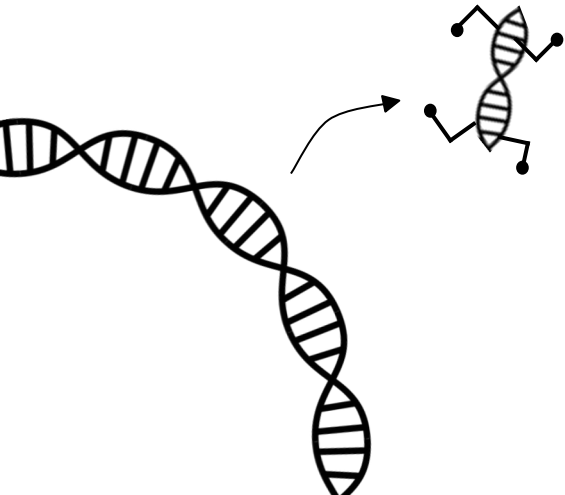
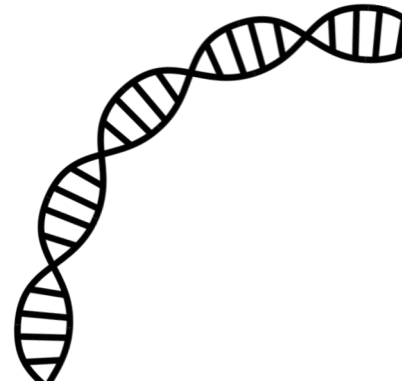


# The importance of transposable element curation



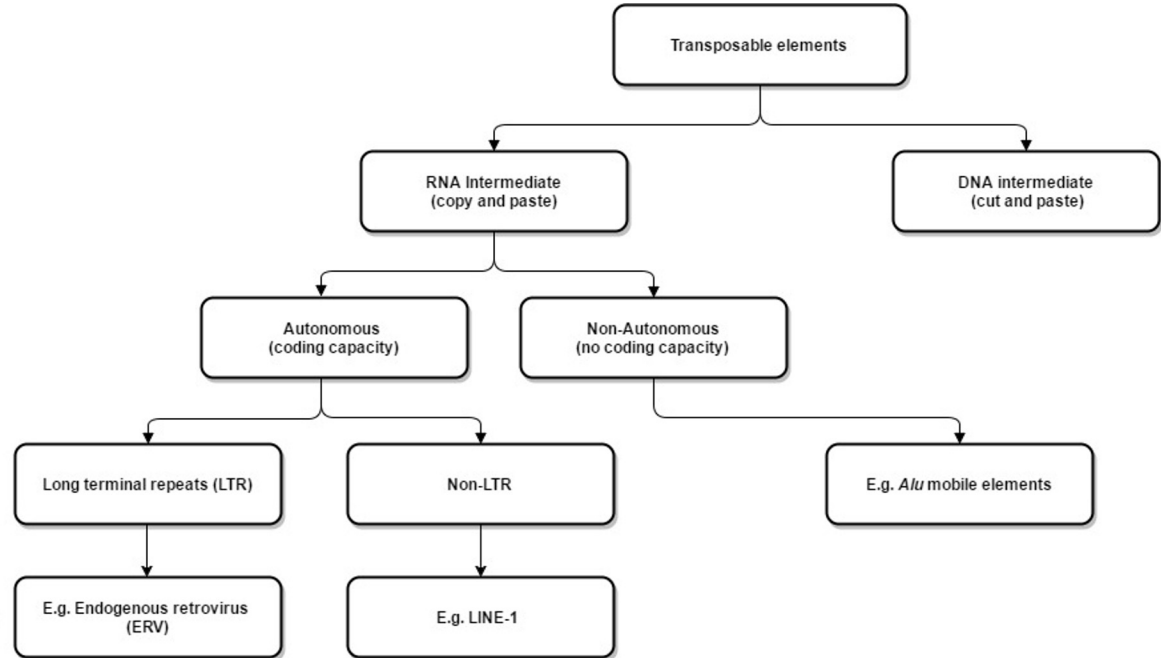
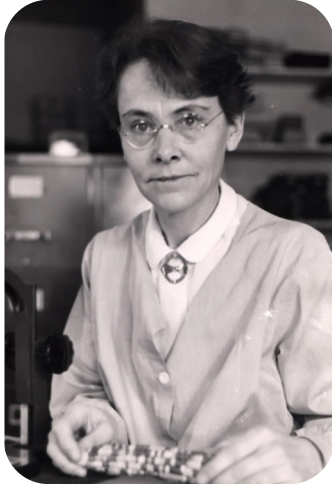
Jessica M. Storer  
Associate Research Scientist  
O'Neill Lab  
12/2/2023





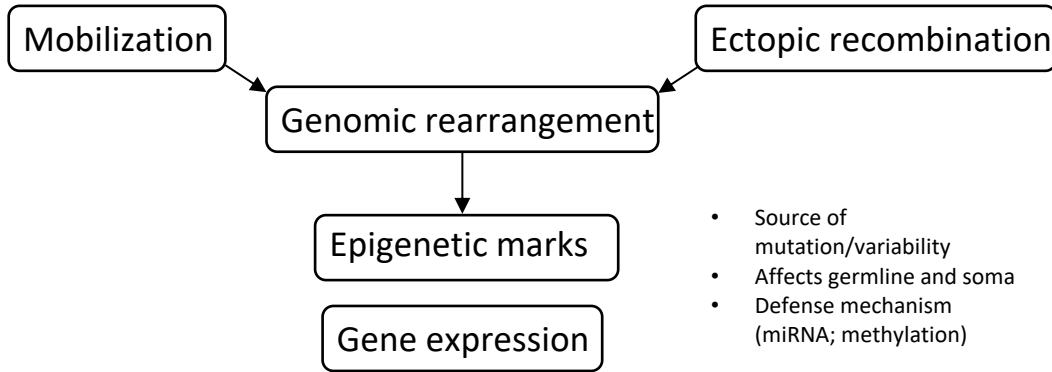
# Transposable elements (TEs)

- Discovered by Barbara McClintok
  - *Zea mays*: kernel color
  - Nobel Prize in 1983



# Transposable elements (TEs)

## Biological impact



Disease phenotypes  
(e.g. cancer, Alzheimer's, etc.)

New regulatory pathways  
(e.g. exaptation of TE  
transposase, producing  
RAG1 & RAG2 proteins)

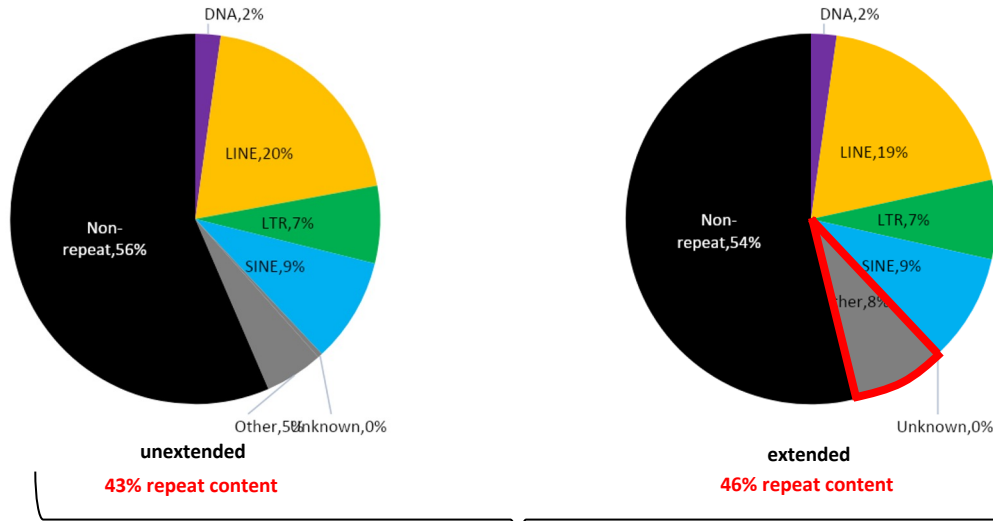
## Utility of TEs

- Genome sequencing projects
- Transgenic organisms
  - E.g., zebrafish, *Arabidopsis*
    - T-DNA seed lines
    - Generation of new mutants
- Phylogenetically informative
  - E.g. *Alu* elements
    - Homoplasmy-free
    - Mode of evolution is unidirectional, i.e., they do not revert to their ancestral state

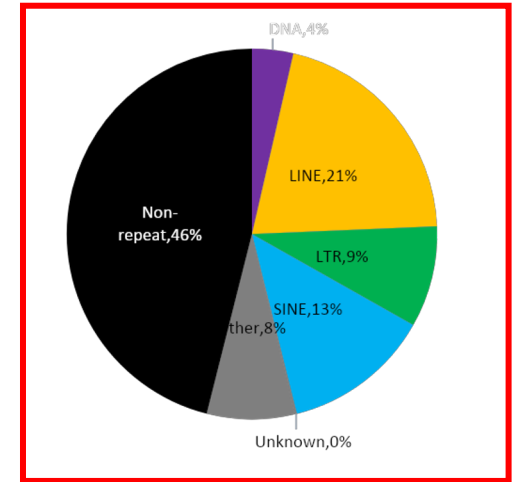
# Why do we need manual curation?

- **Fragmentation** due to large scale deletions, interruptions by nested insertions of additional TEs, and through poor insertion fidelity.
- Because of their mostly **neutral decay**, there are no conserved regions that can anchor the alignment nor are there open reading frames free from indel accumulation
- Copies are often derived from a TE rapidly evolving in a genome, so that they represent a **mixed bag of ancestral forms**.
- **Low complexity regions** and internal repetition are common features.
- The oldest detectable TE copies have accumulated over 35% substitutions since their arrival and given their neutral decay have a substitution level of more than **70% between each other**

