Module 10: Genome Annotation

Eugene Gardner

Postdoctoral Fellow, Wellcome Sanger Institute

e : eg15@sanger.ac.uk

: @DrGeneUK

We finished genome assembly… what next?

Genome Assembly

- Genome assembly identifies the DNA sequence of an organism

We finished genome assembly… what next?

Genome Assembly

- Genome assembly identifies the DNA sequence of an organism
- What do the letters actually tell us?
  - Function?
  - Evolution?
  - Development?

Z	V	E	Z	N	O	I	L	Z	O	E	A
C	E	N	N	A	G	A	Z	E	L	L	E
G	O	H	T	R	A	W	O	D	A	K	L
H	A	K	E	B	B	T	G	E	F	U	L
Y	C	M	L	E	B	A	C	E	F	D	E
E	H	O	E	Z	S	A	B	R	U	U	C
N	E	N	P	L	P	G	E	O	B	E	R
A	E	K	H	E	R	I	A	L	O	H	O
E	T	E	A	O	I	R	B	E	A	N	C
C	A	Y	N	P	N	A	R	H	I	N	O
I	H	N	T	A	G	F	E	N	E	T	D
V	H	N	I	R	B	F	P	Y	P	I	I
E	O	E	K	D	O	E	K	T	E	T	L
T	K	B	I	G	K	H	I	P	P	O	E

We finished genome assembly… what next?

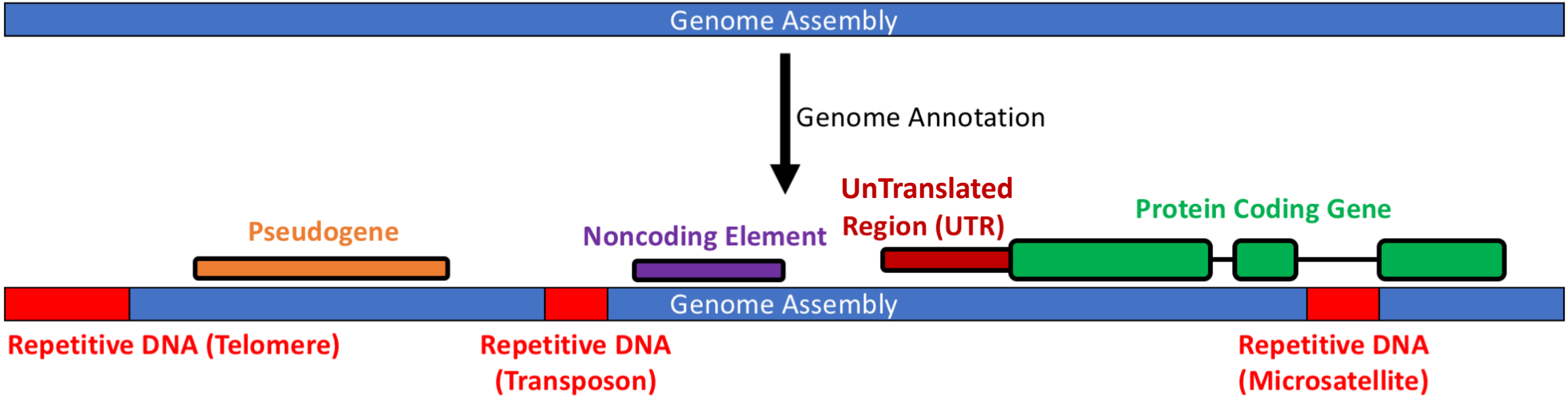
Genome Assembly

- Genome assembly identifies the DNA sequence of an organism
- What do the letters actually tell us about an organism?
  - Function?
  - Evolution?
  - Development?

Z	V	E	Z	N	O	I	L	Z	O	E	A
C	E	N	N	A	G	A	Z	E	L	L	E
G	O	H	T	R	A	W	O	D	A	K	L
H	A	K	E	B	B	T	G	E	F	U	L
Y	C	M	L	E	B	A	C	E	F	D	E
E	H	O	E	Z	S	A	B	R	U	U	C
N	E	N	P	L	P	G	E	O	B	E	R
A	E	K	H	E	R	I	A	L	O	H	O
E	T	E	A	O	I	R	B	E	A	N	C
C	A	Y	N	P	N	A	R	H	I	N	O
I	H	N	T	A	G	F	E	N	E	T	D
V	H	N	I	R	B	F	P	Y	P	I	I
E	O	E	K	D	O	E	K	T	E	T	L
T	K	B	I	G	K	H	I	P	P	O	E

Z	V	E	Z	N	O	I	L	Z	O	E	A
C	E	N	N	A	G	A	Z	E	L	L	E
G	O	H	T	R	A	W	O	D	A	K	L
H	A	K	E	B	B	T	G	E	F	U	L
Y	C	M	L	E	B	A	C	E	F	D	E
E	H	O	E	Z	S	A	B	R	U	U	C
N	E	N	P	L	P	G	E	O	B	E	R
A	E	K	H	E	R	I	A	L	O	H	O
E	T	E	A	O	I	R	B	E	A	N	C
C	A	Y	N	P	N	A	R	H	I	N	O
I	H	N	T	A	G	F	E	N	E	T	D
V	H	N	I	R	B	F	P	Y	P	I	I
E	O	E	K	D	O	E	K	T	E	T	L
T	K	B	I	G	K	H	I	P	P	O	E

