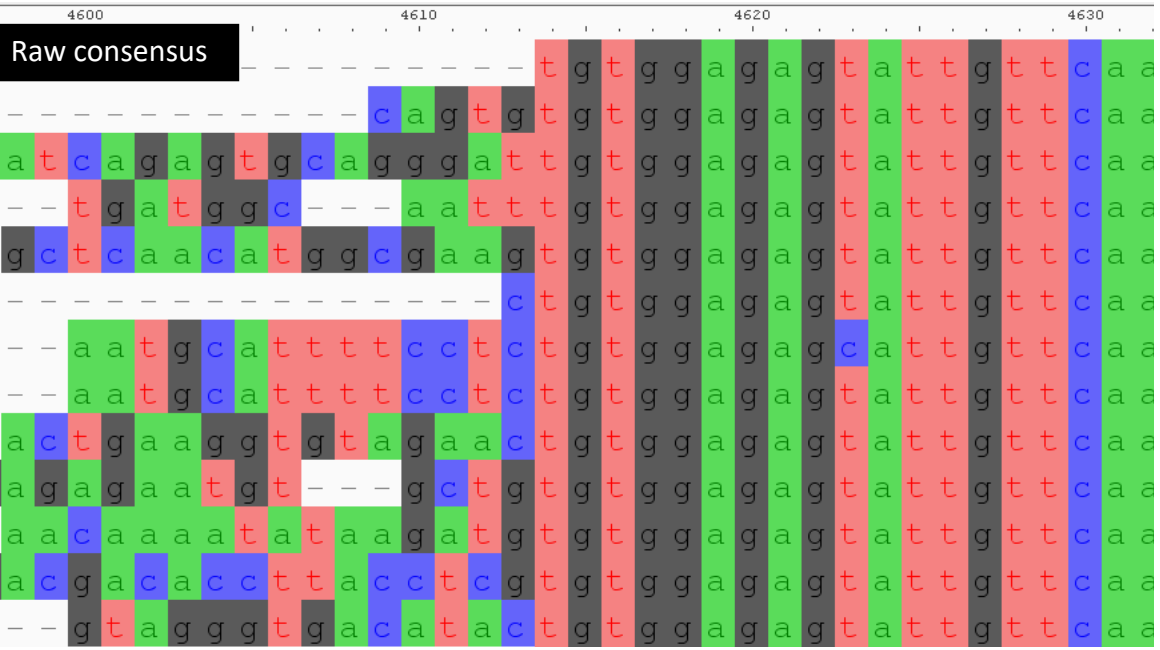
Annotate that the consensus sequence has a 6 nt TSD in column J

