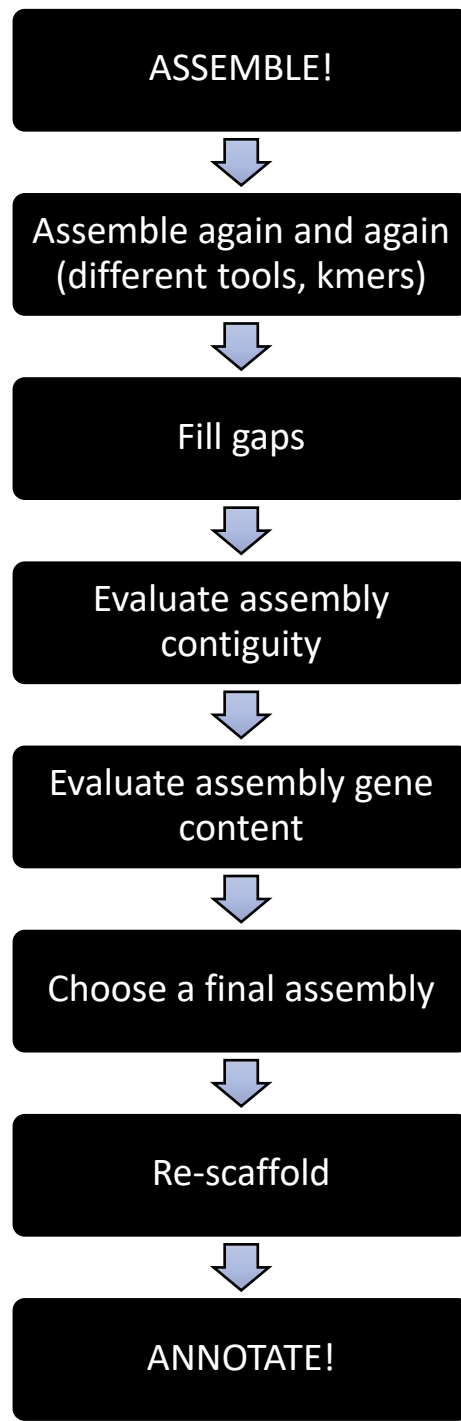
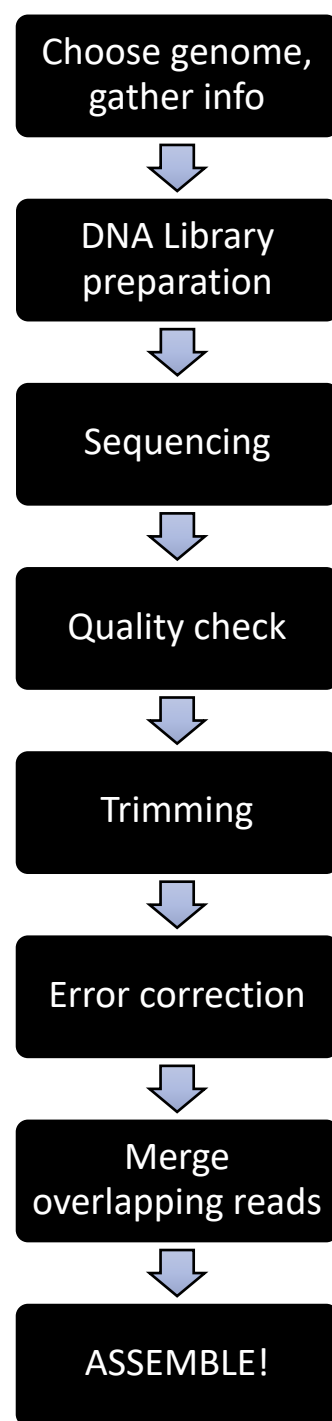
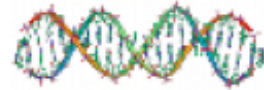


Genome Annotation





Sequence the genome



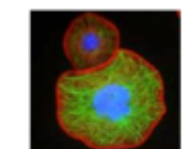
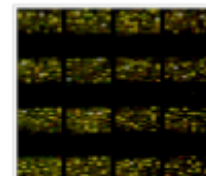
Annotate the genome



Build Database and Distribute Annotations



```
>Smg5
MEVTFSSGGSSNASSECAIDGGTNRRCRL
EPNNGTCILSQEVKDLYSLYTASKQLDD
AKRNVQSVGQLFQHEIEEKRSLLVQLCKQ
IIFKDYQSVGKKVREVMWRRGYEYFIAPV
```



Design Experiments



SUCCESS



FAILURE

ACGGGTTTCGCTACAGATGAACTGAATTTATACACGGACAACCTCATCGCCCATTTTGGGGCGTGGGGCACCGGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACCTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTTCATAGATATTTCTTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACCTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG

A GENOME ASSEMBLY IS JUST A TEXT FILE IN FASTA FORMAT.

CTGTTTCAGGGCTCTGCTTTGGTATCACTCAATATATTTAGACCCAGACAAAGTGGCAAAATTTGCTGCGGCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAAATTGCGGGCGAAAATAGAGGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGGCCGATTTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG

ACGGGTTTCGCTACAGATGAACTGAATTTATACACGGACAACATCATCGCCCATTTTGGGGCGTGGGGCACCGGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACCTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTTCATAGATATTTCTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACCTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG

IT IS MUCH MORE INTERESTING TO HAVE AN ANNOTATION.