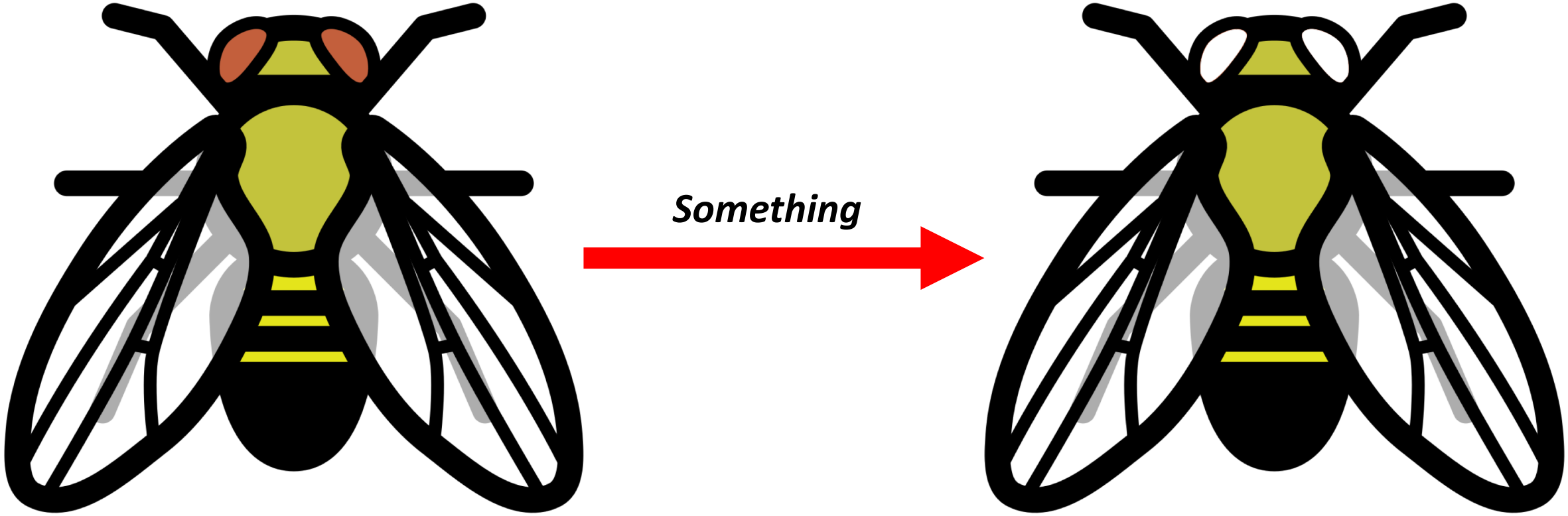
We finished genome assembly… what next?



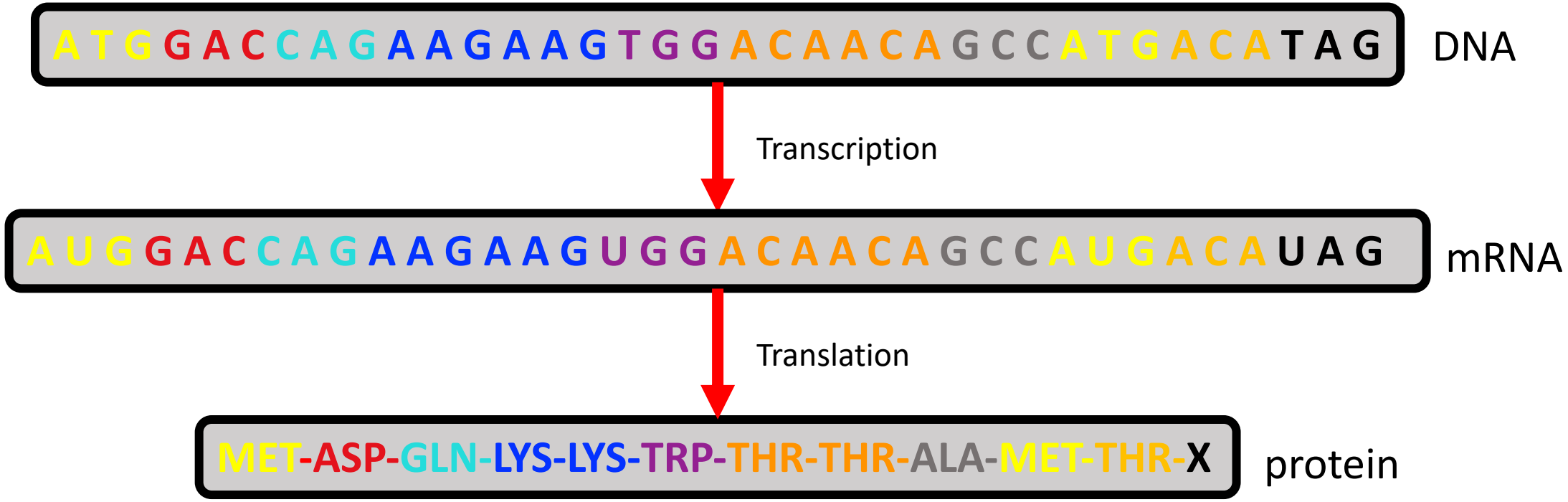
Genome Annotation – The process of assigning meaning to the sequence of a genome

- Genomes have many different types of sequences
- Some sequences can be assigned to a specific category
  - Genes
  - Repetitive Sequence
  - Functional elements (i.e. enhancers / silencers)

What is a Gene?