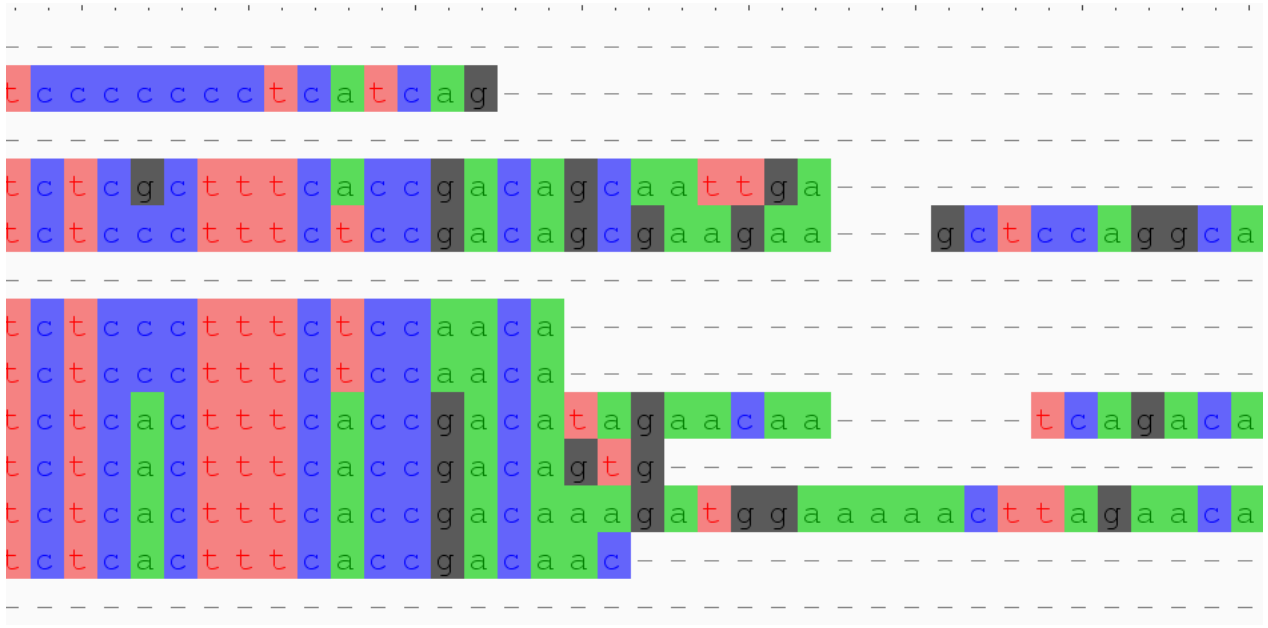
Find the TSDs



5

Find the TSDs

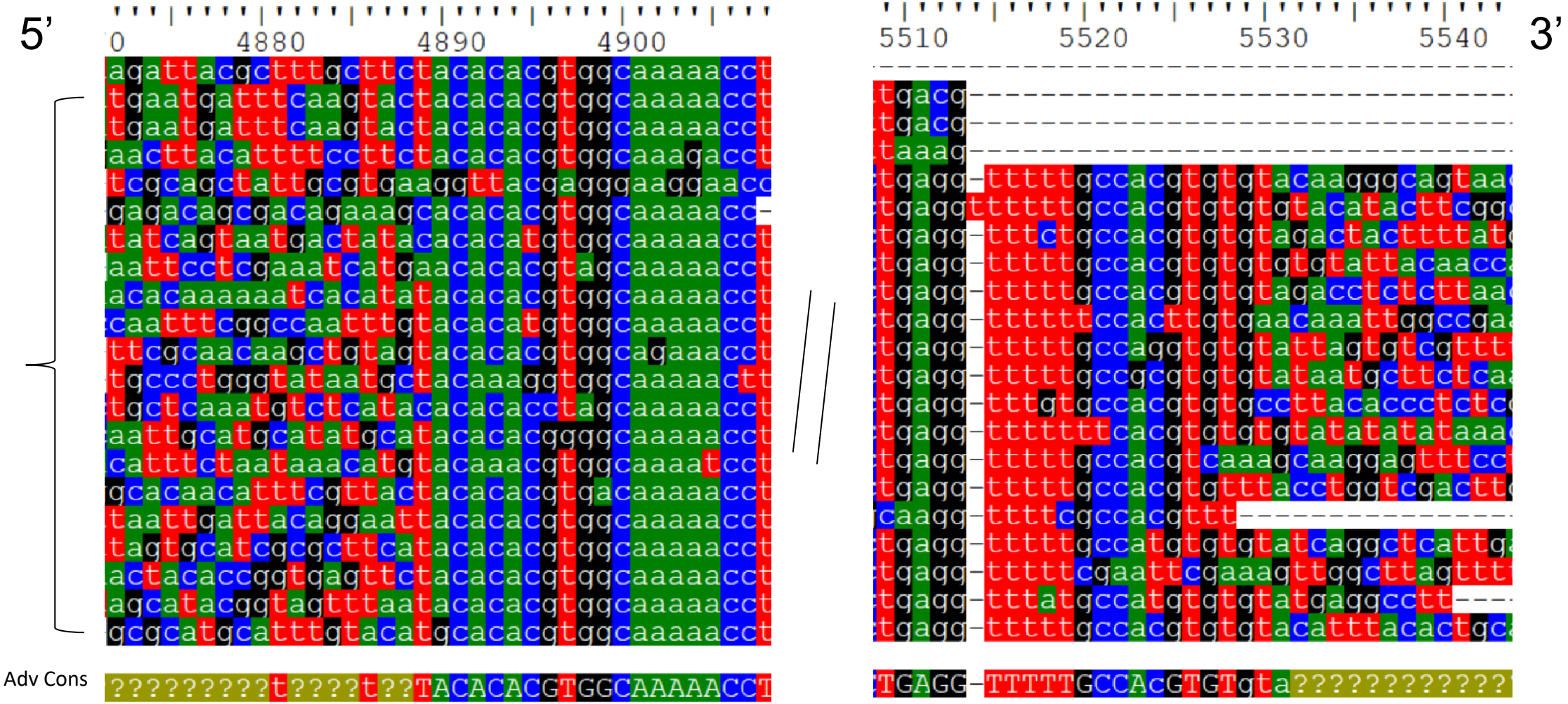




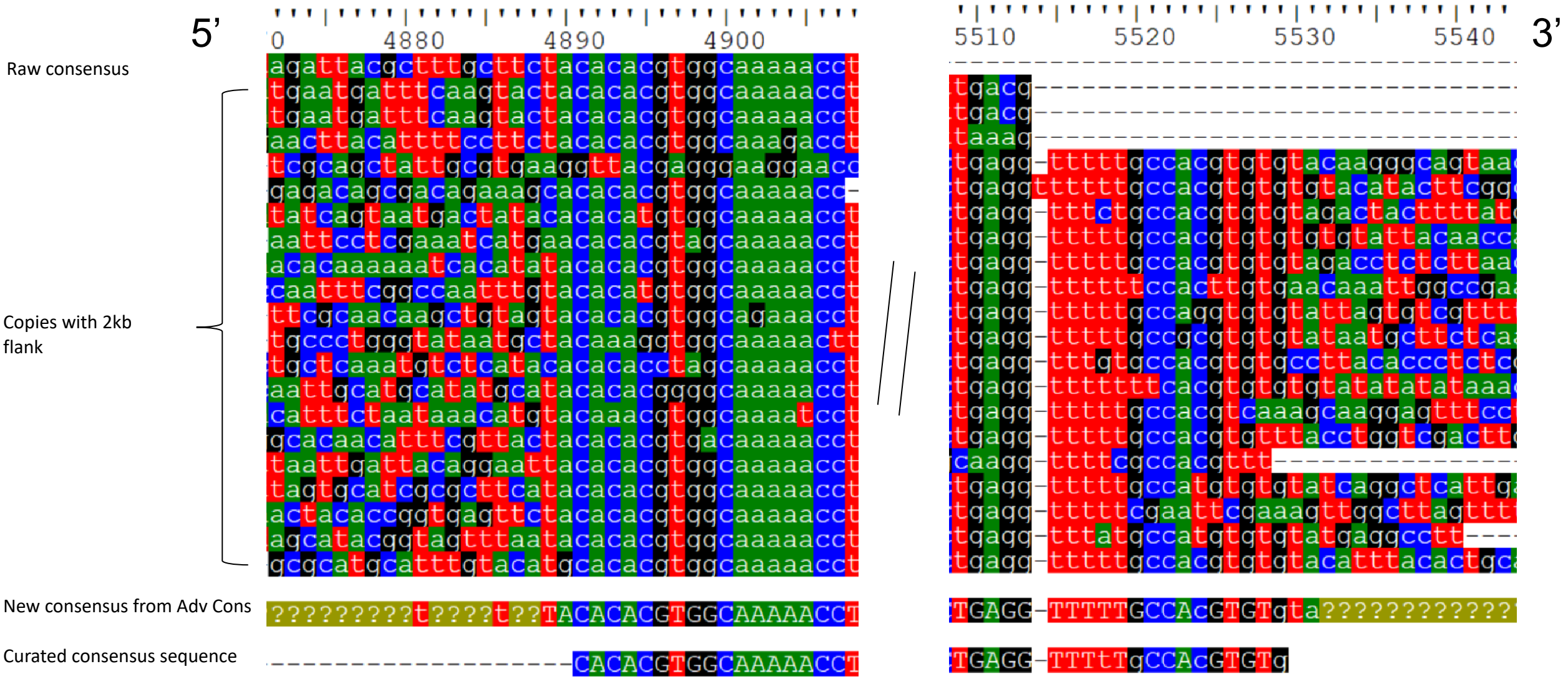
In the presence of TSD motifs, annotate the motif in column K