CTGTTTCAGGGCTCTGCTTTGGTATCACTCAATATATTTAGGACCCAGACAAAGTGGCAAAATTTGCTGCGGCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAAATTGCGGGCGAAAATAGAGGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGGCCGATTTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG

Annotation Goals

Identifying repeats

- Biologically interesting
- Technically very important. Repeats are often “masked” in downstream analyses to avoid false positives and other issues

Identifying protein coding genes

- Build inventory of genes
- Identify boundaries of introns, exons, promoters
- Predict mRNA structure (remember Central Dogma)

Identifying other regions

- Noncoding RNAs
- Promoters
- Cis-regulatory regions

Understanding genome structure

- Centromeres
- Telomeres
- Mapping scaffolds to chromosomes

Community

- Creating a useful resource
- Output needs to be in a standard format



BED Format

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold. Any valid seq_region_name can be used, and chromosome names can be given with or without the 'chr' prefix.
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates
4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".
5. **score** - A score between 0 and 1000. See [track lines](#), below, for ways to configure the display style of scored data.
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **itemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value of itemRgb set to "on" (case-insensitive).
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

<https://m.ensembl.org/info/website/upload/bed.html>

BED Format

BED (9-column):

chr7	127471196	127472363	Pos1	0	+	127471196	127472363	255,0,0
chr7	127472363	127473530	Pos2	0	+	127472363	127473530	255,0,0
chr7	127473530	127474697	Pos3	0	+	127473530	127474697	255,0,0
chr7	127474697	127475864	Pos4	0	+	127474697	127475864	255,0,0

GFF-3 Format

Fields **must** be tab-separated. Also, all but the final field in each feature line must contain a value; "empty" columns should be denoted with a '.'

1. **seqid** - name of the chromosome or scaffold; chromosome names can be given with or without the 'chr' prefix.
2. **source** - name of the program that generated this feature, or the data source (database or project name)
3. **type** - type of feature. Must be a term or accession from the SOFA sequence ontology
4. **start** - Start position of the feature, with sequence numbering starting at 1.
5. **end** - End position of the feature, with sequence numbering starting at 1.
6. **score** - A floating point value.
7. **strand** - defined as + (forward) or - (reverse).
8. **phase** - One of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on..
9. **attributes** - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent