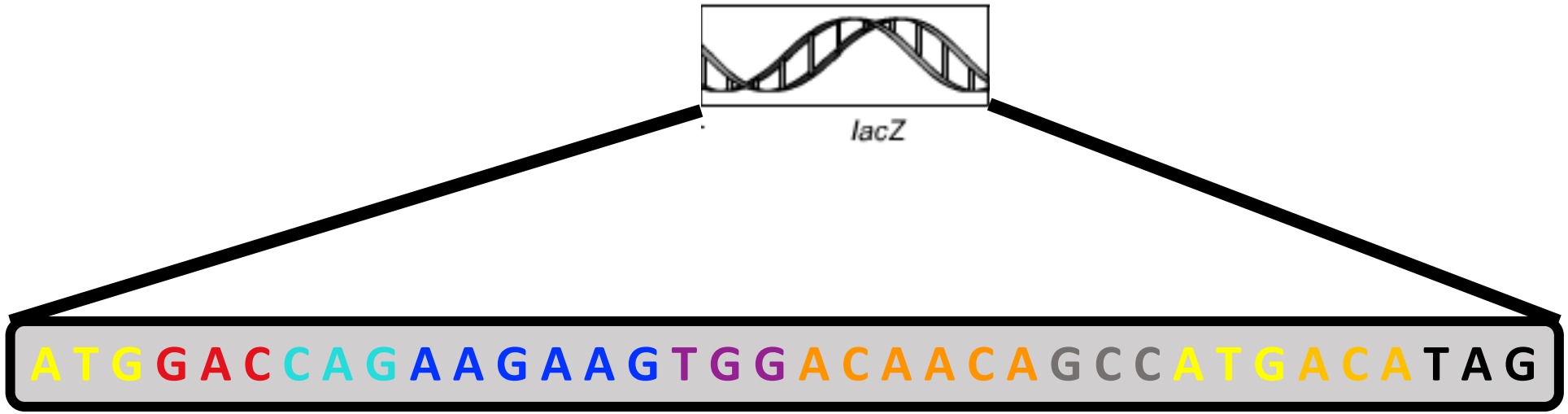
What is a Gene?