5

Find the TSDs

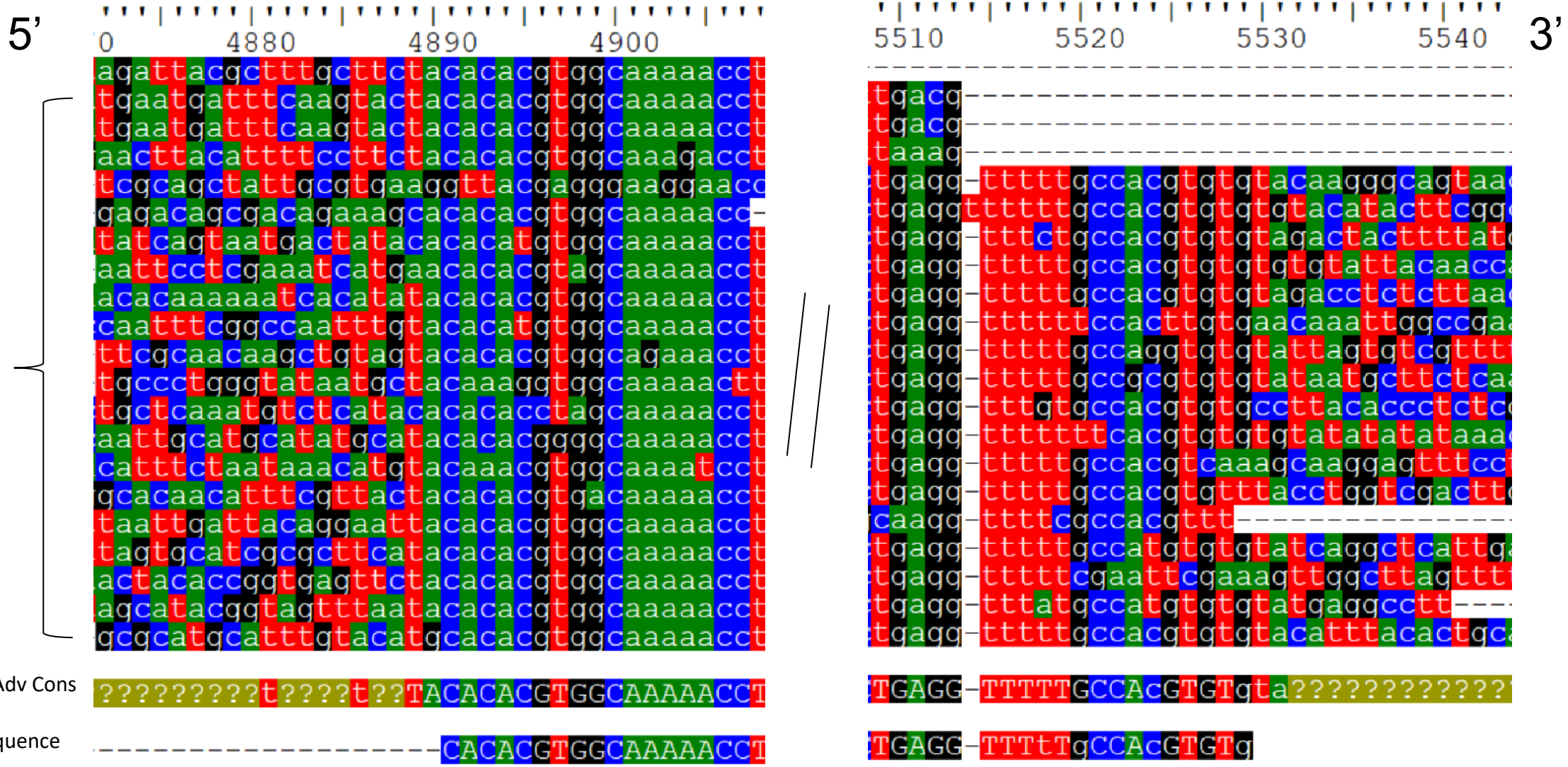


Find the TSDs



5

Find the TSDs



In the presence of TSD motifs, annotate the motif in column K

DNA transposons and TSDs

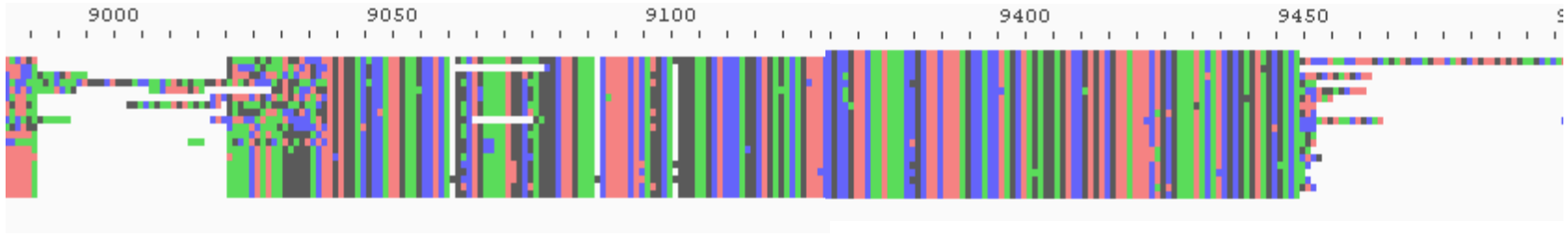
Table 1 Classification and characteristics of eukaryotic DNA transposons

Superfamily	Related IS	TSD	Length[1] (kb)	TIRs[1] (bp)	Terminal motif (5′-3′)	TPase[1] (aa)	Catalytic motif	DNA-binding motif	Additional proteins
Tc1/*mariner*	IS630	TA	1.2–5.0	17–1100	Variable	300–550	DD(30–41)D/E	HTH (cro/paired)	
hAT	nd	8 bp	2.5–5	10–25	YARNG	600–850	D(68)D(324)E[2]	ZnF (BED)	
P element	nd	7/8 bp	3–11	13–150	CANRG	800–900	D(83)D(2)E(13)D[3]	ZnF (THAP)	
MuDR/ Foldback	IS256	7–10 bp	1.3–7.4	0-sev. Kb	Variable	450–850	DD(~110)E	ZnF (WRKY/GCM1)	
CACTA	nd	2/3 bp	4.5–15	10–54	CMCWR	500–1200	nd	nd	TNPA (DNA-binding protein)
PiggyBac	IS1380	TTAA	2.3–6.3	12–19	CCYT	550–700	DDE?	nd	
PIF/ Harbinger	IS5	TWA	2.3–5.5	15–270	GC-rich	350–550	DD(35–37/ 47–48)E	HTH	PIF2p (Myb/SANT domain)
Merlin	IS1016	8/9 bp	1.4–3.5	21–462	GGNRM	270–330	DD(36–38)E	nd	
Transib	nd	5 bp	3–4	9–60	CACWATG	650–700	DD(206–214)E	nd	
Banshee	IS481	4/15 bp	3–5	41–950	TGT	300–400[4]	DD(34)E	HTH	
Helitron	IS91	none	5.5–17	none	5′-TC…CTAR-3′	1400–3000[5]	HHYY (“REP motif”)	ZnF-like	RPA (in plants)
Maverick	none	5/6 bp	15–25	150–700	Simple repeat	350–450[4]	DD(33–35)E	ZnF (HHCC)	4–10 DNA virus-like proteins

[1]Refers to a potentially complete, autonomous element.
[2]Motif in *Hermes* TPase.
[3]Motif in Drosophila P element TPase.
[4]RVE integrase-like.
[5]REP-Helicase.
nd = not determined.