# RepeatModeler vs. curated TE families - human



RepeatModeler2

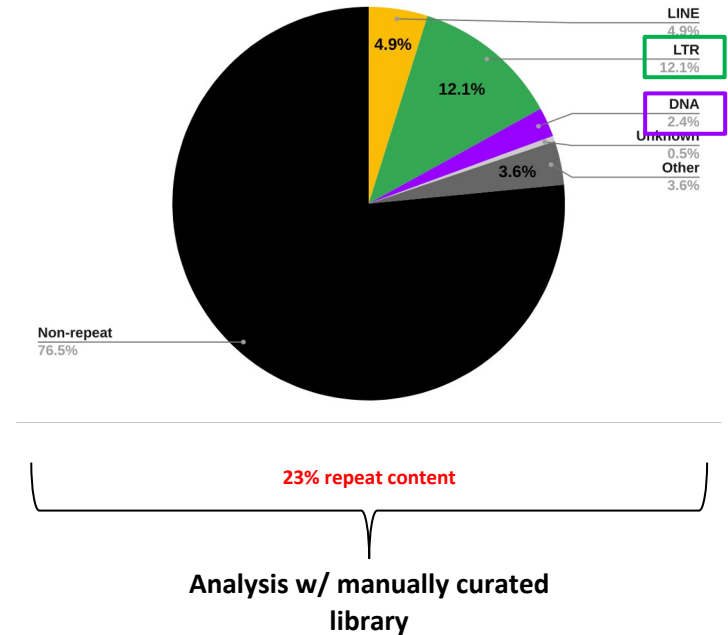
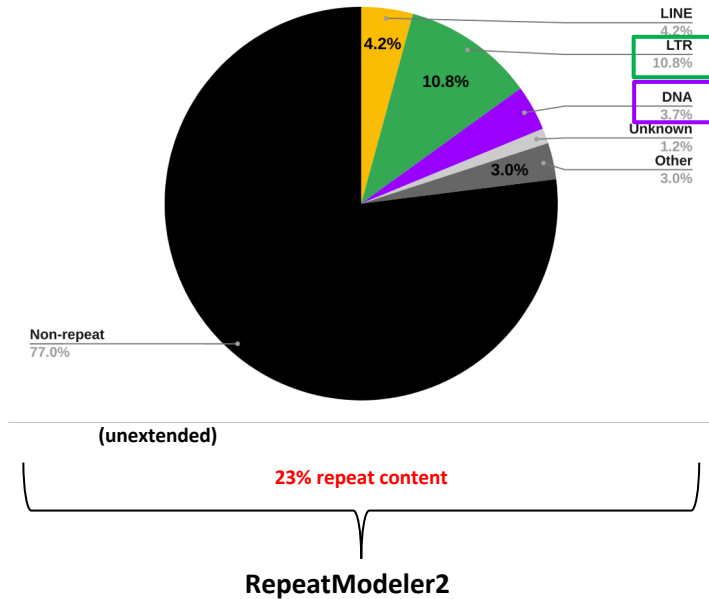


Analysis w/ manually curated library

Enhanced ability to classify and identify repeats with carefully-curated dataset

\*Other: satellites and simple repeats

# RepeatModeler vs. curated TE families - fruit fly



Enhanced ability to classify and identify repeats with carefully-curated dataset

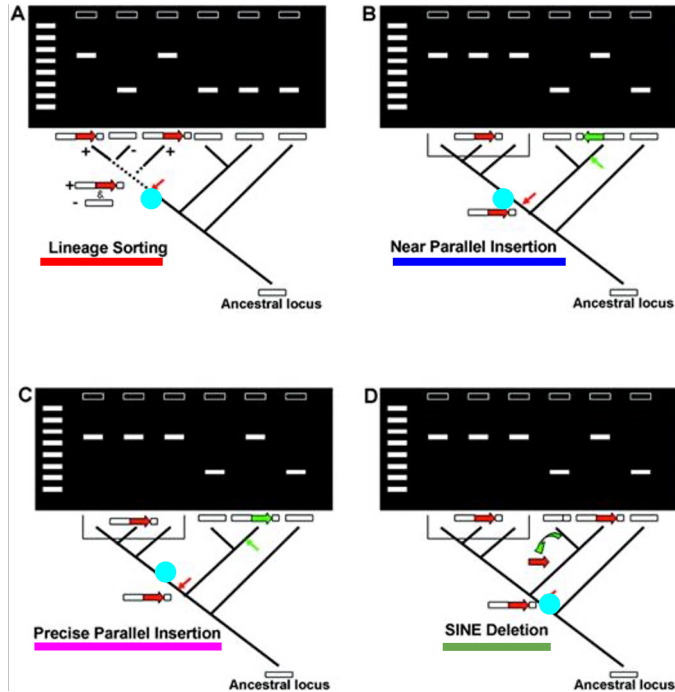
# Alu mobile elements



- SINEs (short interspersed element)
- Non-autonomous
- ~300 bp
- ~1 million copies in the human genome
- Transcribed by RNA polymerase III
- Derived from 7SL RNA
- Homoplasy-free
  - No known mechanism for the specific removal of SINE elements from the genome
  - Mode of evolution is unidirectional, i.e. they do not revert to their ancestral state
- Known ancestral state = absence of *Alu* element
- Easy to genotype
- Elements facilitate a comprehensive analysis of phylogeny



# Homoplasy-free (nearly)



Ray et al. 2006 "SINEs of a Nearly Perfect Character"  
*Systematic Biology* 55(6):928-935

A: Incomplete lineage sorting:  $\sim 0.0006$  ★  
events/insertion

Knowledge of primate behavior

B: Near parallel insertions:  $\sim 0.0004$   
events/insertion

Sequencing

C: Precise parallel insertions:  $\sim 0.005$   
events/insertion

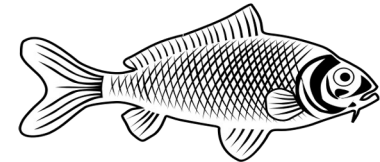
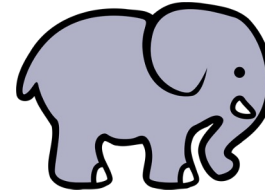
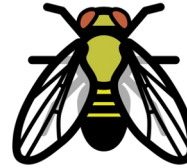
Sequencing - subfamily analysis

D: SINE deletion: No known mechanism

● Insertion event

# Filtering your data

- High copy gene families
- Processed pseudogenes
  - Generally from highly transcribed genes
- Simple repeats/low complexity
- Redundancy
  - There is not any single % divergence cut off or general strategy that will work for all TE types across all organisms
    - Unique repertoire of elements in each species
    - Genomic gain and loss
    - Adaptations
      - Flight
    - Generation time
    - Diet
    - .....





# A beginner's guide to manual curation of transposable elements

Clement Goubert<sup>1,2</sup>, Rory J. Craig<sup>3</sup>, Agustin F. Bilat<sup>4</sup>, Valentina Peona<sup>5</sup>, Aaron A. Vogan<sup>5</sup> and Anna V. Protasio<sup>6,7\*</sup>

Beginner

Advanced

## Curation Guidelines for *de novo* Generated Transposable Element Families

Robert Hubley,<sup>1</sup> Jeb Rosen,<sup>1</sup> and Arian F. A. Smit<sup>1</sup>

<sup>1</sup>Institute for Systems Biology, Seattle, Washington

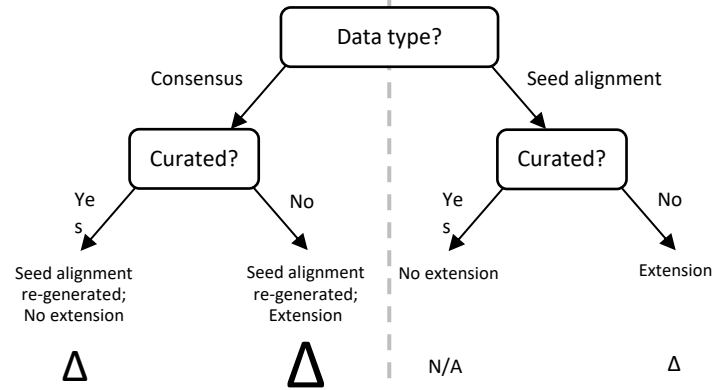
<sup>2</sup>Corresponding author: [jessica.storer@isbscience.org](mailto:jessica.storer@isbscience.org)

### Beginner vs. Advanced:

- Starting material
  - Consensus vs. stockholm/alignment
- Collecting copies/insertions
  - BLAST vs. RepeatMasker
- All models vs. individual models
- Alignment
  - MAFFT vs. Refiner
- Consensus generation
  - EMBOSS vs. Refiner

# Input data type

- Less maintenance of data; no provenance
  - Cannot troubleshoot in downstream processes
- Greater chance of original consensus sequence changes
- Subfamily splitting may no longer be maintained



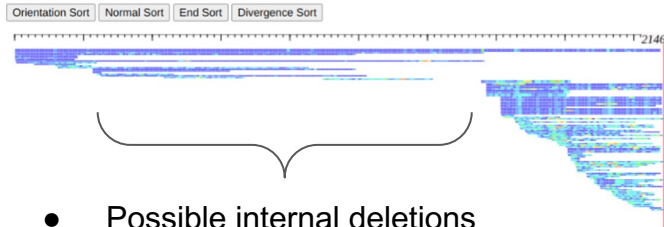
- Provenance of consensus sequence derivation is maintained
  - More data maintained
- Less chance of major changes to consensus sequence made at the end of the curation process

# VISUALIZE, VISUALIZE, VISUALIZE!

**ALWAYS** check your data BEFORE doing ANY ADDITIONAL STEPS

```
$ viewMSA.pl -stockholm <file.stk>
```

#2

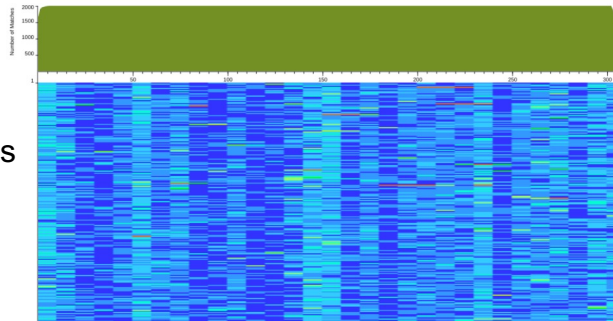


- Possible internal deletions
- Low(er) divergence at terminal ends

Edge to edge matches

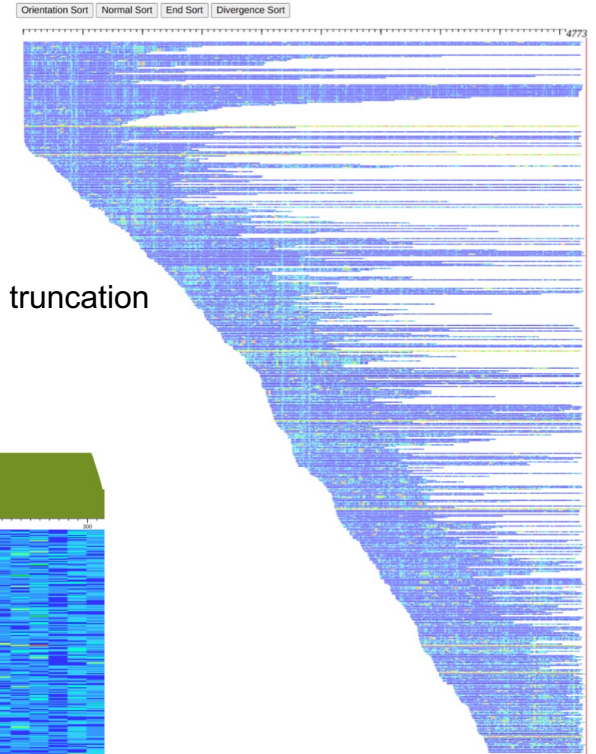
#3

SINE (*Alu*) (could also be soloLTR)