Genes in Bacteria

Images from the public domain

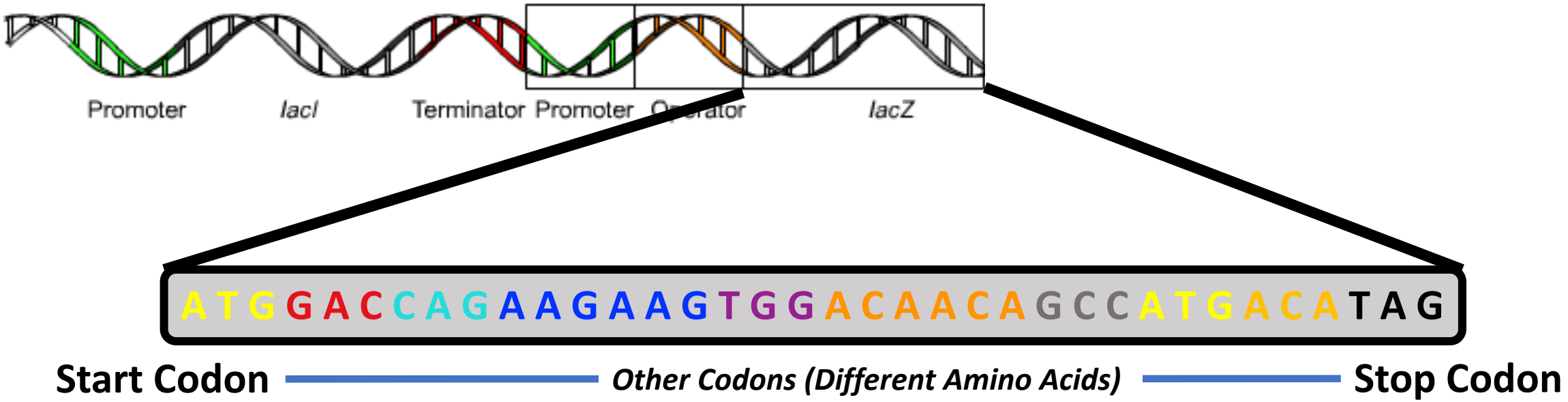
Genes in Bacteria



Start Codon — ***Other Codons (Different Amino Acids)*** — **Stop Codon**

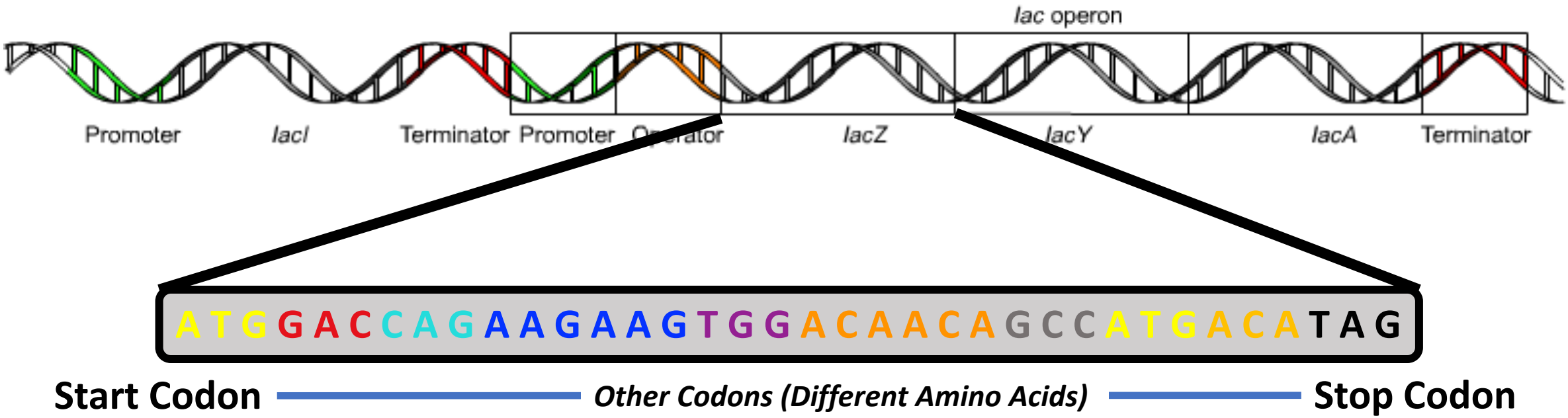
Images from the public domain

Genes in Bacteria