What to look for and fix during curation

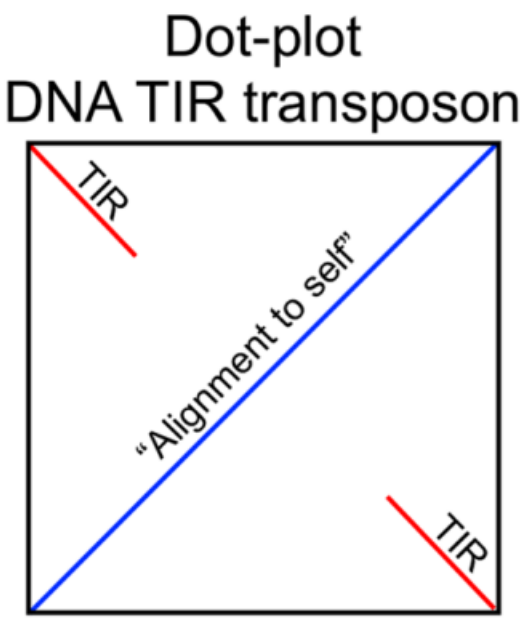
Part 2

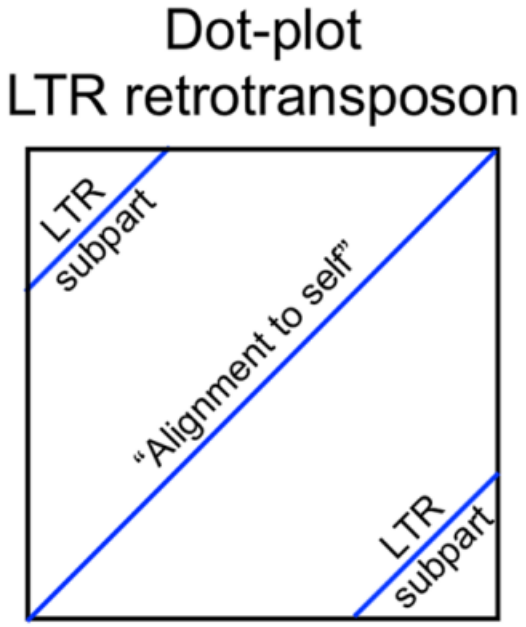


1. Check for the presence of LTRs or TIRs (e.g., dotplots)
2. Check for protein domains
3. Check for nucleotide homology
4. Re-classify the element

Check for LTRs and TIRs

How to read the dotplot of a self-alignment





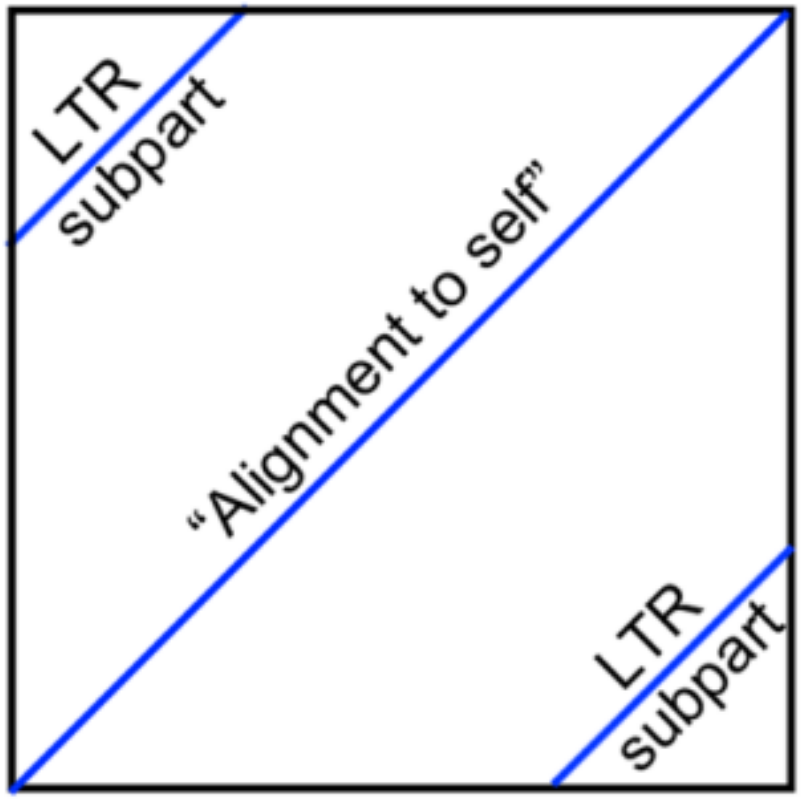
Use MAFFT for dotplots!

https://mafft.cbrc.jp/alignment/server/

1

Check for LTRs

LTRs = Long Terminal Repeats (direct repeats)



Separate the LTRs from the internal sequence (one of the two LTRs – they should be nearly identical)

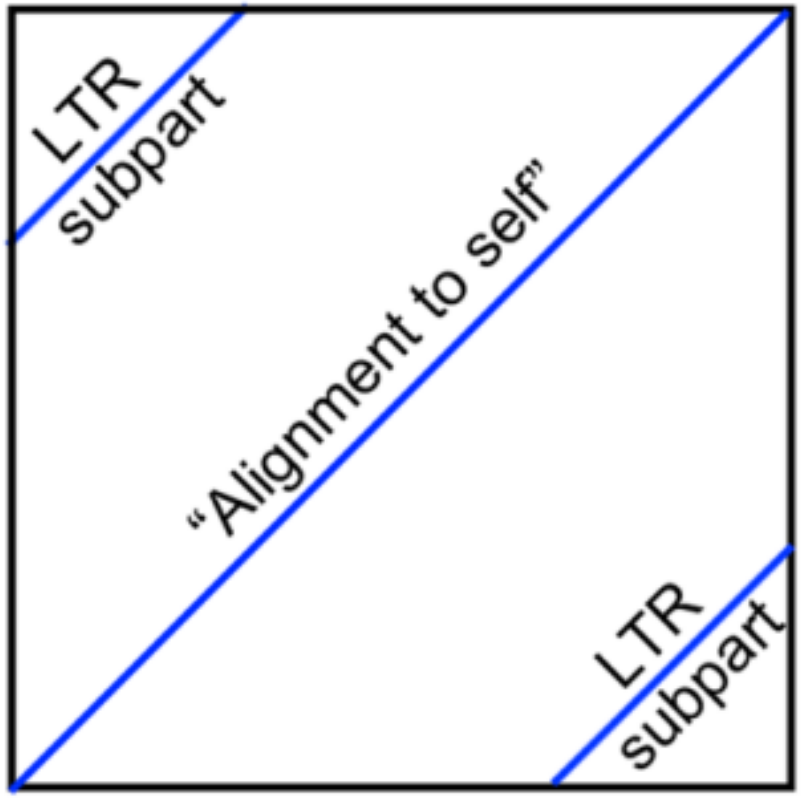
You will have a row with the LTR and a second row with the internal region in the Excel table

The name of the consensus sequence remain the same but you add the suffiy '_LTR' and '_int' (or '_I') just before the classification tag

1

Check for LTRs

LTRs = Long Terminal Repeats (direct repeats)



Separate the LTRs from the internal sequence (one of the two LTRs – they should be nearly identical)

You will have a row with the LTR and a second row with the internal region in the Excel table

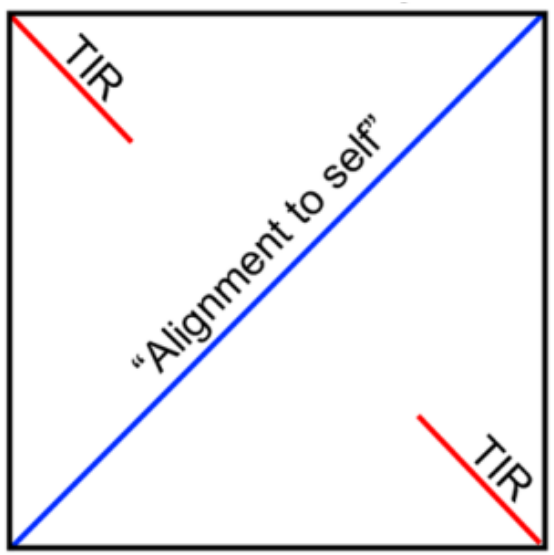
The name of the consensus sequence remain the same but you add the suffiy '_LTR' and '_int' (or '_I') just before the classification tag