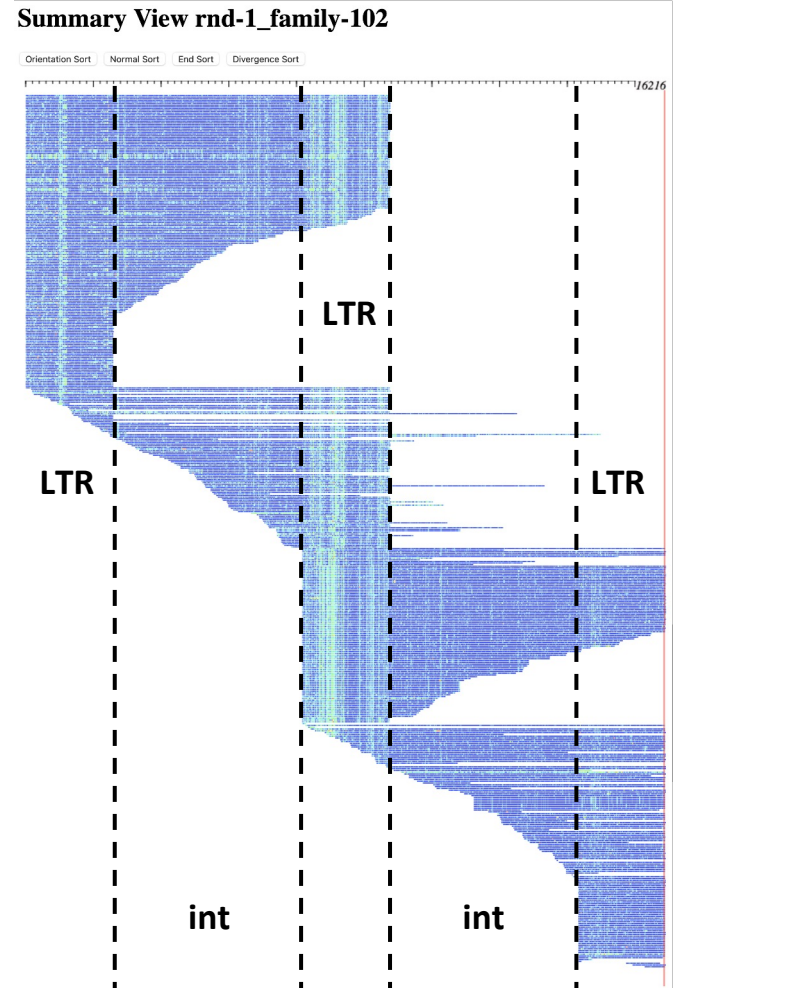
#1

5' truncation



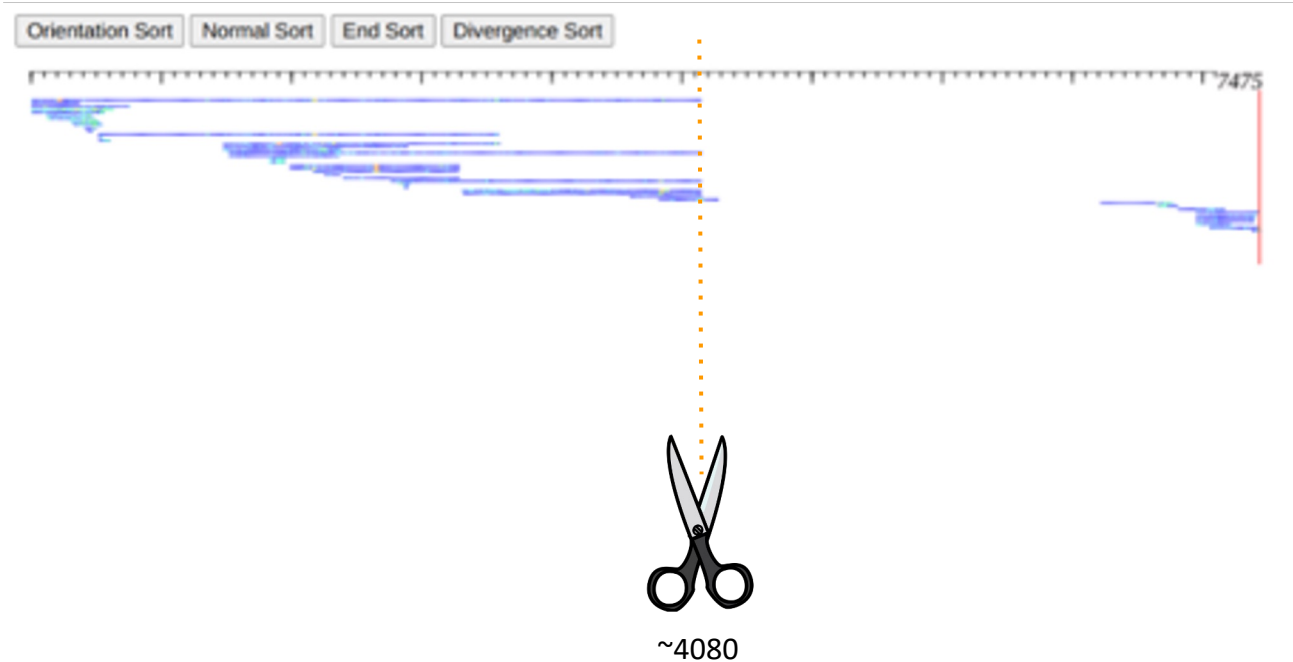
# LTR/ERV over-extension

- 1939 taro models submitted; 1173 (60.5%) LTR/ERV elements
- Types of LTR alignments observed



\*Example from taro validated by RM, crossmatch and dotplot

# On to the manual part and a lot more detail



# Matrix

- Corresponds to the ratio of the nucleotide's observed frequency given an ancestral (consensus) base over the nucleotide's frequency in the background
  - 20 kb flanking the transposon was used for background frequency
- Substitution frequency is dependent upon:
  - Age of the repeat
  - GC content of a given locus

nucleotide in the query sequence (derived state)

	A	G	C	T	A	G	C	T	A	G	C	T
<b>14p</b> 49g.matrix	10	-10	-16	-19	9	-7	-13	-15	9	-5	-12	-13
<b>20p</b> 49g.matrix	-6	10	-18	-18	-3	9	-15	-14	-2	8	-13	-12
<b>25p</b> 49g.matrix	-18	-18	10	-6	-14	-15	9	-3	-12	-13	8	-2
	-19	-16	-10	10	-15	-13	-7	9	-13	-12	-5	9
20p <b>35g</b> .matrix	8	-7	-15	-17	9	-8	-15	-17	9	-6	-13	-14
20p <b>43g</b> .matrix	-4	11	-13	-14	-4	10	-15	-15	-3	9	-14	-14
20p <b>51g</b> .matrix	-14	-13	11	-4	-15	-15	10	-4	-14	-14	9	-3
	-17	-15	-7	8	-17	-15	-8	9	-14	-13	-6	9

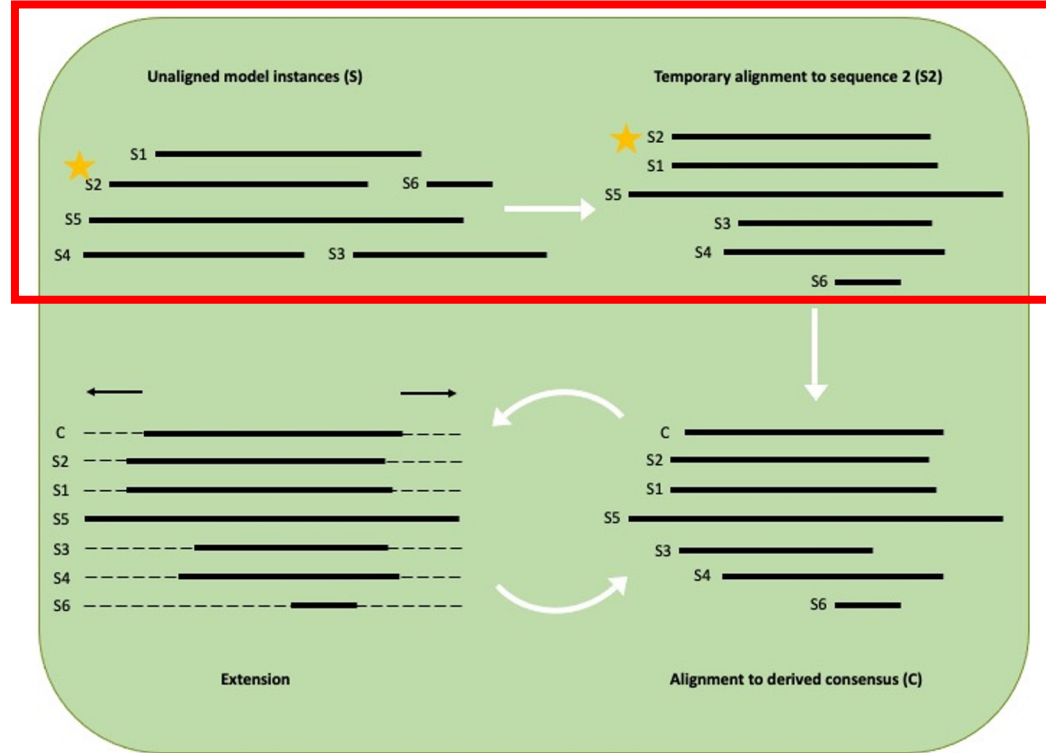
Differing ages; same GC content

Same age; differing GC content



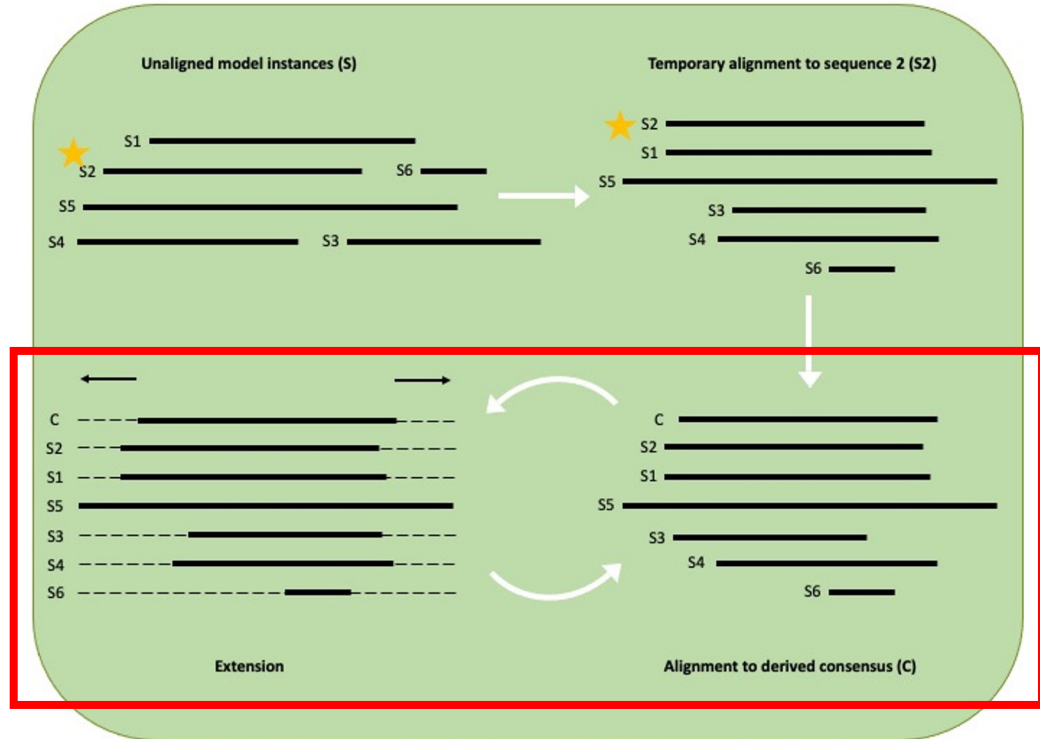
# Alignment

- Sensitive alignment matrices and gap parameters for neutrally—evolving DNA
  - Developed for TE annotation and used in RepeatMasker for years
  - Derived from ancient DNA transposon data delimited by divergence and CG background
- Multiple sequence alignment method
  - Iterative transitive search, bootstrapped with the best matching sequence