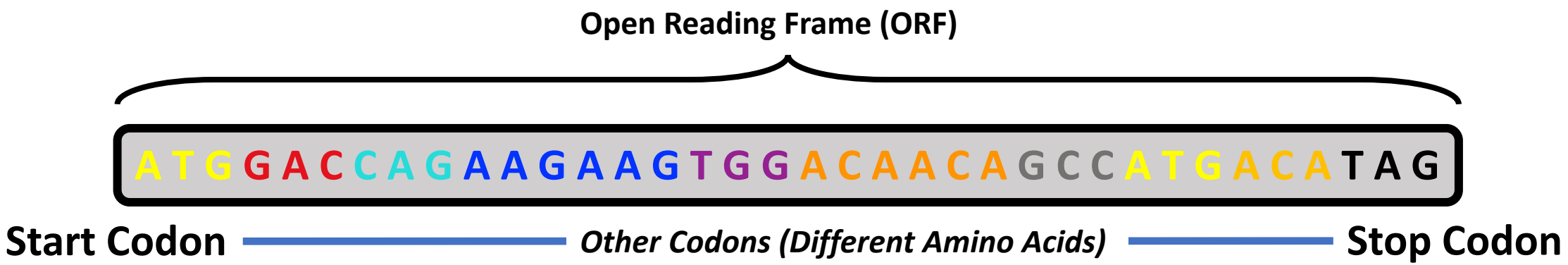
Images from the public domain

Genes in Bacteria

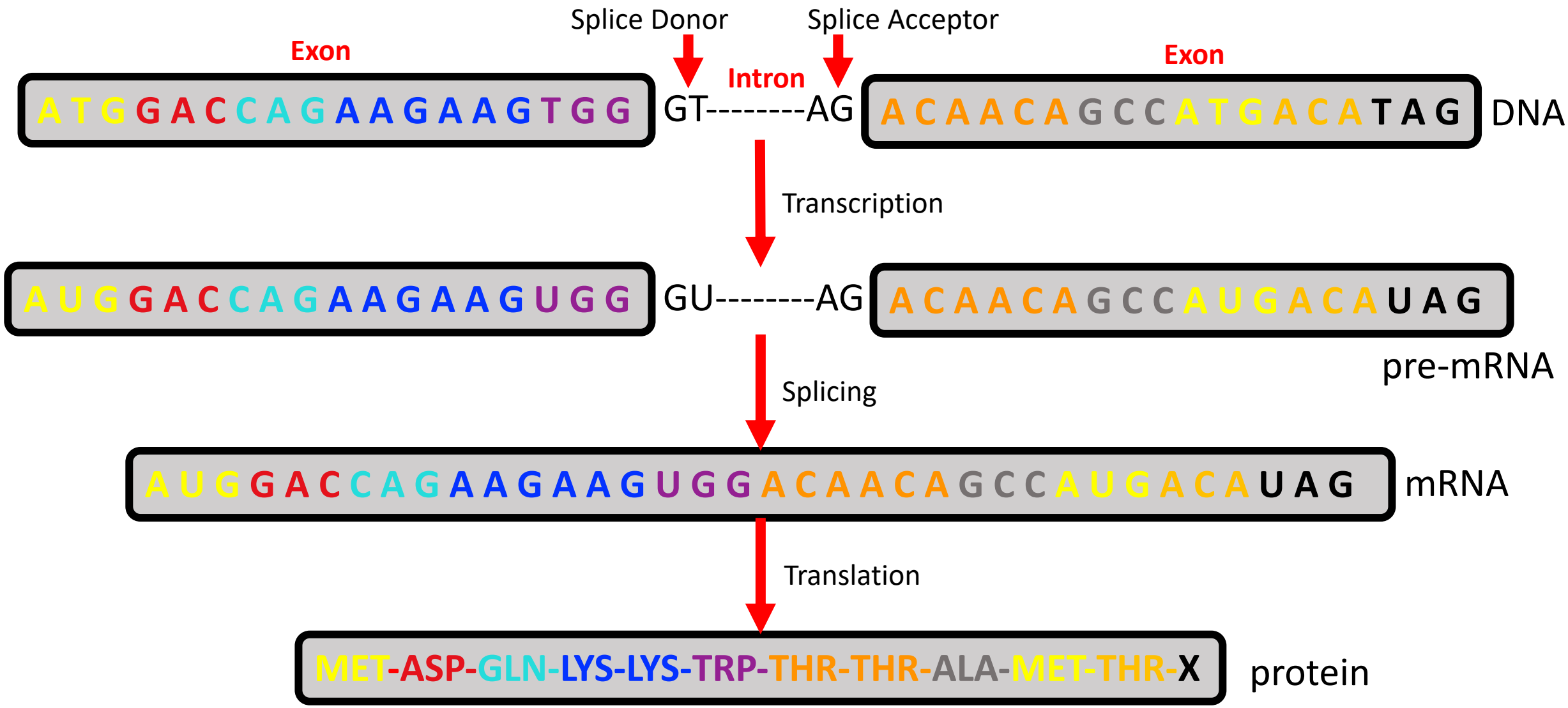


Images from the public domain

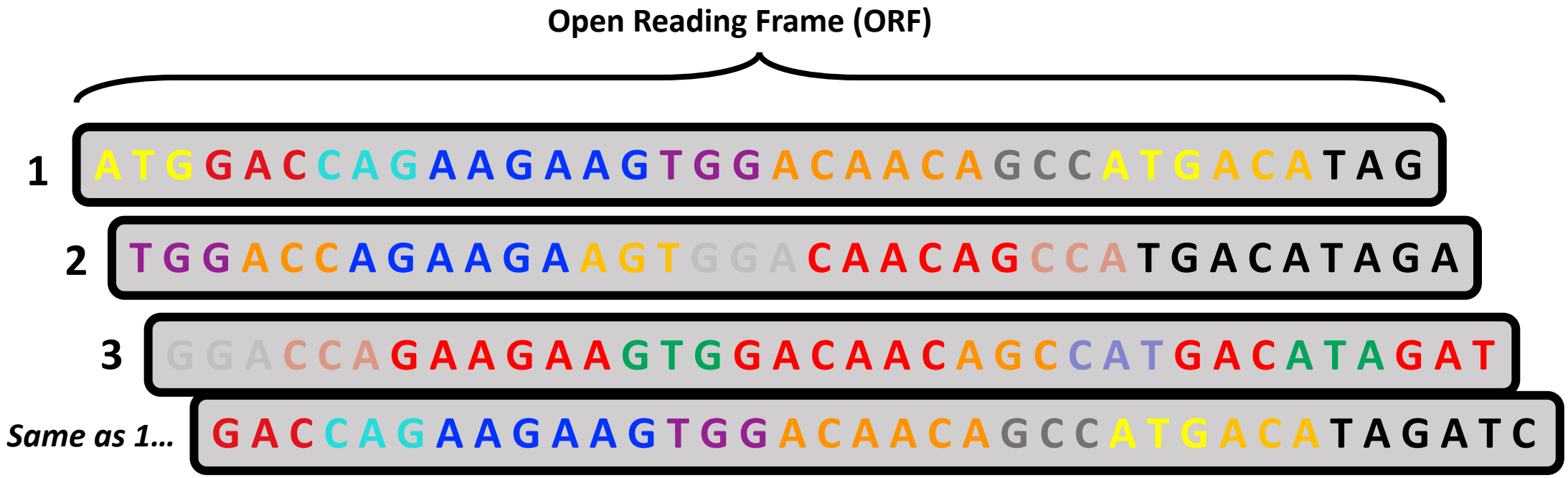
Simplest Gene Discovery



Genes in Eukaryotes

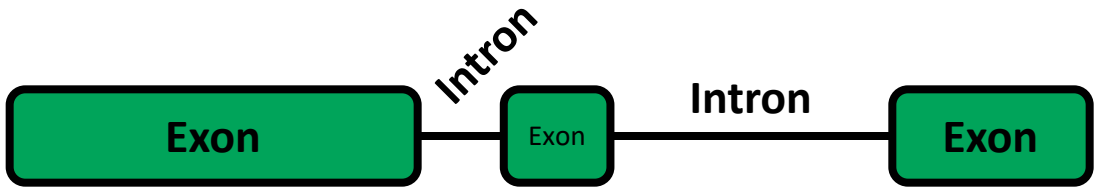