Solo-LTR cannot give this signal! Remember that LTRs usually start with TG and ends with CA

1

Check for TIRs

TIRs = Terminal Inverted Repeats



Sometimes TIRs are too small to be seen in a dotplot like this

Copy consensus twice in a new alignment

Reverse complement one of them and look for homology

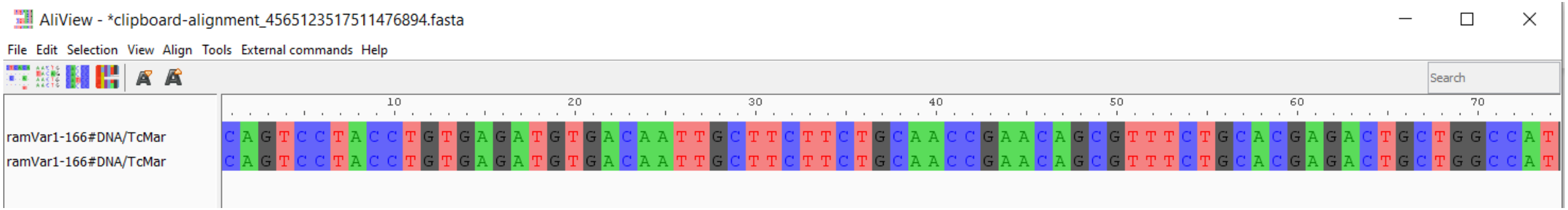
Annotate the length of the TIR in column L