# Refinement

- Iterative process of extension and re-alignment to the consensus sequence until the consensus sequence stabilizes



# Refinement

- Interactive
  - Easily visualize the process
    - Matrix used
    - Search engine used
    - Divergence of your alignment
    - Areas of possible extension

```
A
## alignAndCallConsensus (aka dothensimple/dothemultiple.pl)
## Version 2.0.2-beta-2
##
## Single Family Mode
# Engine: rmblast | Matrix: 25p1g-Hpad.matrix | Bandwidth: 40,
# Minmatch: 7, Minscore: 200,
# Maxdiv: 60, GapInit: -25,
# InsGapExt: -5, DelGapExt: -4
-----
ITERATION: 1
Working on example1_con
Unique aligned sequences: 126
Total Crossmatch Score: 259954
Per Base Average: 4.51
Kinura Divergence: 0.117704123552758 60886 aligned bps ]

B
## alignAndCallConsensus (aka dothensimple/dothemultiple.pl)
## Version 2.0.2-beta-2
##
## Single Family Mode
# Engine: rmblast | Matrix: 14p41g-Hpad.matrix | Bandwidth: 40,
# Minmatch: 7, Minscore: 200,
# Maxdiv: 60, GapInit: -25,
# InsGapExt: -7, DelGapExt: -6
# Extension Mode, 1 sequences have Hpads
# Starting Round Index: 2
-----
ITERATION: 1
Working on example1_con
Unique aligned sequences: 126
Total Crossmatch Score: 274004
Per Base Average: 4.68
Kinura Divergence: 0.127627183695182 57462 aligned bps )
Changes:
consensus      1  AAGSNNNCDAAC-TGTCTGTGTGGAGACCTTGTG-GGC-----CG-CCDC-----C-CA-----G--CC-AC-GTGGAAAGNCCCT 65
               1  ???????
ref:example1_con 1  |HHHHHHHCDAAAC-T-GTCTGTGTGGAGACCTTGTG-GGC-----CG-CCDC-----C-CA-----G--CC-AC-GTGGAAAGNCCCT 65
consensus      66  AACTTCCCGATGA-GGA-AAGCCCTCTCTCCCC-C-GCAGGGAGGGTCCCTT-A-TCTCACTCT-----GT-----C-----C-T-----G 129
ref:example1_con 66  AACTTCCCGATGA-GGA-AAGCCCTCTCTCCCC-C-GCAGGGAGGGTCCCTT-A-TCTCACTCT-----GT-----C-----C-N-----G 129
consensus      130  GCCCGG-CCCBANNCBACCACTTC-CCGCC-CGGCAA-----CC-CC-CTC-----GCAGGAGGGTCCCT-T 189
ref:example1_con 130  GCCCGG-CCCBAGGCCBACCACTTC-CCGCC-CGGCAA-----CC-CC-CCN-----GCNA-AGGGCCCTT-T 187
consensus      190  -CT-CACTCCGGCC-CC-----GTT-CCC-C-GG-CCA-C-GT-G-AAA-GGGCC-TA-AC-----TTCCGATAGAAGGA-AGCC 258
ref:example1_con 188  -CT-CACTCCGGCC-CC-----GTT-CCC-C-GG-CCA-C-GG-CCA-C-GG-CCA-GGNC-TA-AC-----CTTCCGATAGAAGGA-AGCC 245
consensus      340  CCG-A-----ACCA-ATC-A-----CC-CC-G-CCC-----AT-C-AGC-T-C-T-C-CCAGTAA- 378
ref:example1_con 335  CCG-A-----ACCA-ATC-A-----CC-CC-G-CCC-----AT-C-AGC-T-C-T-C-CCAGTAA- 373
consensus      420  C-----CC-----GG-C-AA-C-C-AACC-TGGCCATC-C-CCACCCCGAGATC-A-GC-TCTCC-C-AC-C-----C 470
ref:example1_con 415  C-----CC-----GG-C-AA-C-C-AACC-TGGCCATC-C-CCACCCCGAGATC-A-GC-TCTCC-C-CC-C-----C 465
consensus      471  CC-----CCAG-----TCC-CT-C-T-----G-CCC-----C-----TA-----TA-AA 494
ref:example1_con 466  CC-----CCAG-----NCC-CT-C-T-----G-CCC-----C-----TA-----TA-AA 489
consensus      495  AACCGA-CDBAAC-AA-AGAAA-GC-CGGCCBAGACTGCTACTTCCD-CCGGC-AGGATCCAGTCCGGCCGG-AGACTCTCCAAT-AAA 579
ref:example1_con 490  AACCGA-CDBAAC-AA-AGAAA-GC-CGGCCBAGACTGCTACTTCCD-CCGGC-AGGATCCAGTCCGGCCGG-AGACTCTCCAAT-AAA 574
consensus      580  -GCC-TGNA-C-TGG-T-CACCACGCT-CT-CCGCT-G-GTGTAACTCGGTG-CG-GC-CTGGGG-TCCAACACTAGAGGGTCCGGGC- 685
ref:example1_con 575  -GCC-TGNA-C-TGG-T-CACCACGCT-CT-CCGCT-G-GTGTAACTCGGTG-CG-GC-CTGGGG-TCCAACACTAGAGGGTCCGGGC- 658
consensus      656  AGGTACAGACACCGGGTCCACCA 679
ref:example1_con 651  AGGTACAGACACCGGGTCCACCA 674

s(kip),c(changeinbetweenfs),x(pandandchange),b(beginexpand) or 5('),e(indexpend) or 3('),##-## (range),d(one)
A range only works if the new and old consensus have the same positions at the start and end of the range.
|
```

## • Interactive mode

- Choose the option that best fits the data

- Accept all changes
- Pad the sequence and accept all changes
- Expand the sequence in the 5' direction
- Expand the sequence in the 3' direction

```
s(kip),c(hangeinbetweenHs),x(pandandchange),b(eginexpand) or 5('),e(ndexpand) or 3('),##-## [range],d(One)
5
-----
Keeping only 5' H-pad changes.
ITERATION: 6
Working on example1_con
Unique aligned sequences: 131
Total Crossmatch Score: 304572
Per Base Average: 4.98
Kimura Divergence: 0.118839360124563 ( 61182 aligned bps )
Changes:
consensus          1  ACCCAGTAGGAAAGC---CGCCGCC---TTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC-----CC-C----- 74
   ???????       1?       ?
ref:example1_con   1  HHHHHHAGGAAAGC---CGCCNCCC---TTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC-----CC-C----- 74
consensus          689  GGGTCCGGCC---AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
ref:example1_con   689  GGGTCCGGCC---AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
```

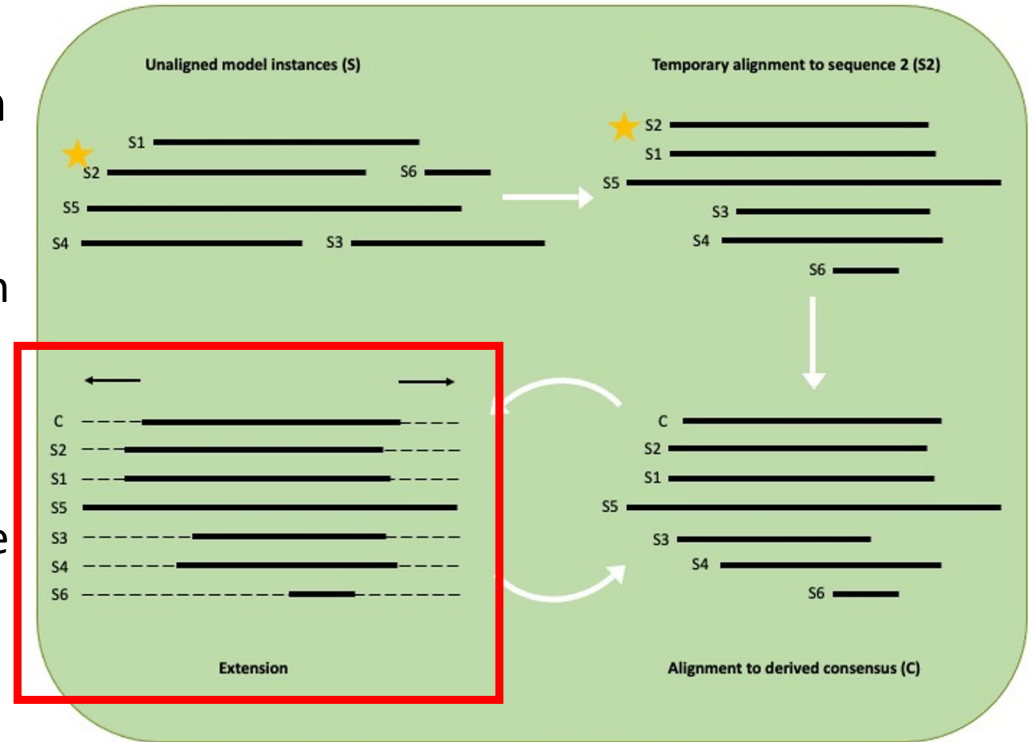
```
s(kip),c(hangeinbetweenHs),x(pandandchange),b(eginexpand) or 5('),e(ndexpand) or 3('),##-## [range],d(One)
5
-----
Keeping only 5' H-pad changes.
ITERATION: 7
Working on example1_con
Unique aligned sequences: 131
Total Crossmatch Score: 305440
Per Base Average: 4.91
Kimura Divergence: 0.118743378&19929 ( 61282 aligned bps )
Changes:
consensus          1  AGGAAGTAGCCAGI--AAGAAAGCCGCC-CCC-TTTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC-----CC-C---- 81
   ???????       1
ref:example1_con   1  HHHHHHAGCCAGA--AAGAAAGCCGCC-CCC-TTTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC-----CC-C---- 81
consensus          708  CCGGCC-AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
ref:example1_con   708  CCGGCC-AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
```

```
s(kip),c(hangeinbetweenHs),x(pandandchange),b(eginexpand) or 5('),e(ndexpand) or 3('),##-## [range],d(One)
5
-----
Keeping only 5' H-pad changes.
ITERATION: 8
Working on example1_con
Unique aligned sequences: 131
Total Crossmatch Score: 306066
Per Base Average: 4.92
Kimura Divergence: 0.118531931190716 ( 61354 aligned bps )
Changes:
consensus          1  AGTCAGCAGGAA--GTAGCCAGA--AAGAAAGCCGCC-CCC-TTTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC----- 85
   ???????
ref:example1_con   1  HHHHHHAGGAA--GTAGCCAGA--AAGAAAGCCGCC-CCC-TTTTTCTCTTTAAGGAGTT-G---GGANT-GTCTGGTGGAGG-ACCTTTGGCC----- 85
consensus          699  ACGAGGTCGGCC---AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
ref:example1_con   699  ACGAGGTCGGCC---AGGTCAGACAACCGGGTCCACANNNNCHN
   ???????
```

```
s(kip),c(hangeinbetweenHs),x(pandandchange),b(eginexpand) or 5('),e(ndexpand) or 3('),##-## [range],d(One)
d
Done! Consensus file (example1_con.fa) has been updated with any previously made selections.
```