GFF-3 Format

GFF3:

```
##gff-version 3
```

```
ctg123 . mRNA          1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123 . exon          1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123 . exon          1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123 . exon          3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123 . exon          5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123 . exon          7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

Annotations can be made

With ***ab initio*** methods based on the understanding of particular properties of different genome features

Based on direct evidence, such as RNA-seq

Based on comparison to a reference of similar sequences

- Looking for repetitive DNA elements
- Blasting known protein coding genes

BLAST – Basic Local Alignment Search Tool

Requirements

- Query (sequence you want to identify)
- Database (reference)

Local alignment

- proteins are modular, genes contain exons and introns
- versus global alignment

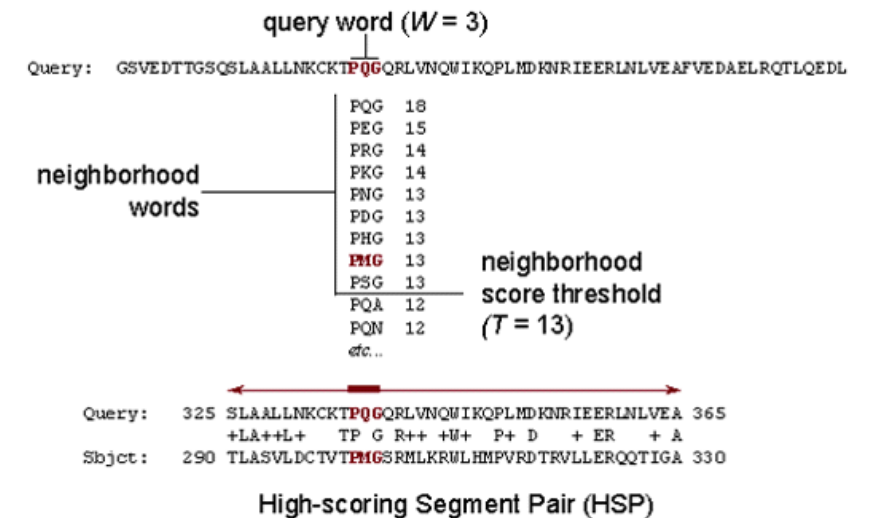
Neighborhood

- BLAST considers exact words, but also similar ones according to BLOSUM26 matrix
- these are aligned and then *extended*
- cumulative score is tallied

Goal

- when score drops significantly, extension is trimmed
- this results in the high scoring segment pair

The BLAST Search Algorithm



BLAST is used for many things.

BLAST searching is fundamental to understanding the relatedness of any favorite query sequence to other known proteins or DNA sequences.

Applications include:

- identifying orthologs and paralog
- discovering new genes and proteins
- discovering variants of genes or proteins
- investigating expressed sequence tags (ESTs)
- exploring protein structure and function

Types of BLAST searches

blastn (nucleotide BLAST) -> DNA query to DNA database

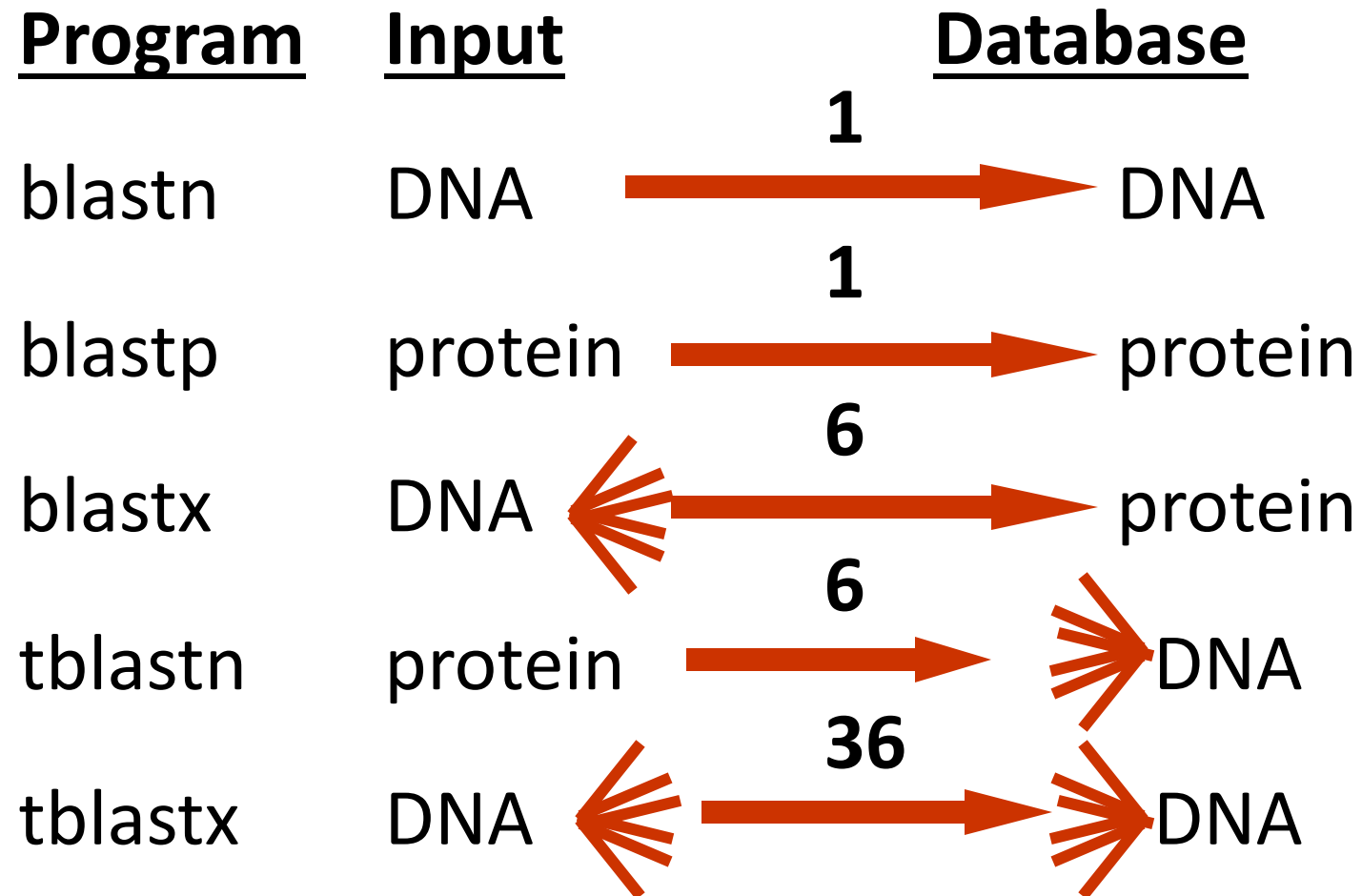
blastp (protein BLAST) -> protein query to protein database

blastx (translated BLAST) -> DNA query to protein database