1

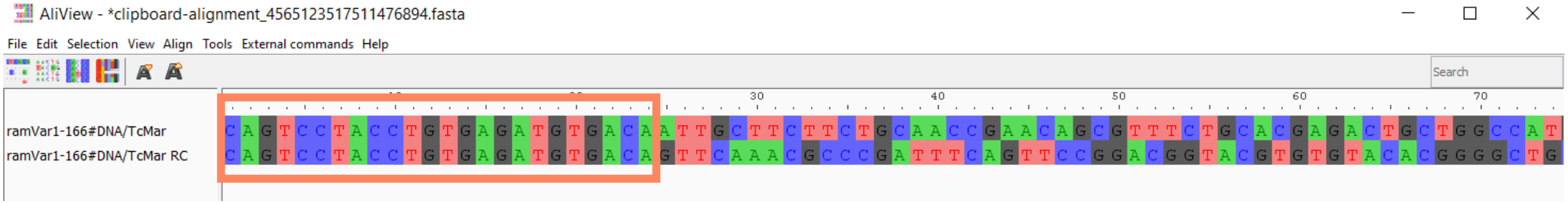
Check for TIRs

TIRs = Terminal Inverted Repeats

Copy consensus twice in a new alignment



Reverse complement one of them and look for homology

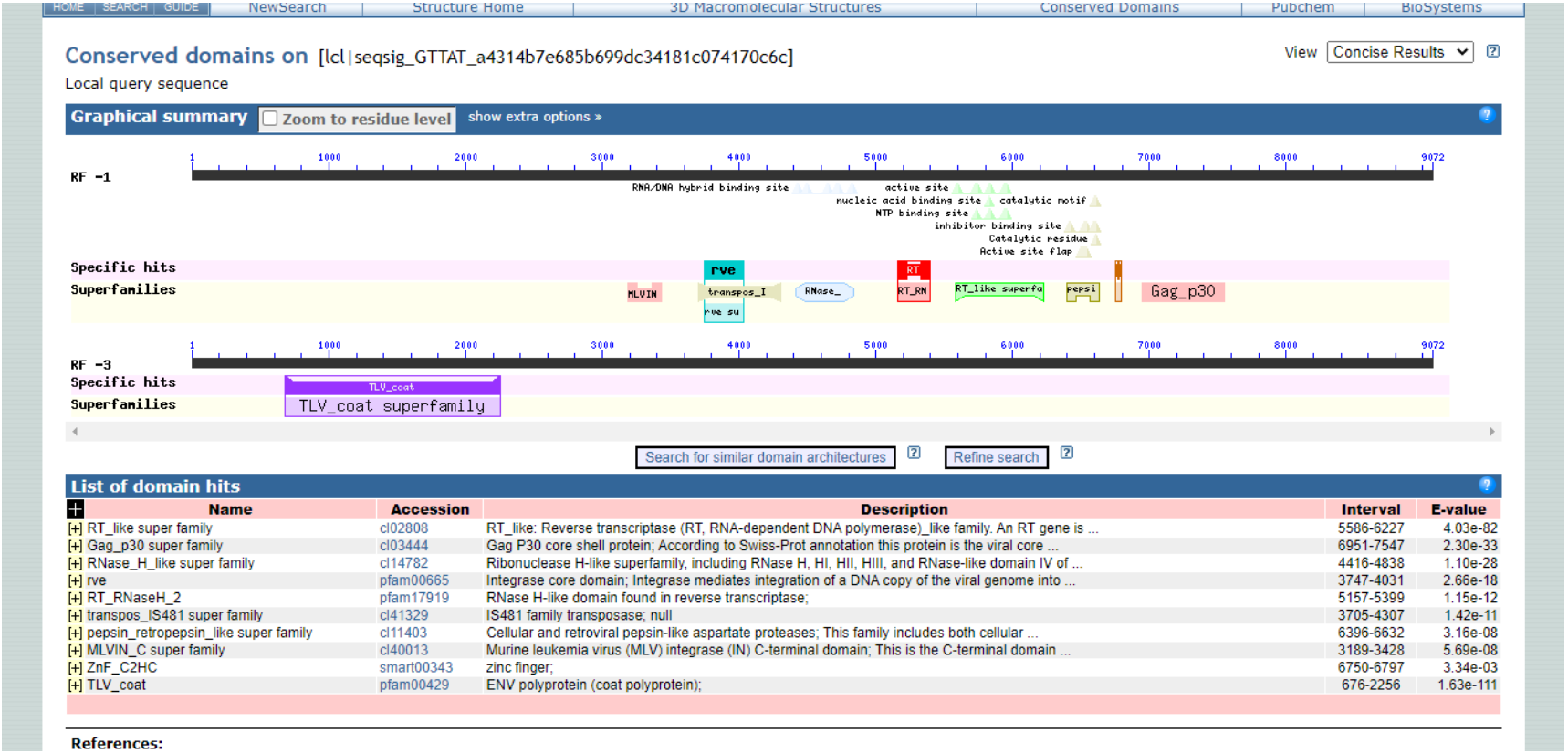


Annotate the length of the TIR in column L

Check for protein domains

Conserved Domains Database

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

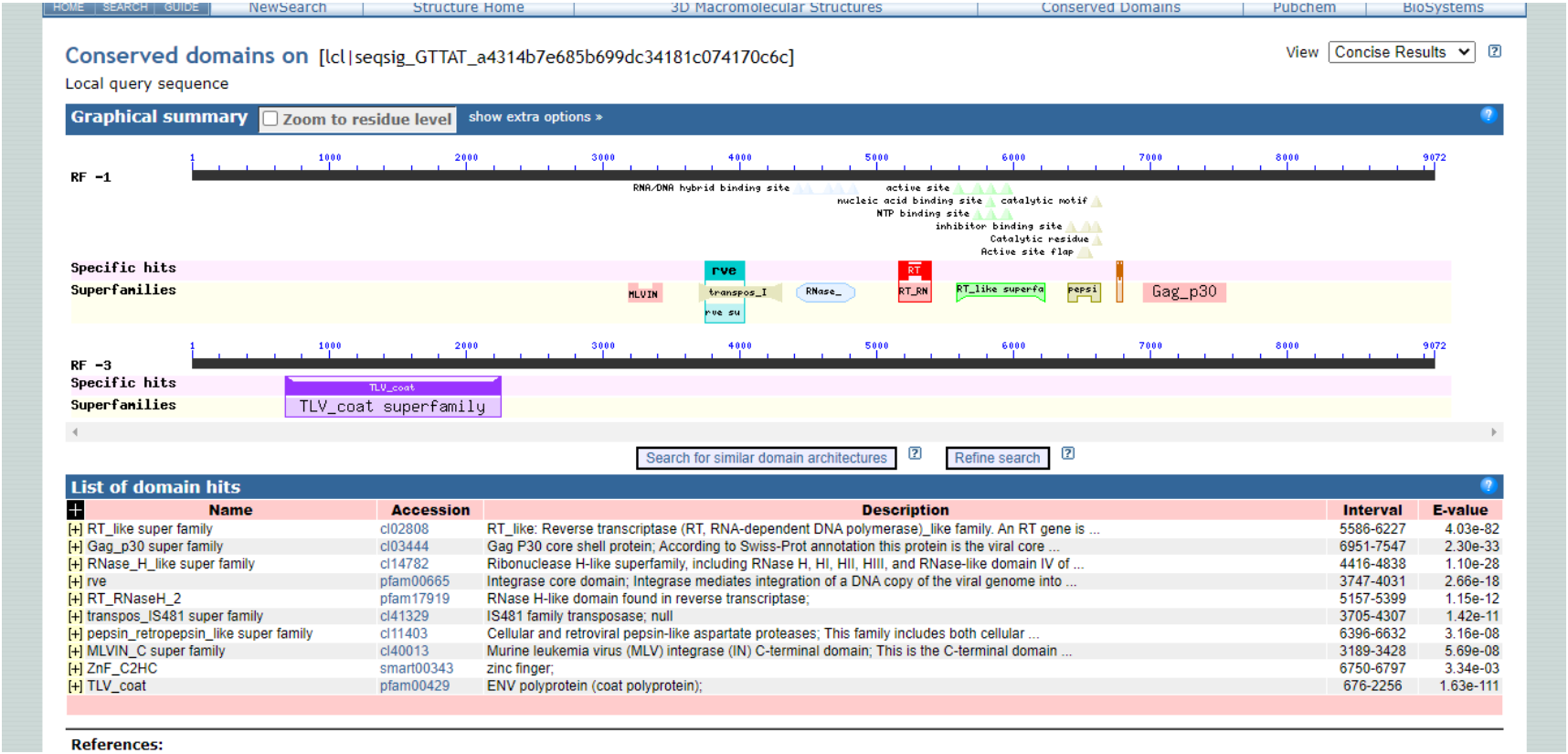


2

Check for protein domains

Conserved Domains Database

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi



HOME | SEARCH | GUIDE | NewSearch | Structure Home | 3D Macromolecular Structures | Conserved Domains | Pubchem | BioSystems

Conserved domains on [lcl|seqsig_GTTAT_a4314b7e685b699dc34181c074170c6c]

Local query sequence

View: Concise Results

Graphical summary | Zoom to residue level | show extra options »

RF -1

RNA/DNA hybrid binding site
active site
nucleic acid binding site
catalytic motif
NTP binding site
inhibitor binding site
Catalytic residue
Active site flap

Specific hits: rve, RT

Superfamilies: MLVIN, transpos_I, rve su, RNase_, RT_RN, RT_like superfa, pepsi, Gag_p30

RF -3

Specific hits: TLV_coat

Superfamilies: TLV_coat superfamily

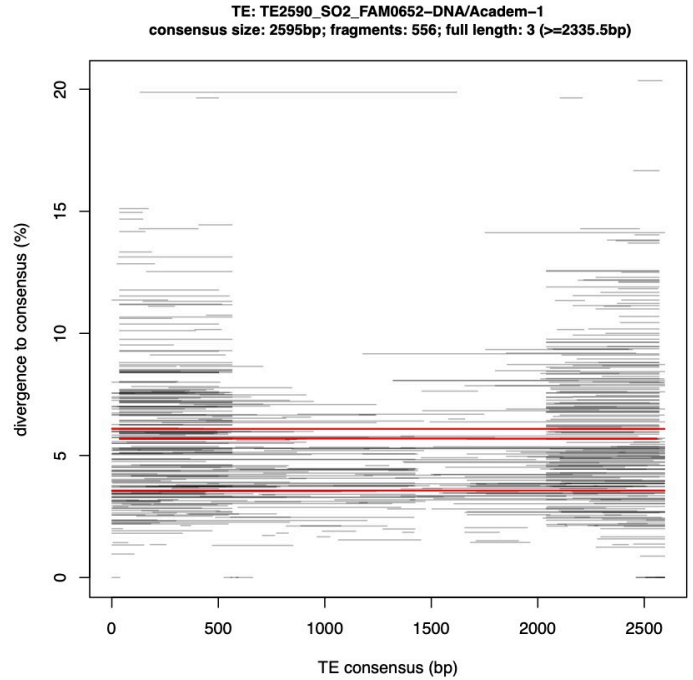
Search for similar domain architectures | Refine search