Frame



And also the reverse complement…

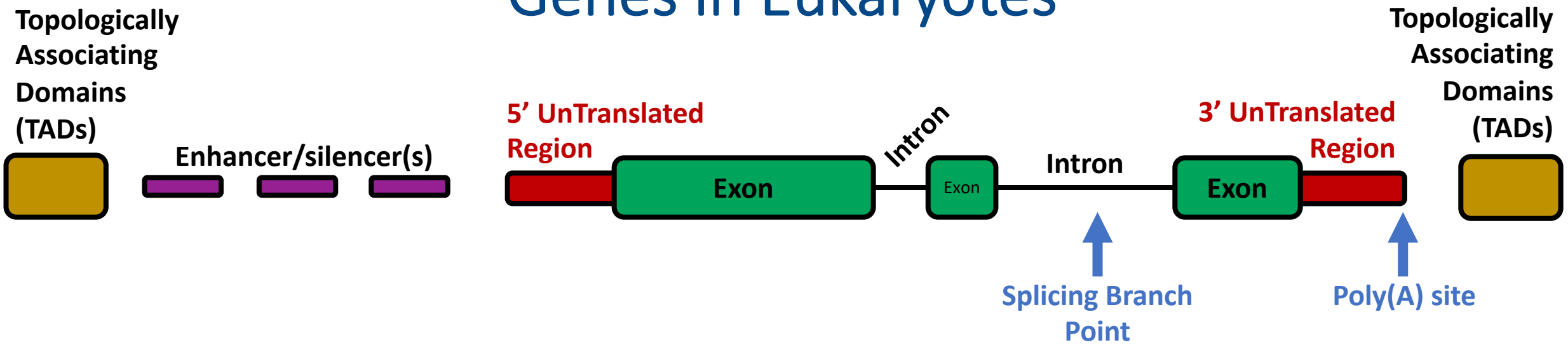
Genes in Eukaryotes



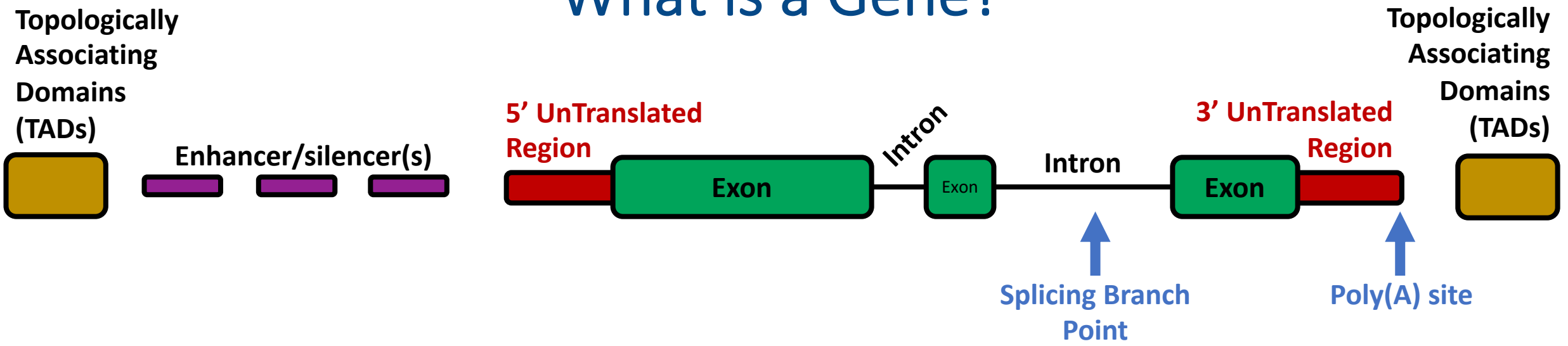
Genes in Eukaryotes

Genes in Eukaryotes

Genes in Eukaryotes

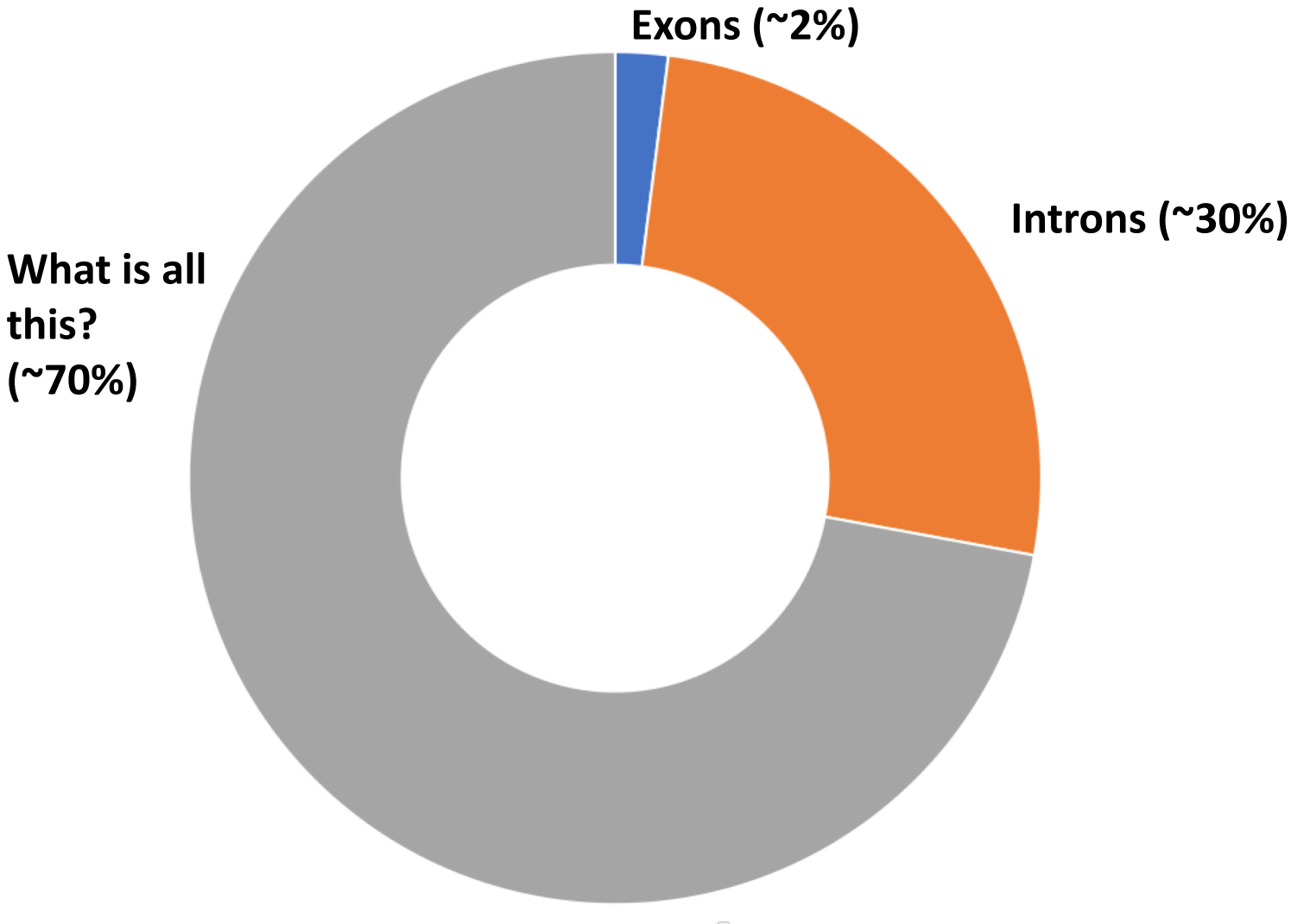


What is a Gene?