tblastn (translated BLAST -> protein query to DNA database

tblastx (translated BLAST) -> translated DNA query to translated DNA database

Choose a BLAST program based on your needs



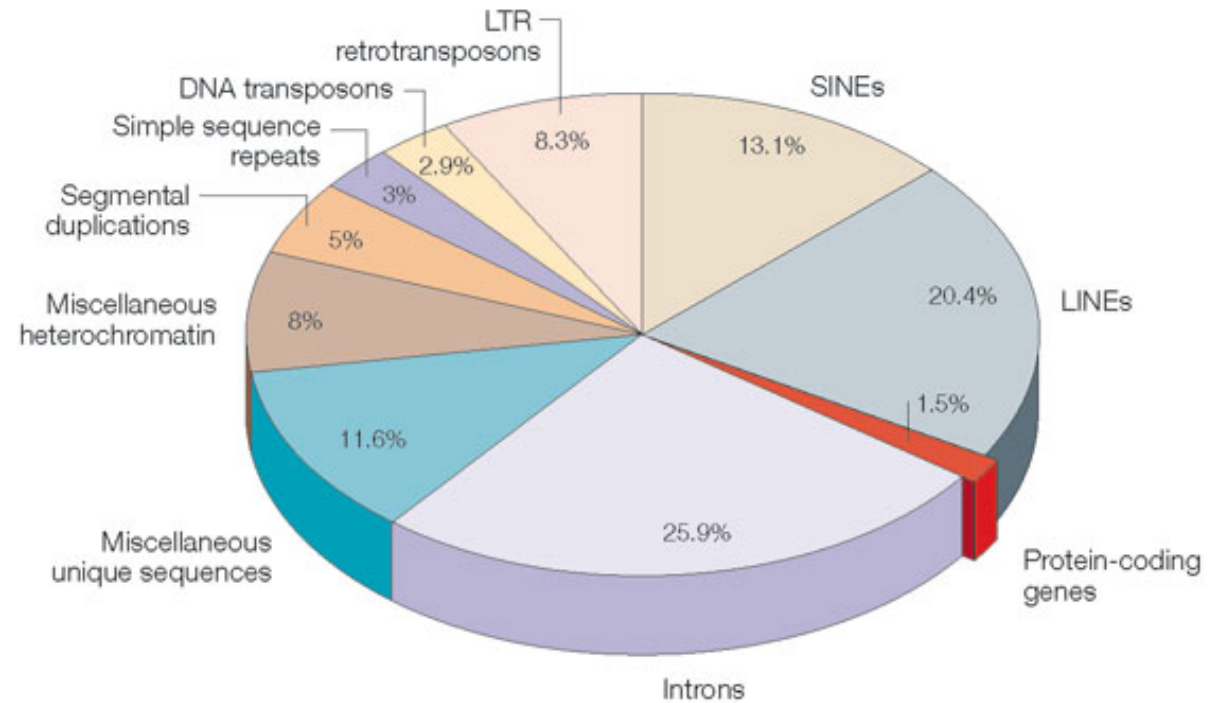
Repetitive DNA

Tandem repeats

- Adjacent along the chromosome
- Satellite DNA
 - minisatellites – 10-60bp
 - microsatellites - <10bp

Interspersed repeats

- Transposable elements



Repeat Masking

Essential step before gene annotation

Many repeats contain ORFs, can be mistaken as genes.

Repeats should be “masked” before you try to annotate genes

- “NNNNN” = hardmasked

- “atcg” = softmasked

 - Better for downstream BLAST or genome alignment

Two Methods of Repeat Finding

- ***Database method***

- RepeatMasker (repeatmasker.org)
 - Blast RepBase elements to your genome
 - Good for mammals or model organisms
 - Ascertainment bias – species have unique repeats
 - Human: 50% masked with “homo sapiens” repeats.