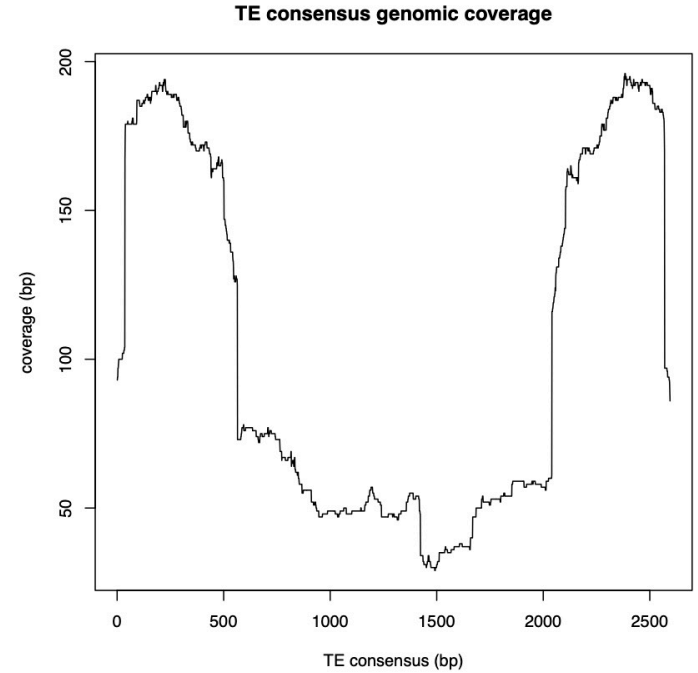
List of domain hits

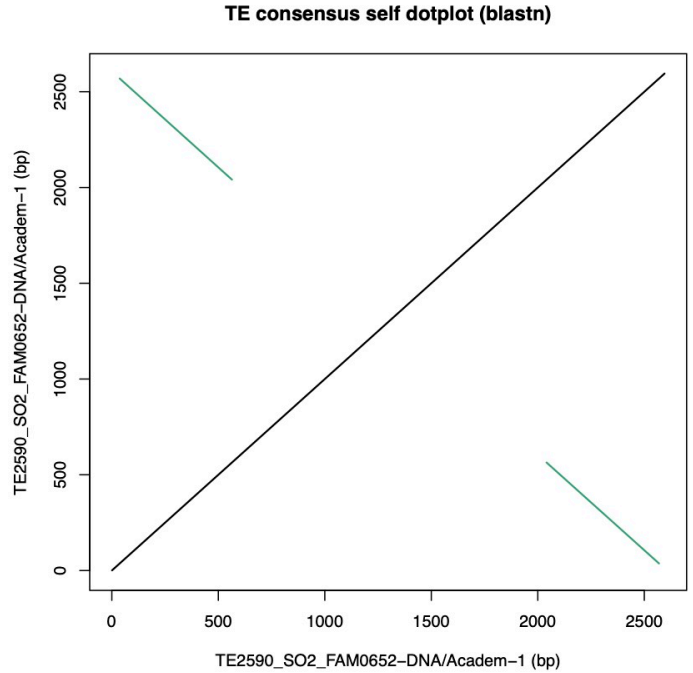
Name	Accession	Description	Interval	E-value
[+] RT_like super family	cl02808	RT_like: Reverse transcriptase (RT, RNA-dependent DNA polymerase)_like family. An RT gene is ...	5586-6227	4.03e-82
[+] Gag_p30 super family	cl03444	Gag P30 core shell protein; According to Swiss-Prot annotation this protein is the viral core ...	6951-7547	2.30e-33
[+] RNase_H_like super family	cl14782	Ribonuclease H-like superfamily, including RNase H, HI, HII, HIII, and RNase-like domain IV of ...	4416-4838	1.10e-28
[+] rve	pfam00665	Integrase core domain; Integrase mediates integration of a DNA copy of the viral genome into ...	3747-4031	2.66e-18
[+] RT_RNaseH_2	pfam17919	RNase H-like domain found in reverse transcriptase;	5157-5399	1.15e-12
[+] transpos_IS481 super family	cl41329	IS481 family transposase; null	3705-4307	1.42e-11
[+] pepsin_retropepsin_like super family	cl11403	Cellular and retroviral pepsin-like aspartate proteases; This family includes both cellular ...	6396-6632	3.16e-08
[+] MLVIN_C super family	cl40013	Murine leukemia virus (MLV) integrase (IN) C-terminal domain; This is the C-terminal domain ...	3189-3428	5.69e-08
[+] ZnF_C2HC	smart00343	zinc finger;	6750-6797	3.34e-03
[+] TLV_coat	pfam00429	ENV polyprotein (coat polyprotein);	676-2256	1.63e-111

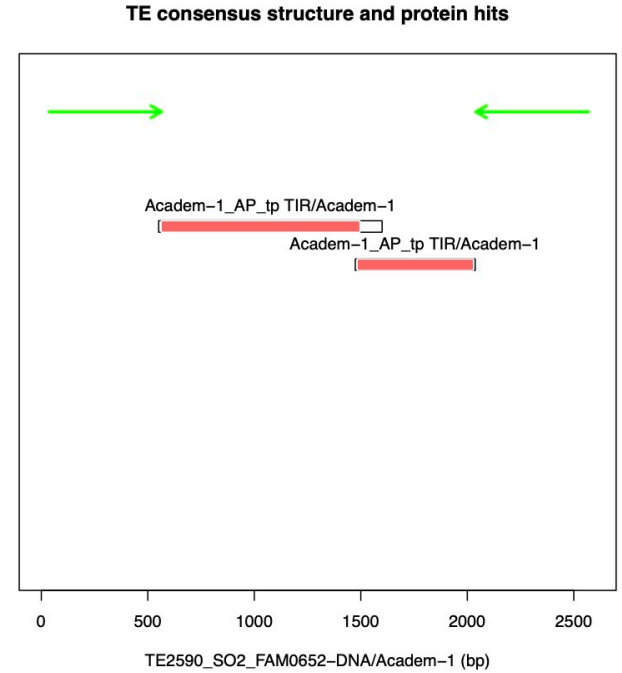
References:

The presence and order of protein domains can help for classification – annotate the presence of important domains in column G


TE: TE2590_SO2_FAM0652–DNA/Academ–1
consensus size: 2595bp; fragments: 556; full length: 3 (>=2335.5bp)
divergence to consensus (%)
0
5
10
15
20
0
500
1000
1500
2000
2500
TE consensus (bp)


TE consensus genomic coverage
coverage (bp)
50
100
150
200
0
500
1000
1500
2000
2500
TE consensus (bp)


TE consensus self dotplot (blastn)
TE2590_SO2_FAM0652–DNA/Academ–1 (bp)
0
500
1000
1500
2000
2500
TE2590_SO2_FAM0652–DNA/Academ–1 (bp)


TE consensus structure and protein hits
Academ–1_AP_tp TIR/Academ–1
Academ–1_AP_tp TIR/Academ–1
0
500
1000
1500
2000
2500
TE2590_SO2_FAM0652–DNA/Academ–1 (bp)

3

Check for nucleotide homology

Example: Censor and Repbase

https://www.girinst.org/censor/index.php

Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

chr20:12962364-12971435 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
chr20:12962364-12971435	1	276	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7451	1.7097	797
chr20:12962364-12971435	288	394	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261
chr20:12962364-12971435	532	8544	TguERV6_I	1	8078	ERV/ERV1	c	0.7176	1.7271	26725
chr20:12962364-12971435	8546	8821	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7490	1.7333	814
chr20:12962364-12971435	8833	8939	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261

3

Check for nucleotide homology

Example: Censor and Repbase

https://www.girinst.org/censor/index.php

Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

chr20:12962364-12971435 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
chr20:12962364-12971435	1	276	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7451	1.7097	797
chr20:12962364-12971435	288	394	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261
chr20:12962364-12971435	532	8544	TguERV6_I	1	8078	ERV/ERV1	c	0.7176	1.7271	26725
chr20:12962364-12971435	8546	8821	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7490	1.7333	814
chr20:12962364-12971435	8833	8939	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261