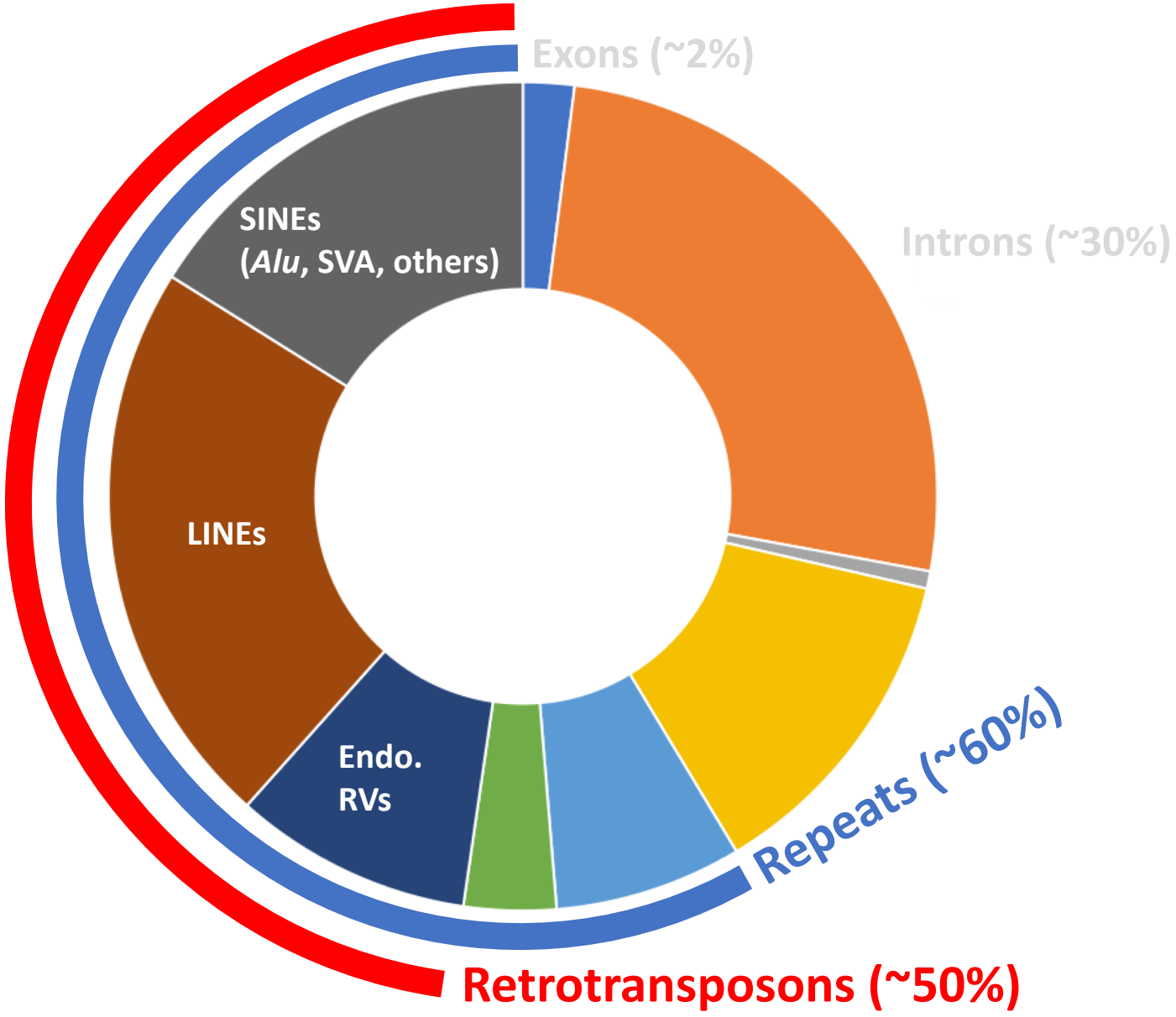


There is no one right answer!