 - Humpback whale: 38% masked with “mammalia” repeats.
 - Glass lizard: 13% masked with “vertebrate” repeats

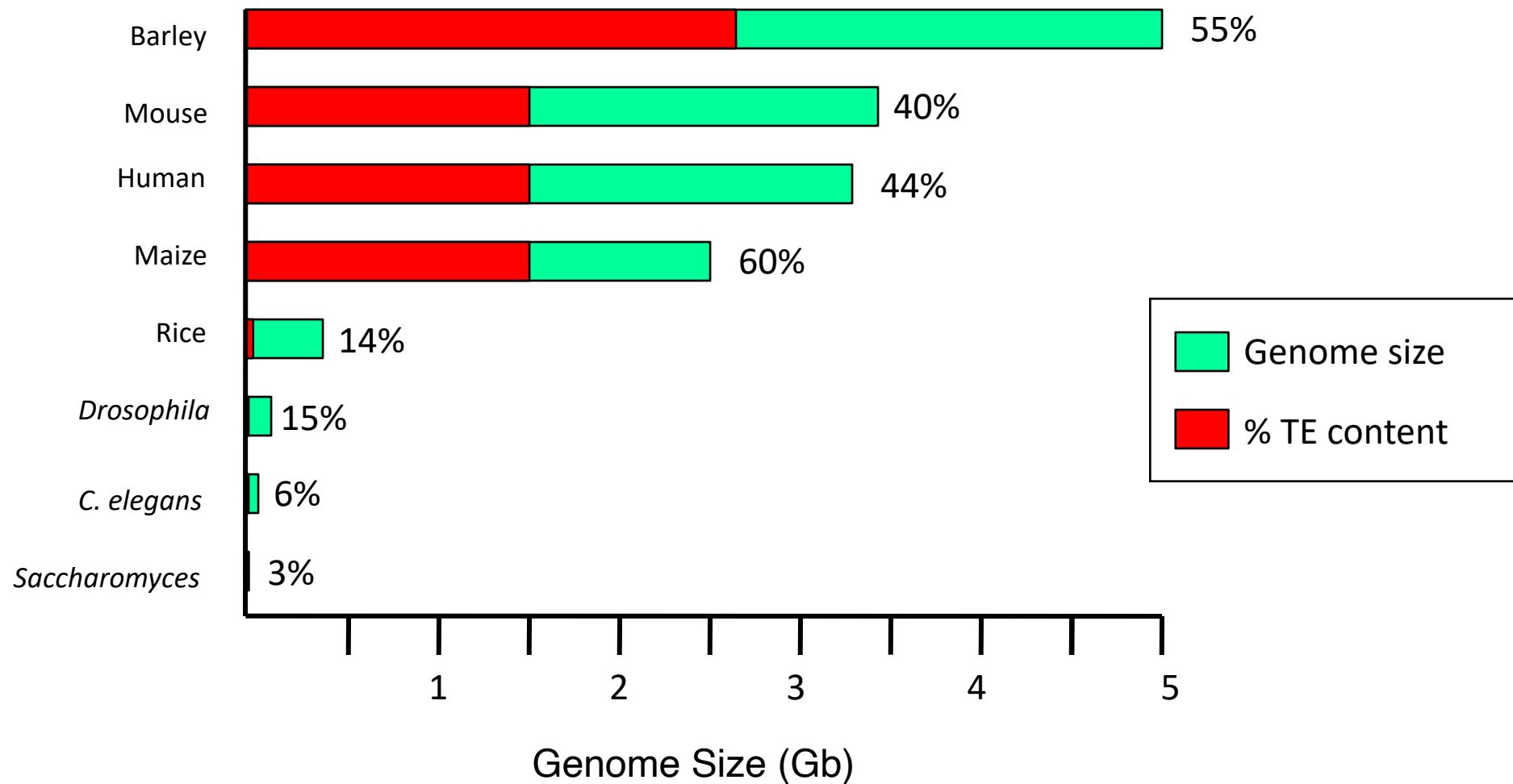
- ***De novo method***

- RepeatModeler (repeatmasker.org)
 - Blast your genome to itself
 - Models repeats without *a priori* knowledge
 - Good for finding species-specific repeats
 - May miss low-copy number repeats

What are Transposable Elements?

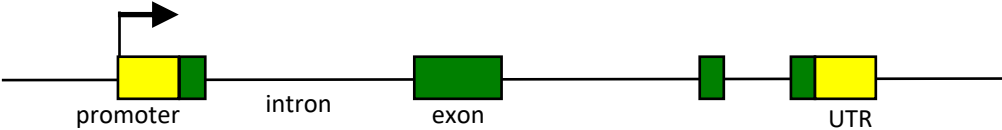
- DNA sequences that move about the genome
- Transposition
- “jumping genes”
- “junk DNA”, “selfish genes”

Large genomes have a lot of transposable elements

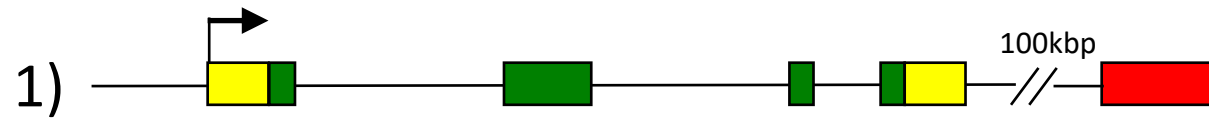


TEs Effect Can Affect Gene Function and Regulation

Transposable element: 

Average eukaryotic gene: 

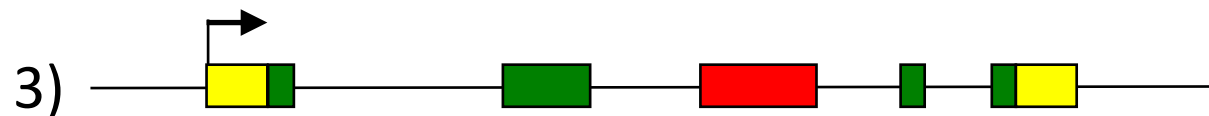
promoter intron exon UTR



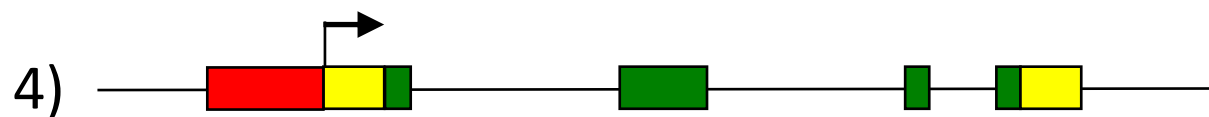
No effect



Gene disruption



Likely no effect?



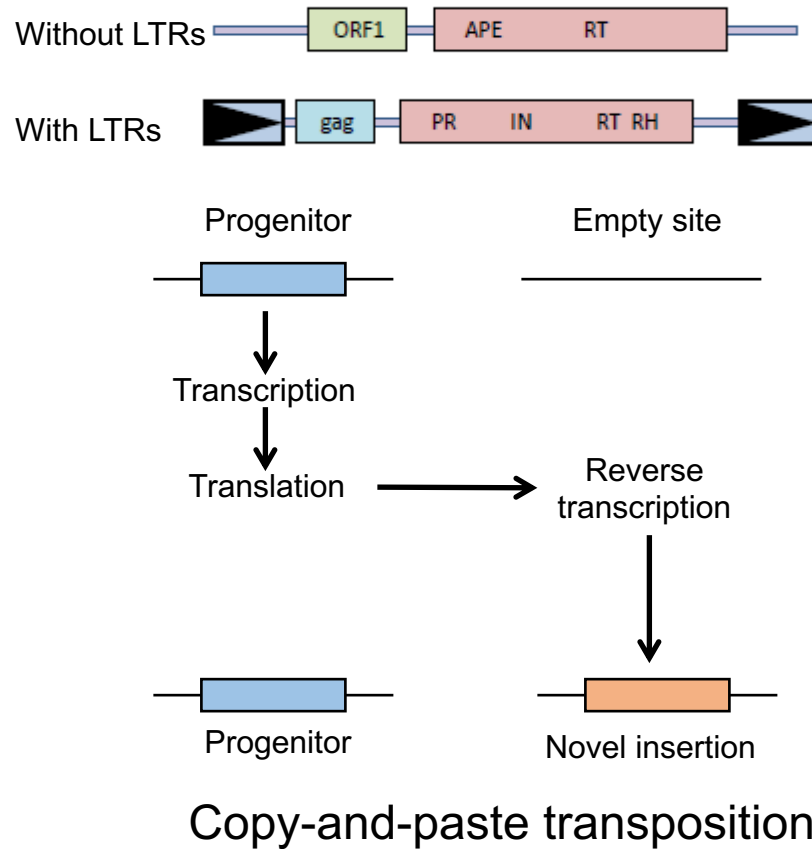
Effects gene regulation?

Examples of Human Disease Caused by TE Mutation