3

Check for nucleotide homology

Example: Censor and Repbase

https://www.girinst.org/censor/index.php

Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

chr20:12962364-12971435 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
chr20:12962364-12971435	1	276	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7451	1.7097	797
chr20:12962364-12971435	288	394	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261
chr20:12962364-12971435	532	8544	TguERV6_I	1	8078	ERV/ERV1	c	0.7176	1.7271	26725
chr20:12962364-12971435	8546	8821	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7490	1.7333	814
chr20:12962364-12971435	8833	8939	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261

In general, take into consideration hits that are longer than 50 bp and have a score higher than 500

3

Check for nucleotide homology

Example: Censor and Repbase

https://www.girinst.org/censor/index.php

Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

chr20:12962364-12971435 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
chr20:12962364-12971435	1	276	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7451	1.7097	797
chr20:12962364-12971435	288	394	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261
chr20:12962364-12971435	532	8544	TguERV6_I	1	8078	ERV/ERV1	c	0.7176	1.7271	26725
chr20:12962364-12971435	8546	8821	TguERV6_LTR1	407	665	ERV/ERV1	c	0.7490	1.7333	814
chr20:12962364-12971435	8833	8939	hAT6-2_SP	4951	5043	DNA/hAT	d	0.7216	1.5714	261

In general, take into consideration hits that are longer than 50 bp and have a score higher than 500

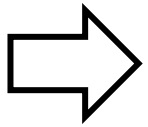
Annotate similarity to known repeats in column F

4

Re-classification

Collect the evidence

- TSD length and motifs
- TIR length
- Protein domains
- Homology to known transposable elements



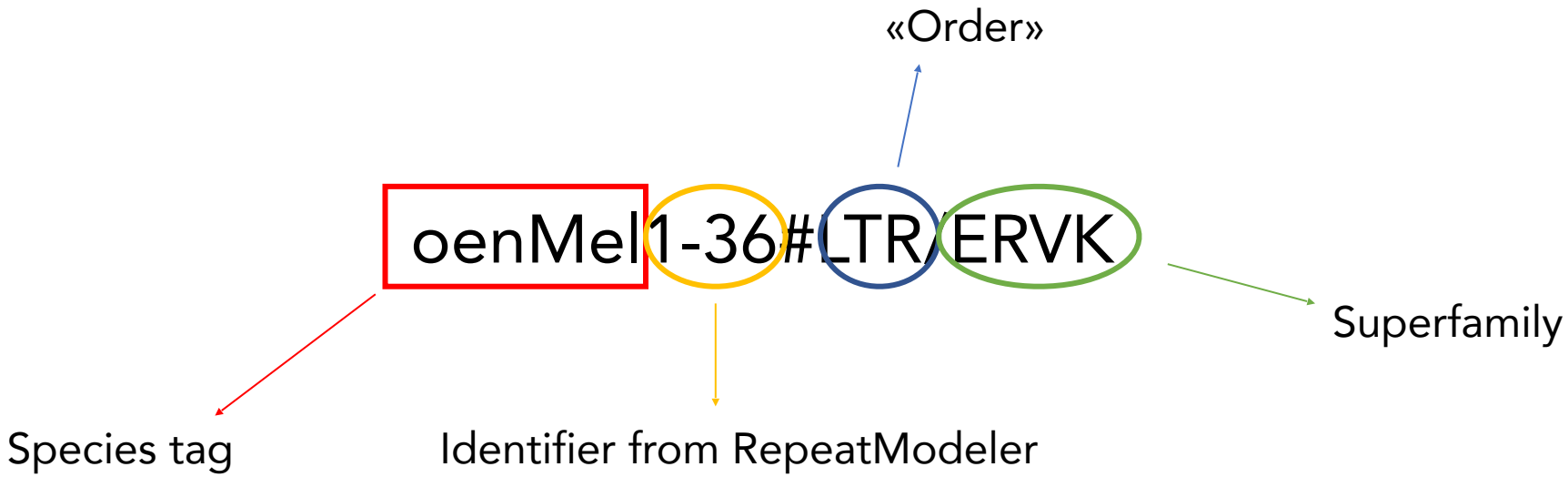
Check classification tables:

- Wicker et al 2007
- Feschotte and Pritham 2007
- Feschotte et al 2020

Change the classification tag according to the collected evidence in column S

It is possible that no better classification can be given to the consensus sequence or that a previously classified repeat becomes an ‘Unknown’

Structure of the consensus sequence name



Other examples

oenMel1-36_LTR#LTR/ERVK

oenMel1-36_int#LTR/ERVK

oenMel1-90.inc#LINE/CR1

ramVar1-626#DNA/Piggybac

Practical 2

Part 1 – Explore the diversity of TEs with TEAid plots

Part 2 – Curation and classification of alignments

- Inspect the alignments in the folder “Practical2_part2 > Alignments”
- Create a new consensus sequence with Advanced Consensus Maker (look at the slides for parameters)
- Fix SNPs and indels (majority rule)
- Find termini and TSDs

Classify the element based on TSDs and termini and use dotplots, Repbase, protein hits to help classification!
In the Practical2 folder you can find papers with classification tables.

Curation tips

- Pay attention to **CpG sites**, especially if you know that methylation is ”used” by the genome to silence TEs.
- Apply a majority rule to decide how **indels/gaps** in the alignments should be represented in the new consensus sequence. Example: in a position we have 5 gaps (“-”, 5 Cs and 5Gs. Break the tie first between insertion and gaps then apply the majority rule on the winner. In this case insertions win over gaps and you write “S” in the consensus (C/G = Strong).
- **Do not leave “?” in the new consensus sequences** because it is not recognised by blast and RepeatMasker as a valid character. If you have a tie between nucleotide frequencies, please use the ambiguity IUPAC characters (e.g., S, W, R, Y, M, K, ...). Also you can look at the nucleotide at that position in the old consensus to decide. In the worst case scenario “N” is a perfectly acceptable character.
- Double check that the new consensus sequences do not contain flanking regions and TSDs (be particularly careful with Mariners).
- **Double check that the features of the curated TEs are recorded** in the Excel file.
- **Adopt a consistent nomenclature within your library and try to be as consistent as possible with repeat libraries from closely related species** (if any) or, in general, to other manually curated repeat libraries.
- Adopt a **nomenclature for non-autonomous elements that is consistent** with other repeat libraries. Always record it in the Excel file.

Curation tips

- When you classify an element **check the naming convention for the class/superfamily on Repbase or Dfam** if possible (it could be inconsistent with RepeatMasker though)
- In case you find **LTR retrotransposons**, please split the consensus in the LTR portion ("_LTR") and internal portion ("_I"); you can check on Repbase examples of LTR naming.
- In case you find **LINE retrotransposons**, don't worry if you cannot find both ends of the element. Because of the 5' truncation phenomenon, it is very difficult to obtain complete consensus sequences of LINEs. Try to extend the consensus as much as you can (do not build a consensus of less than 3 sequences). Try to have at least one end included in the consensus sequence
- In DNA transposons, TIRs must be symmetrical but one or two mismatches are ok
- Don't be afraid to use the "Comment" column to annotate anything you think interesting, you may find unexpected patterns as you keep curating the library.