# Extension of truncated models

- Consensi derived from *de novo* repeat finders are often truncated
  - Need to extend into the flanking sequence in order to get an accurate and full-length model!
- H-pad
  - Positively-scoring
  - Part of IUPAC code, but not in consensi or genomic sequence
- Support protocol
  - Get flanking sequence



# Get flanking sequence

```
$ extendFlankingSeqs.pl -d(atabase) <2bit genome file> -i(nput)
<cross_match file> -o(utput) <fasta file>
```

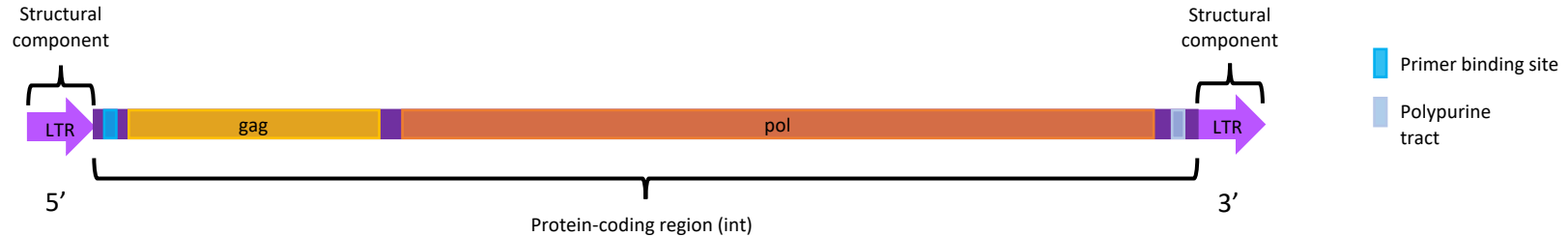
*(command is seen in supplemental protocol of publication!)*

NOTE: the input file is the .out file in most cases (this may trigger RepeatMasker flashbacks....)

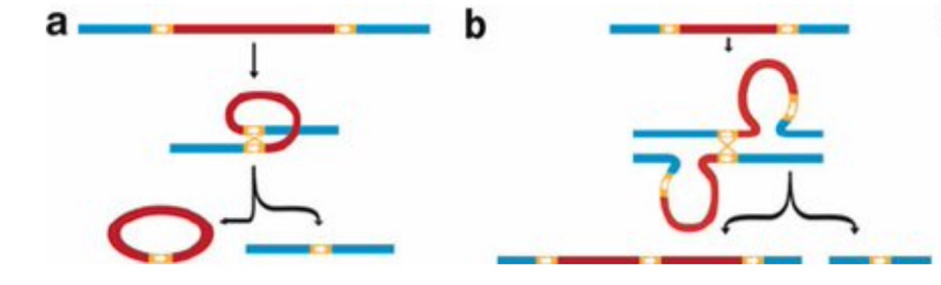
## **Alternatively....**

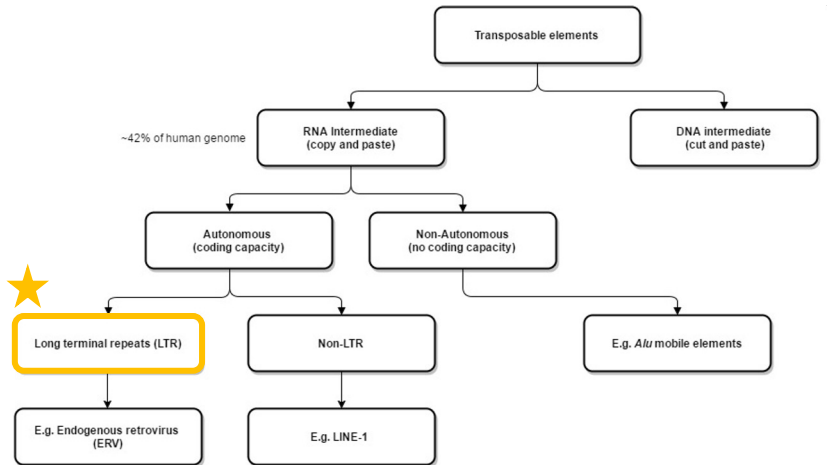
```
$ bedtools slop -i <input bed file> [options] > output.bed
$ bedtools getfasta -fi <genome> -bed <output from bedtools slop> -fo
<output fasta file>
```

# LTR/ERV sequence structure



- LTR = long terminal repeat
  - 200-1000 bp
- Prone to internal deletions in LTR region
- Recombination to form new subfamilies
- Ectopic recombination common
  - Many soloLTRs
  - LTR-int-LTR-int-LTR structures





-42% of human genome



- LTR structural features
  - 5' TG
  - 3' CA
  - Possible subfamily structure?



```

consensus
ref:example1_con
KZ285162.1:1790864-781926_R
KZ281810.1:19831485-9832517
KZ197485.1:4286980-4288526_R
KZ282280.1:26936096-26938313_R
KZ285256.1:1496984-4968734_R
KZ281897.1:5148760-5142888_R
KZ286965.1:6536487-6538595
KZ285945.1:1757228-782888
KZ196241.1:110857712-18864329
KZ281190.1:8694675-8782652_R
KZ282677.1:5819911-5828414_R
KZ288347.1:3215669-3223889
KZ282344.1:122278662-22291853
KZ288605.1:12542173-2542973_R
KZ285789.1:1929785-1930424_R
KZ289381.1:1254649-2625782
KZ284391.1:6828668-6822183_R
KZ288747.1:537838-537554
KZ197479.1:8092430-8092619_R
KZ198231.1:3316230-3316793_R
KZ282947.1:4968328-4968833_R
KZ198877.1:781918-783249_R
KZ284105.1:1138088-1382533
KZ281190.1:8694675-8782652_R
KZ287949.1:2729268-2729796_R
KZ282677.1:5819911-5828414_R
KZ288661.1:6627992-6628526_R
KZ197812.1:12034561-12035893_R
KZ281897.1:5146517-5147718_R
KZ283842.1:2738721-2736887
KZ287968.1:178376-78874
KZ198769.1:3783342-3783859
KZ284984.1:4216817-4216595
KZ287498.1:437896-438158_R
KZ198867.1:11837137-11838619
KZ283811.1:2625469-2625782
KZ197452.1:1912911-1913443_R
KZ281674.1:11885354-1889553
KZ282428.1:21438-27572
KZ283842.1:2841627-2842208_R

```

```

1 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAHT-GTCTGGTGGAGG-ACCTTTGG 98
1 HHHHHHAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAHT-GTCTGGTGGAGG-ACCTTTGG 98
383 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTT-CCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-GCTTTTGG 475
566 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTT-CCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-GCTTTTGG 654
1002 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTT-CCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-GCTTTTGG 1890
1194 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCT-TTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-TACCTTTGG 1286
1205 CGCCCAAATTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-GTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-ACCTTTGG 1294
1478 CGCCCAAATTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-ACCTTTGG 1566
1639 CGCCCAAATTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-ACCTTTGG 1729
5893 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-ACCTTTGG 5182
6832 CGCCCAAAGTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTAGAGA-ACCTTTGG 6120
7472 CGCCCAAATTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTCTTCT-TTAAGGAGTT-G--GGAJT-GTCTGGCAGAGG-ACCTTCGG 7851
7812 CGCCCAAATTCAG-CAGGAA-GTAGCCAGA-AAGAAAGCCGCCG-CCC-TTTTTCCTCTTTAAGGAGTT-GGGAAT-T-GTCTG-TGGAGG-ACCTTTGG 7983
7519 TCAG-CAGGAA-GTAGCCAGA-AAAGAG-CGATG-TCC-TTCTCTG-----AAAAAGT-----GGAJT-GT 7576
11694 -----CCD-TTTTTCCTCTTTAAGGAGTT-G--GGAJT-GTCTGGTGGAGG-ACCTTTGG 11742
16 TTAAGG-TG-A--GGAJT-GTCTGGTGGAGG-ACCTTTGG 49
231 AGTACT-TG-A--GGAJT-GTCTGGCAGAGG-ACCTTTGG 33
AGGAGT-TG-A--GGAJT-GTCTGGTGGAGG-ACCTTTGG 288
2 GAGCT-TG-A--GGAJT-GTCTGGTGGAGG-ACCTTTGG 33
6 TT-A--GTA-T-GTCTGGTGGAGG-ACCTTTGG 33
9 GAAJ-TGCTGGTGGAGG-ACCTTTGG 32
11 GACT-GTCTGGTGGAGG-ACCTTTGG 34
11 GAAJ-TGCTGGTGGAGG-ACCTTTGG 34
4 ACT-TGCTGGTGGAGG-ACCTTTGG 26
9 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 AJ-TGCTGGCAGAGG-ACCTTTGG 31
9 ACT-TGCTGGTGGAGG-ACCTTTGG 32
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
12 ACT-TGCTGGTGGAGG-ACCTTTGG 34
14 ACT-TGCTGGTGGAGG-ACCTTTGG 35
15 AT-TGCTGGTGGAGG-ACCTTTGG 37
18 T-GTCTGGTGGAGG-ACCTTTGG 38
10 T-GTCTGGTGGAGG-ACCTTTGG 38
11 T-GTCTGGTGGAGG-ACCTTTGG 31
11 T-GTCTGGTGGAGG-ACCTTTGG 33
13 T-GTCTGGTGGAGG-ACCTTTGG 34
15 T-GTCTGGTGGAGG-ACCTTTGG 35
2713 T-GTCTGGTGGAGG-ACCTTTGG 3733
22 TGGTGGAGG-ACCTTTGG 48

```

```

consensus
ref:example1_con
KZ198877.1:781918-783249_R
KZ282344.1:122278662-22291853
KZ281674.1:11885354-1889553
KZ284984.1:4216817-4216595
KZ282677.1:5819911-5828414_R
KZ283842.1:2738721-2736887
KZ198231.1:3316230-3316793_R
KZ282785.1:56276-56785
KZ196431.1:463521-464837_R
KZ197736.1:4292491-292981_R
KZ282280.1:13962962-13963452
KZ199231.1:2988555-2989116_R
KZ286466.1:11569437-11569927