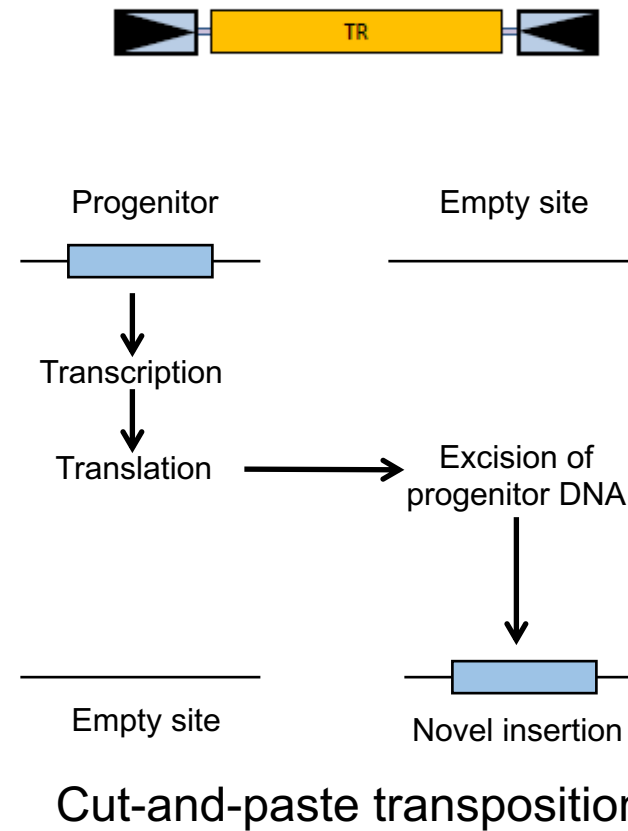
	Gene	Disorder	Element	Mechanism
Alu	NF1	Neurofibromatosis	Alu Ya5	Intron/skipping
	BCHE	Acholinesterasemia	Alu Yb8	Exon insertion
	F9	Hemophilia B	Alu Ya5	Exon insertion
	CASR	Familial hypocalciuric hypercalemia	Alu Ya4	Exon insertion
	ADD1	Huntington's disease	Alu	Exon insertion
LINE-1	Factor VIII	Hemophilia A	L1	Exon insertion
	APC	FAP	L1	Exon insertion
	Dystrophin	Muscular Dystrophy	L1	Exon insertion
	Globin	Beta thalassemia	L1	Intron
	RP2	Retinitis Pigmentosis	L1	Intron
	Fukutin	Muscular Dystrophy	L1	Intron/skipping

Types of transposable elements and their mode of transposition

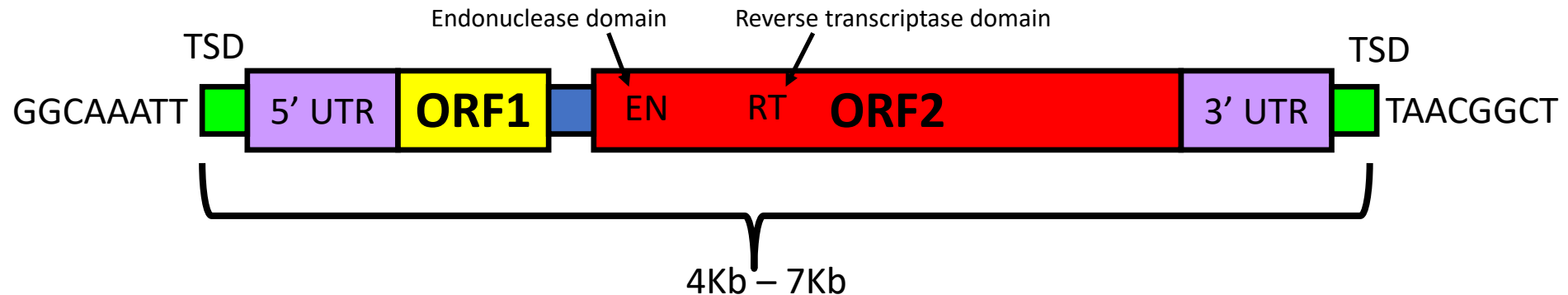
Class I: Retrotransposons



Class II: DNA transposons



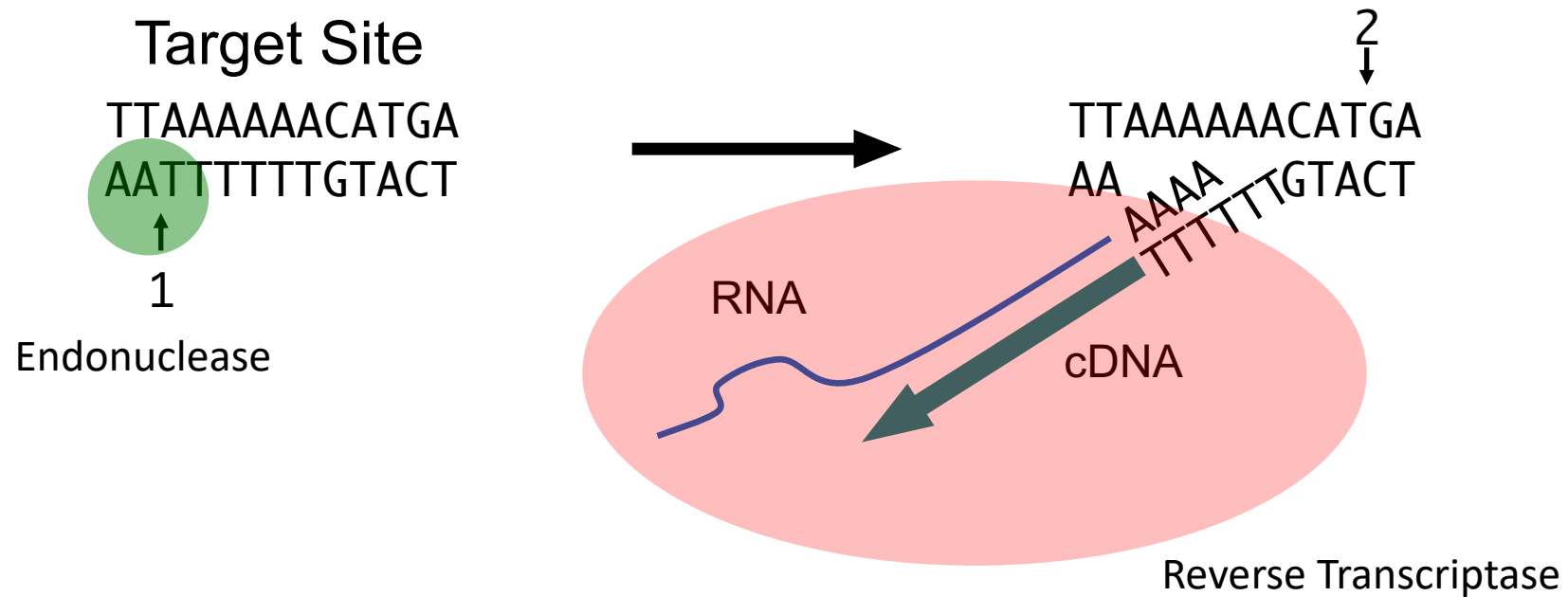
LINE-1: Long Interspersed Nuclear Element (or, L1)



Autonomously replicating non-LTR retrotransposon

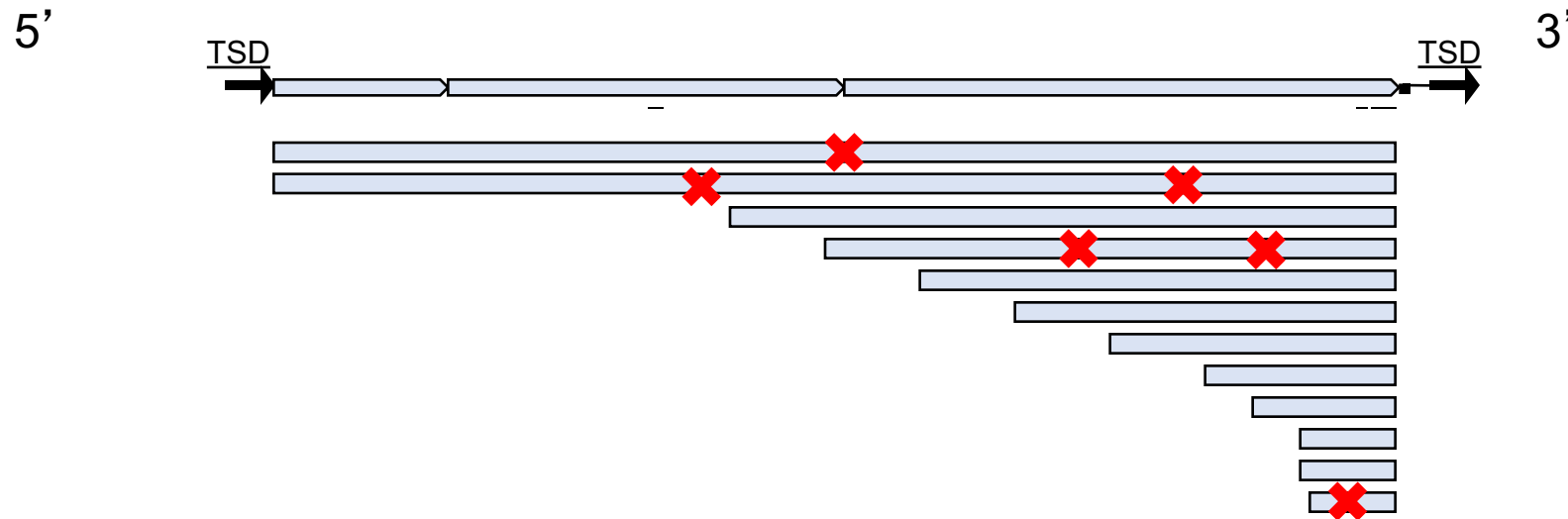
Produces both full-length (FL) and truncated (TR) copies

Target-primed Reverse Transcription (TPRT)



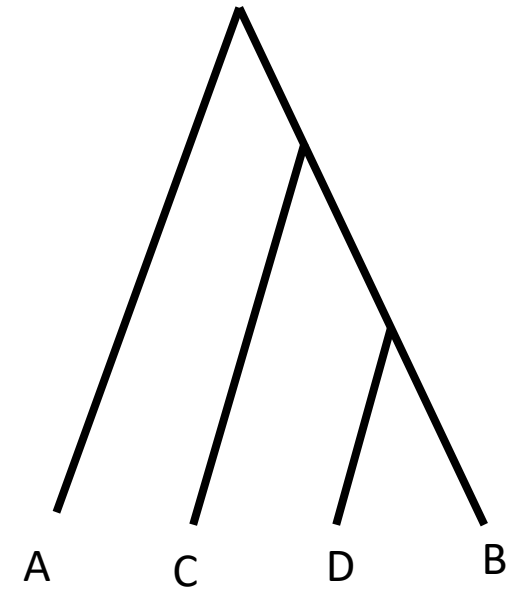
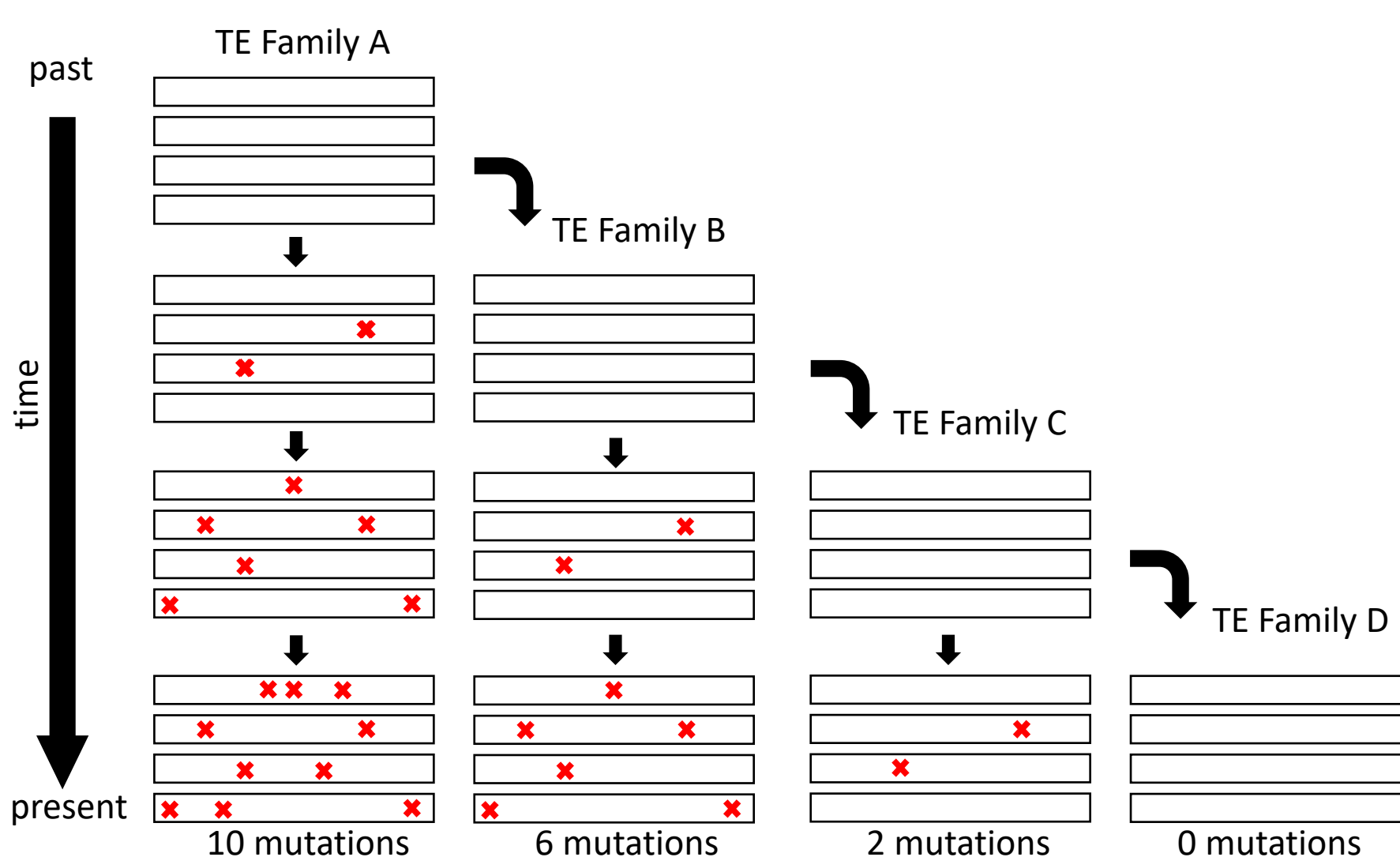
TPRT produces mostly truncated elements

Most of the products of non-LTR retrotransposon replication are truncated and thus not able to replicate.



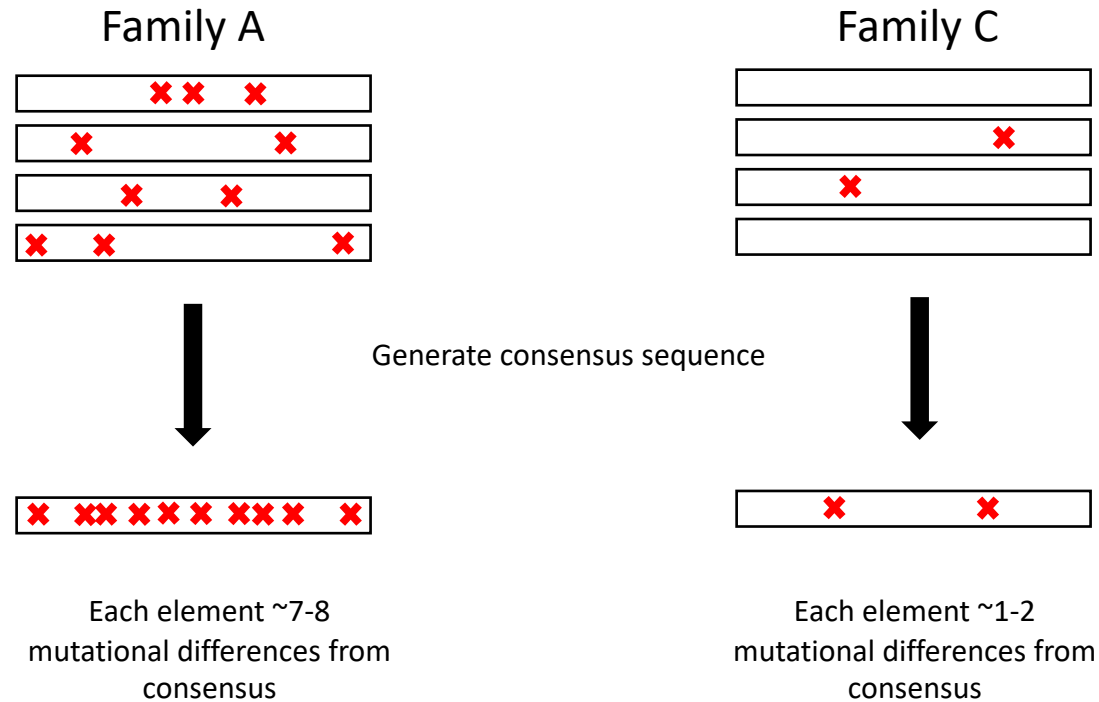
In addition, since replication is error prone, some Full-Length elements may contain disabling mutations.

TEs Accumulate Mutations over Time at the Pseudogene Rate



You can also reconstruct the evolutionary history of TE families in a genome.

Estimating Relative Ages of Elements within TE Families



The calculated distance from consensus is a proxy of the age of the insertion.

Comparative genomics shows that each lineage has it's own repertoire of TEs.

These repertoires can be described in terms of the abundance, diversity and age of TEs in the genome.



Studying the interaction of these intragenomic parasites with their host genomes is a major field of comparative genomics