```

```

689 -TTCCGCTGGTGAATTCGGTCC-CGCGCCG-TGG--GGT-CCAAGTACAGGAGGCTCGGCC-AGSTCAGACAACCGGGGTGCGACANNNNNIN 698
685 -TTCCGCTGGTGAATTCGGTCC-CGCGCCG-TGG--GGT-CCAAGTACAGGAGGCTCGGCC-AGSTCAGACAACCGGGGTGCGACANNNNNIN 698
642 -TTCAATCGGCTAA-TTCG-TCC-CGCTCTGG--GGT-TCCAAAGTATGAGAGGTCGCTGG-AGGTCAGACA 705
12294 -TTTCGCTGATTAATTCGGTCC-CCT-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 2237
524 -ATTGCTGCTGATAAATTCGGTCC-CCT-CTTGG--GGT-TCCAAAGTATGAGAGGTCGAGG-AGGTCAGACA 588
499 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 543
526 -ATTGCTGATAAATTCGGTCC-CCT-CTTGG--GGT-TCCAAAGTATGAGAGGTCGAGG-AGGTCAGACA 609
498 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 567
488 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 552
588 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 571
2143 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 2287
847 -ATTGCTGATAAATTCGGTCC-CCT-CTTGG--GGT-TCCAAAGTATGAGAGGTCGAGG-AGGTCAGACA 849
2845 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 2188
513 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 585
450 -TTTCGCTGGTGAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 589
434 -TTCAATCGGCTAAATTCGGTCC-CGC-CTTGG--GGT-CCAAGTATGAGAGGTCGAGG-AGGTCAGACA 497
441 -CTGCTGCTGTT--AGTCTATTG-AGCCGCTG-----C-----CTCGCTGGG-CGG-----GTCGCCACAGCCGGGGTCAGACAATTACCT 518
411 -CTGCTGCTGTT--TCGATCG-AGCCGCTGCTCTGCT-CC-----CGTGCT-CGG-----GTCGCCACAGCCGGGGTCAGACAATTACCA 484
411 -CTGCTGCTGTT--TCGATCG-AGCCGCTGCT-CC-TGCTCCCATGGCTGG-----GTCGCCACAGCCGGGGTCAGACAATTACCA 484
483 -CTGCTGCTGTT--TCGATCG-AGCCGCTGCT-CC-----CTGCTCCCATGGCTGG-CGG-----GTCGCCACAGCCGGGGTCAGACAATTACCA 556
412 -CTGCTGCTGTT--TCGATCG-AGCCGCTGCTCT-CC-----CGTGCT-CGG-----GTCGCCACAGCCGGGGTCAGACAATTACCA 485

```





# Sequence structures to consider

## ERV/LTR

\$ TSD.p1

- Consistent TSD length
  - True for soloLTRs and FL ERV
- 5' TG; 3' CA
- ORFs?

## LINE

- ORFs?

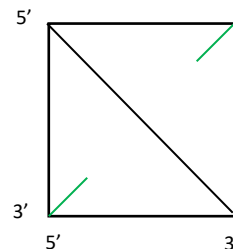
## SINE

- polyA tail
- G/C rich 5' end

- Homology to LINE 3' sequence
  - Exception is *Alu*
- RNA polymerase III A and B box
- Hairpin structure similar to tRNA?

## DNA

- Terminal inverted repeats (TIRs)
- Can be seen via dotplot
- [Curated termini](#)



# Reducing redundancy

```
$ rmbblast.pl <consensus_sequences.fa>  
[options]
```

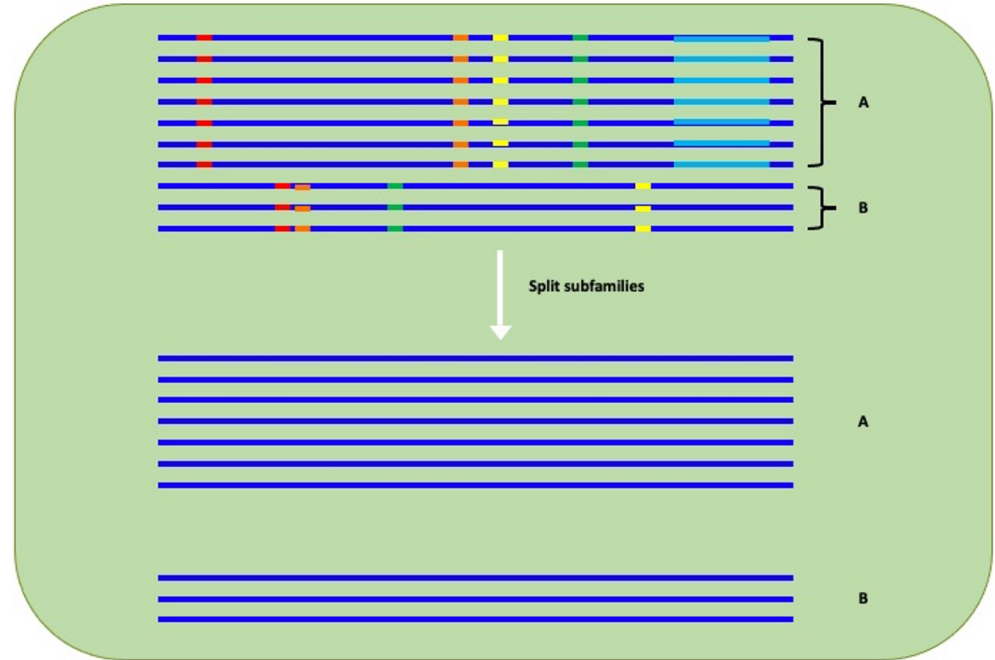
Crossmatch-like output (similar to RepeatMasker output)

**Table 2** Rmbblast.pl Output of the Eight Subfamilies Produced by ClusterPartialMatchingSubs.pl Analysis of example1

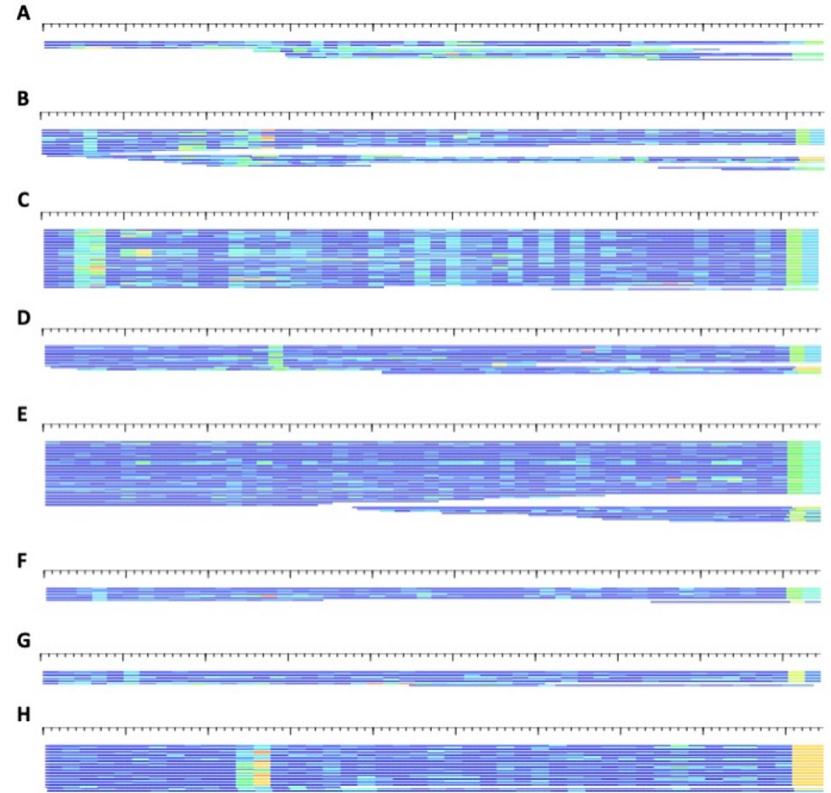
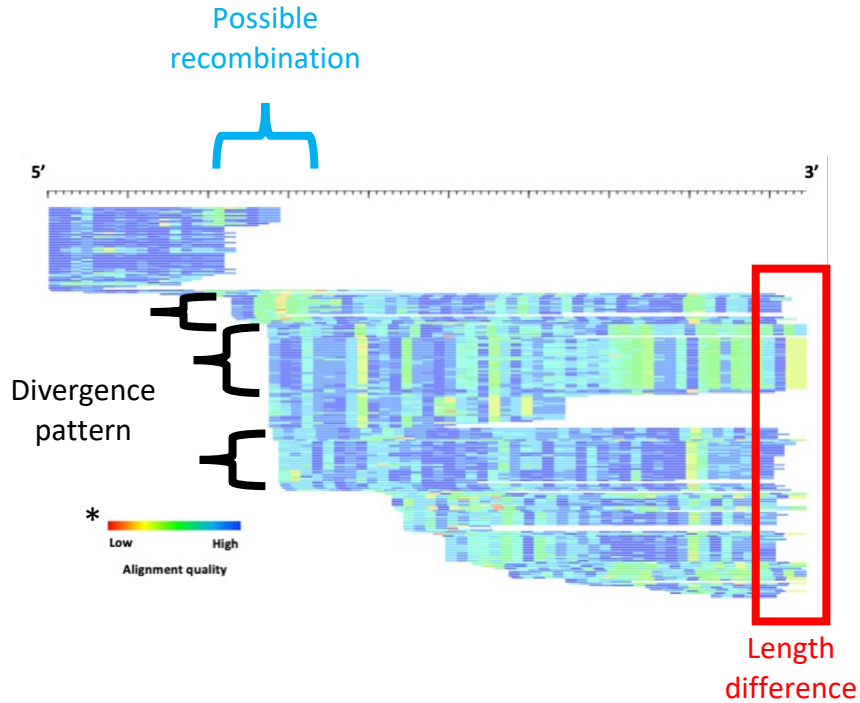
SW	Divergence	% del	% ins	Query	Query start	Query end	A_left	Target	Target start	Target end	T_left
4920	0.88	0	0	Cluster0	1	571	0	Cluster0	1	571	0
2704	2.1	0.35	32.33	Cluster0	1	571	0	Cluster12	1	433	0
2009	11.88	2.21	3.06	Cluster0	210	571	0	Cluster8	132	490	0
4051	0.22	0	0	Cluster10	1	465	0	Cluster10	1	465	0
2941	10.54	7.1	0.2	Cluster10	1	465	0	Cluster7	2	498	0
2725	8.6	13.76	1.34	Cluster10	1	465	0	Cluster6	1	522	0
3822	0.23	0	0	Cluster12	1	433	0	Cluster12	1	433	0
2735	2.77	32.33	0.35	Cluster12	1	433	0	Cluster0	1	571	0
4579	0.19	0	0	Cluster2	1	522	0	Cluster2	1	522	0
<b>3488</b>	<b>1.72</b>	<b>0</b>	<b>14.73</b>	<b>Cluster2</b>	<b>1</b>	<b>522</b>	<b>0</b>	<b>Cluster5</b>	<b>32</b>	<b>486</b>	<b>0</b>
<b>3076</b>	<b>4.62</b>	<b>1.93</b>	<b>14.5</b>	<b>Cluster2</b>	<b>4</b>	<b>522</b>	<b>0</b>	<b>Cluster8</b>	<b>29</b>	<b>490</b>	<b>0</b>
4268	0.21	0	0	Cluster5	1	486	0	Cluster5	1	486	0
3503	5.56	2.06	1.85	Cluster5	1	486	0	Cluster8	4	490	0
<b>3452</b>	<b>1.98</b>	<b>14.73</b>	<b>0</b>	<b>Cluster5</b>	<b>32</b>	<b>486</b>	<b>0</b>	<b>Cluster2</b>	<b>1</b>	<b>522</b>	<b>0</b>
4562	0	0	0	Cluster6	1	522	0	Cluster6	1	522	0
2787	11.3	1.53	6.64	Cluster6	1	522	0	Cluster7	2	498	0
2734	7.66	1.34	13.76	Cluster6	1	522	0	Cluster10	1	465	0
4354	0	0	0	Cluster7	1	498	0	Cluster7	1	498	0
2941	9.86	0.2	7.1	Cluster7	2	498	0	Cluster10	1	465	0
2806	11.87	6.64	1.53	Cluster7	2	498	0	Cluster6	1	522	0
4346	0.41	0	0	Cluster8	1	490	0	Cluster8	1	490	0
3524	5.54	1.85	2.06	Cluster8	4	490	0	Cluster5	1	486	0
<b>3045</b>	<b>5.19</b>	<b>14.5</b>	<b>1.93</b>	<b>Cluster8</b>	<b>29</b>	<b>490</b>	<b>0</b>	<b>Cluster2</b>	<b>4</b>	<b>522</b>	<b>0</b>
2053	11.98	3.06	2.21	Cluster8	132	490	0	Cluster0	210	571	0

# Subfamily assignment

- When to analyze subfamilies
  - Truncation patterns
    - DNA transposon deletion products
    - LTR recombination
  - Divergence
- Coseg
  - Full-length TE instances
- CD-HIT-based script
  - Length differences

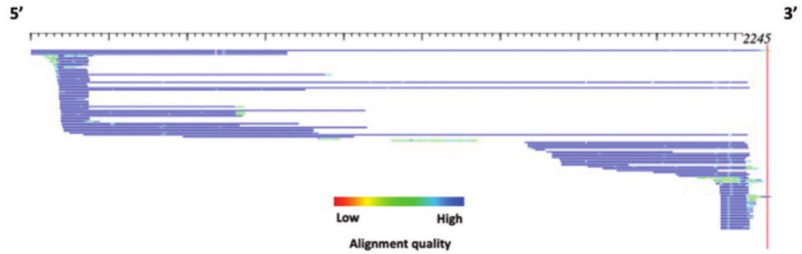


# Subfamily assignment - LTR

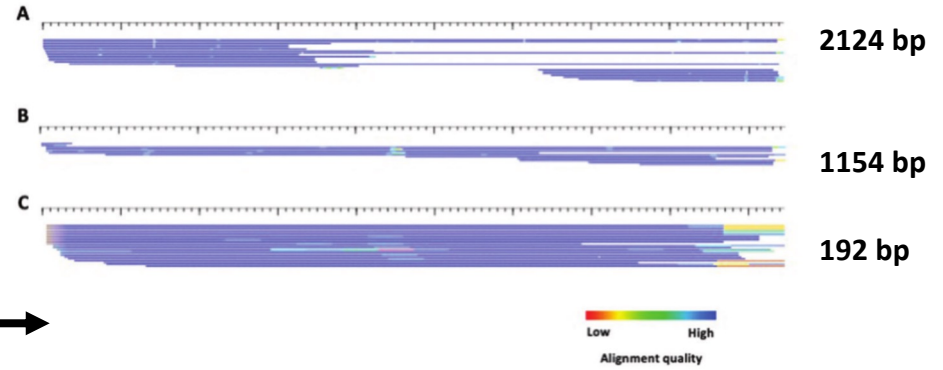


\* Each sequence is represented by a single row (sorted by start position) where the color gradient indicates alignment quality (red=low; blue=high) over 10bp non-overlapping windows.

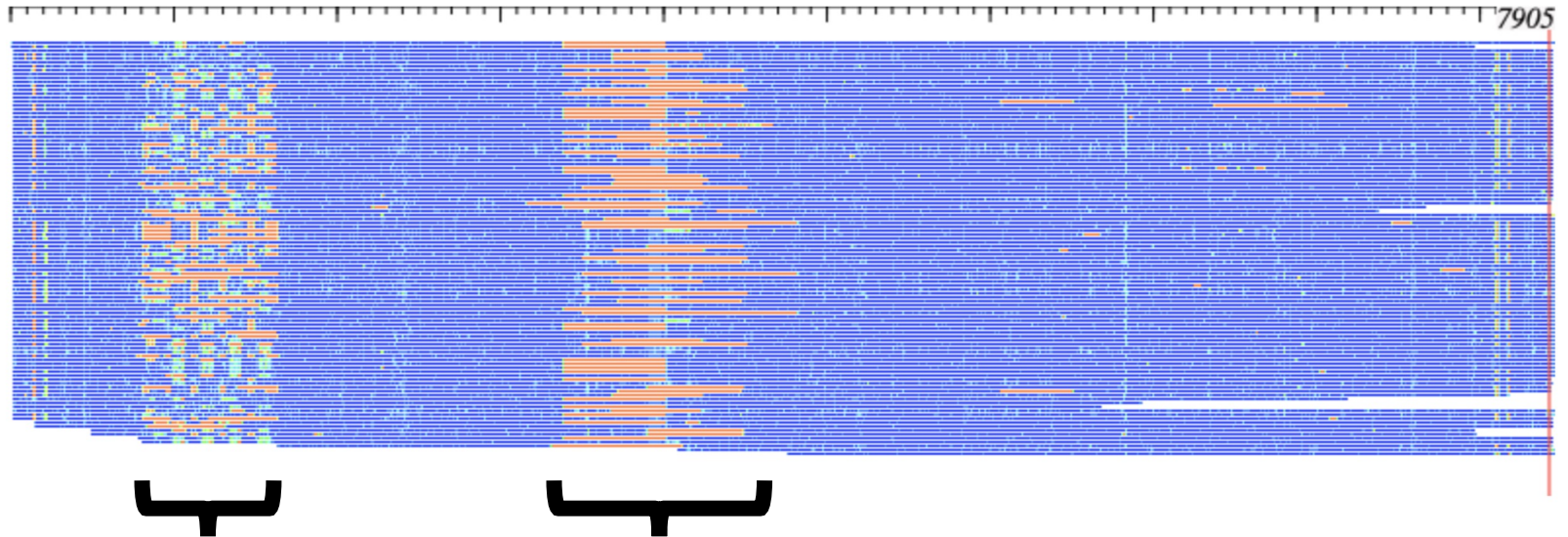
# Subfamily assignment - DNA



Subfamily analysis  
(CD-hit based script)



# PtERV (*Pan troglodytes* endogenous retrovirus) subfamily analysis

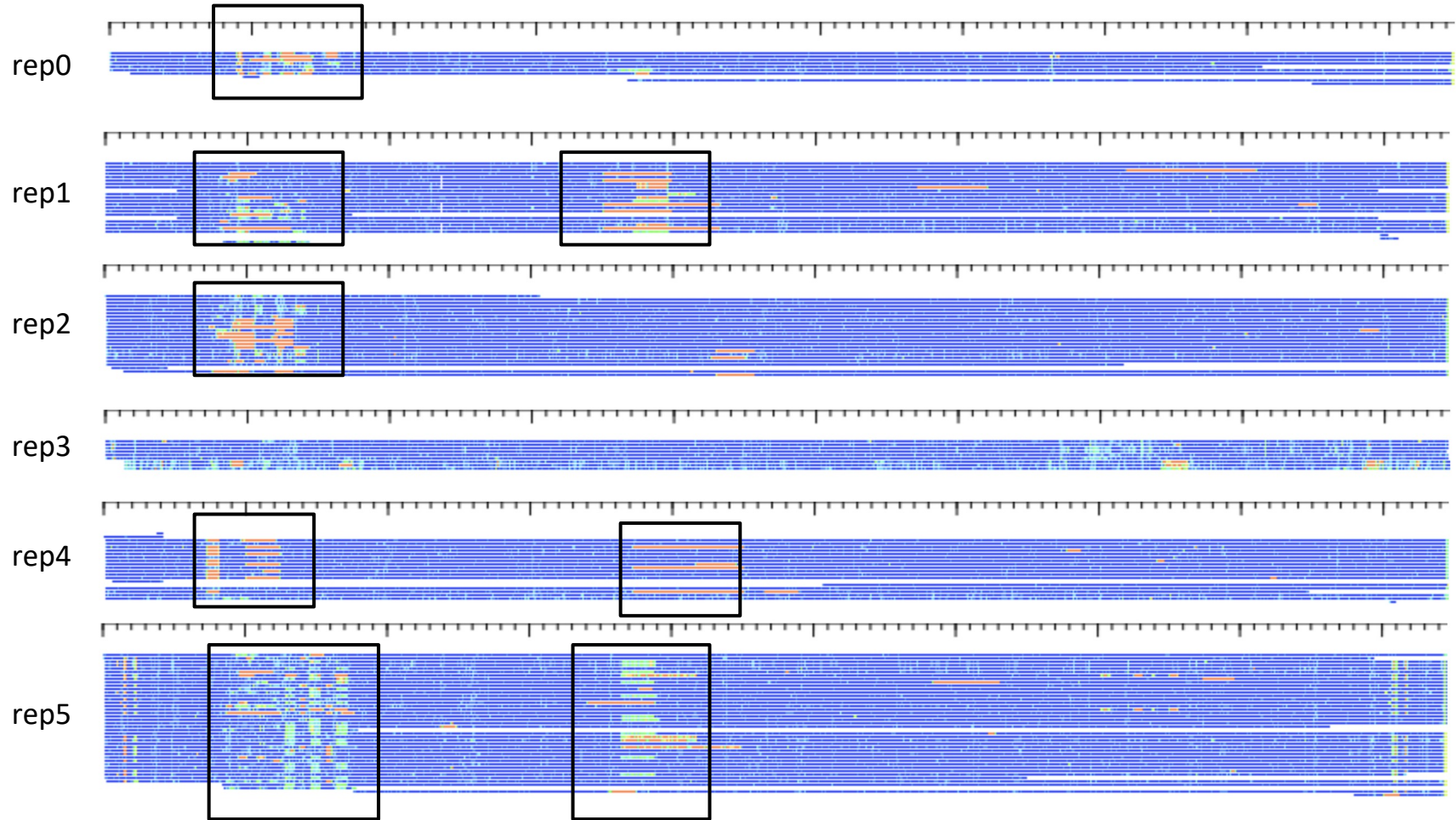


# PtERV subfamily analysis

## Strategy:

1. COSEG - 3 subfamilies produced
2. Separate subfamilies by divergence and/or length polymorphism
  - cross\_match
  - Divergence analysis
    - Split 3 COSEG subfamilies into 6 subfamilies









Subfamily	LTR	Int	LTR	Avg. div. (stdev)	Additional subfamilies?
rep0	1a	1b	1a	1.54 ± 0.9	2
rep1	1c	1c/d	1c/d	1.59 ± 0.55	2
rep2	1a	1a	1a	1.6 ± 1.06	2
rep3	2b/c	2a/b	2b/c	1.87 ± 0.71	2
rep4	1c	1a	1c/d	2.19 ± 0.81	2
rep5	1a/c	1b/c/d	1a/c	4.99 ± 2.18	3





# Telomere-to-telomere (T2T) - CHM13

## From telomere to telomere: the transcriptional and epigenetic state of human repeat elements

Savannah J. Hoyt, Jessica M. Storer, Gabrielle A. Hartley, Patrick G. S. Grady, Ariel Gershman, Leonardo G. de Lima, Charles Limouse, Reza Halabian, Luke Wojenski, Matias Rodriguez,  Nicolas Altemose,  Leighton J. Core, Jennifer L. Gerton,  Wojciech Makalowski, Daniel Olson, Jeb Rosen, Arian F. A. Smit,  Aaron F. Straight,  Mitchell R. Vollger,  Travis J. Wheeler, Michael C. Schatz, Evan E. Eichler, Adam M. Phillippy,  Winston Timp, Karen H. Miga,  Rachel J. O'Neill

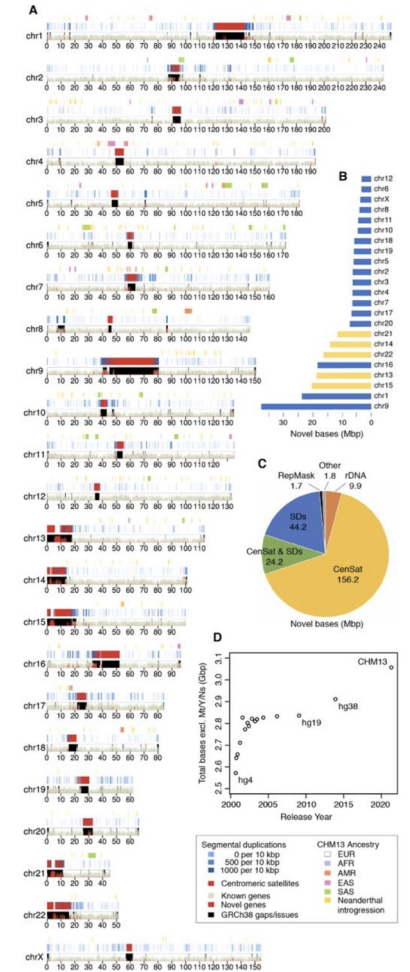
**doi:** <https://doi.org/10.1101/2021.07.12.451456>

# Telo

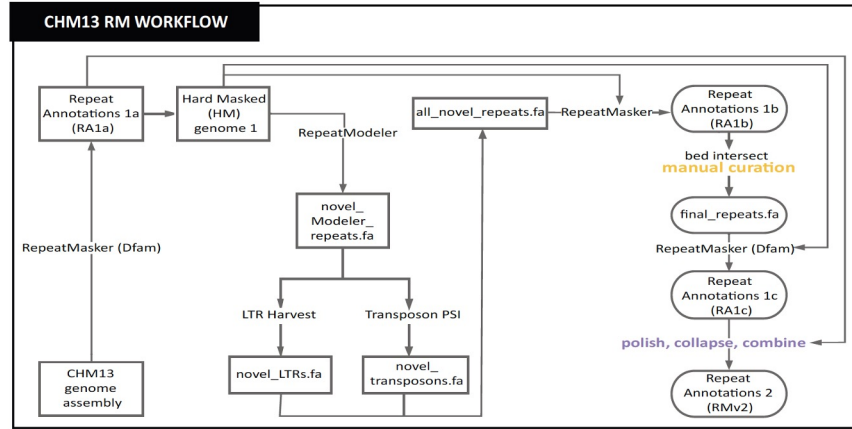
- Gap chr

- ~20

- Tra



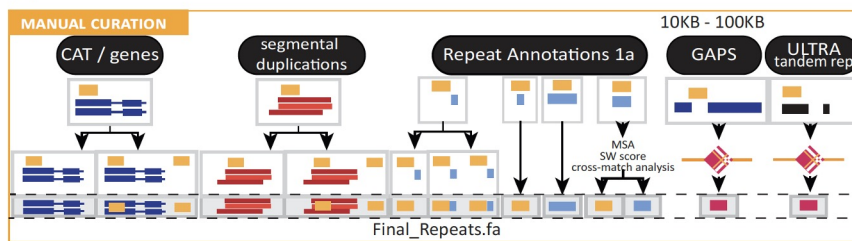
**A.**



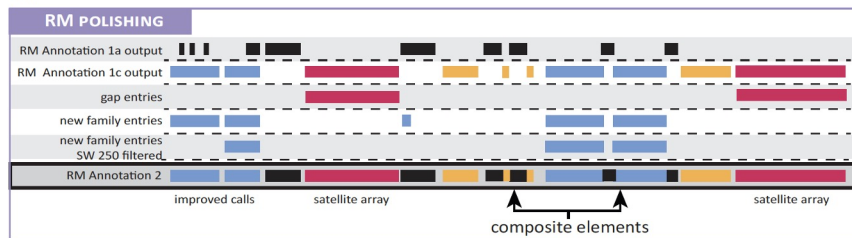
# M13

US

**B.**



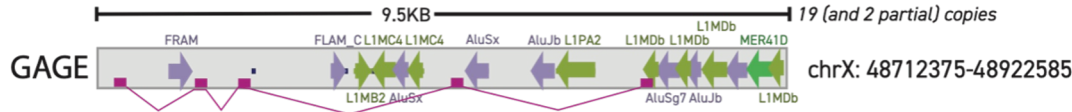
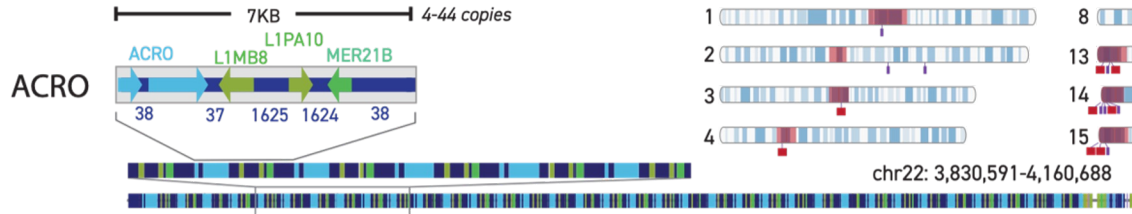
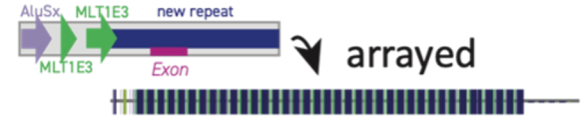
**C.**



# Composite repeats - CHM13

19 composites - 2.8 Mb

## Complex/Composite



**BMC Genomics**



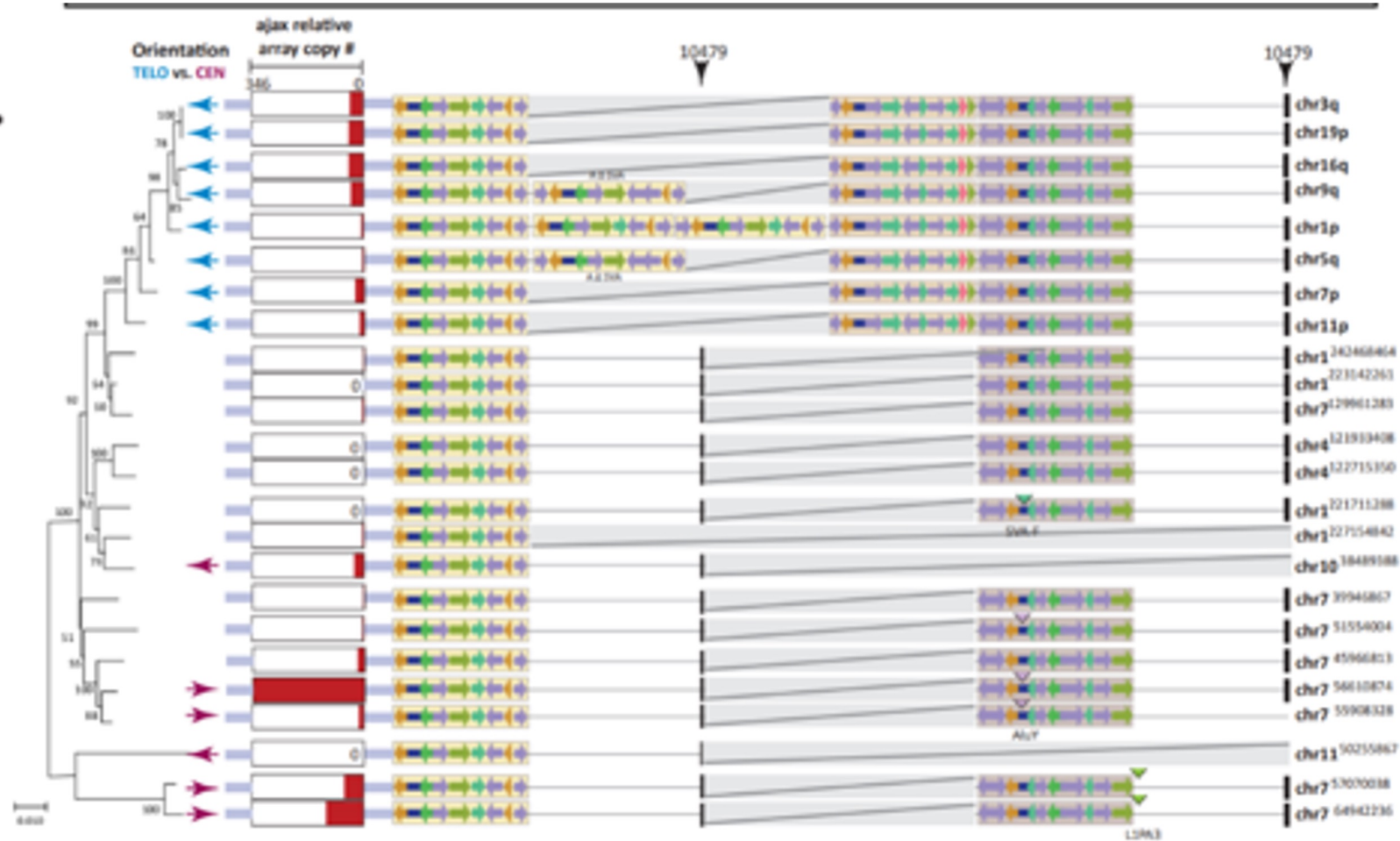
Research article

**Analysis of the largest tandemly repeated DNA families in the human genome**

Peter E Warburton\*, Dan Hasson, Flavia Guillem, Chloe Lescale, Xiaoping Jin and Gyorgy Abrusan

Open Access

C.



# Acknowledgements\*



Rachel O'Neill  
Savannah Hoyt  
Gabby Hartley  
Patrick Grady  
Christine McCann  
Emily Fuller  
Emry Brannan  
Laura Holt  
Michelle Neitzey  
Rich Green  
Nicole Pauloski  
Vel Johnston

EMBL-EBI



Fergal Martin  
Denye Ogeh



Mark Batzer  
Jerilyn Walker

zoonomia



Elinor Karlsson  
Kerstin Lindblad-Toh



TEXAS TECH  
UNIVERSITY

David A. Ray  
Nicole Paulat  
Austin Osmanski



Karen Miga  
Adam Phillippy  
Mark Diekhans



Universität  
Münster

Wojciech Makalowski  
Reza Halabian  
Felix Manske  
Matias Rodriguez  
Michelle Leyers



Arian Smit  
Robert Hubley  
Jeb Rosen  
Anthony Gray



NHGRI grant U24 HG010136  
and R01 HG002939



Family & friends

\***very** abbreviated list!

# Helpful links

## Publications

[Effect of different alignment tools on reconstructing TE sequences](#)

[TE discovery methodologies](#)

[Visualizing annotations](#)

[Advanced curation protocol](#)

[Beginner curation protocol](#)

## Tools

[RepeatModeler utilities](#)

[TE Aid](#)

[SODA](#)

[FlexiDot github](#)