Human Genome Composition



Human Genome Composition



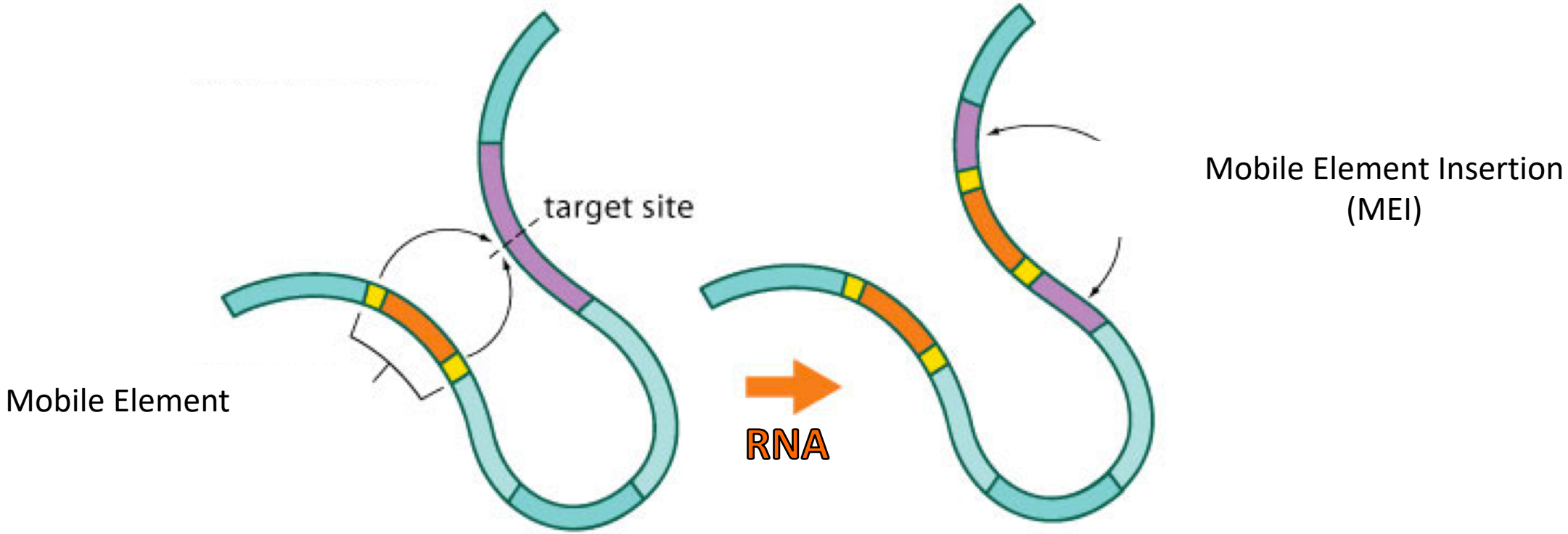
Human Repeat Classes

Simple

Complex

[ATCG] * N	**AAAAAAAA…**	*Homopolymer tract*
[ATCG][ATCG] * N	**ATATATATAT…**	*Low complexity repeat*
([ATCG] * N) * N	**ATGATGATG…**	
([ATCG]* N>2) * N	**TTAGGGTTAGGG…**	*Satellite DNA*
N/A	**GGGCCGGGTGC…**	*Transposons*

Human Retrotransposition



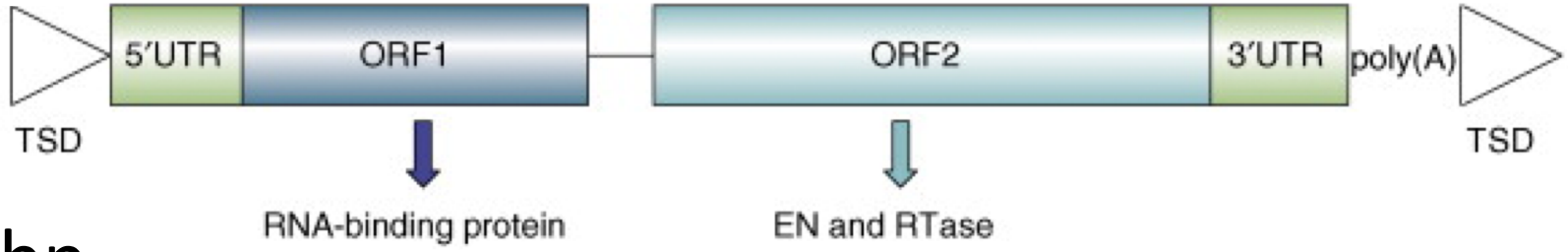
Mobile Element

Mobile Element Insertion (MEI)

Cordaux, R., & Batzer, M. A. *The impact of retrotransposons on human genome evolution.* (2009). *Nat Rev Genet*. PMID: 19763152.
Mills, R. E., Bennett, E. A., Iskow, R. C., et al. *Which transposable elements are active in the human genome?* (2007). *Trends Genet*. PMID: 17331616.

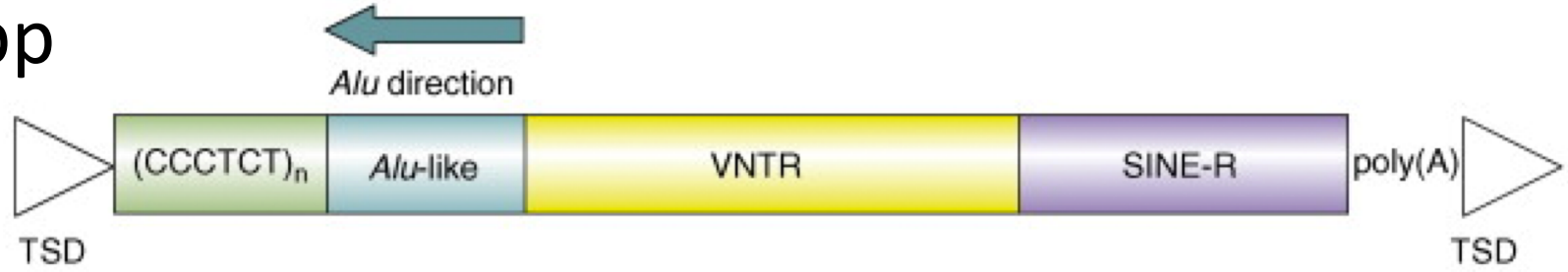
Human Retrotransposition

L1-6kbp



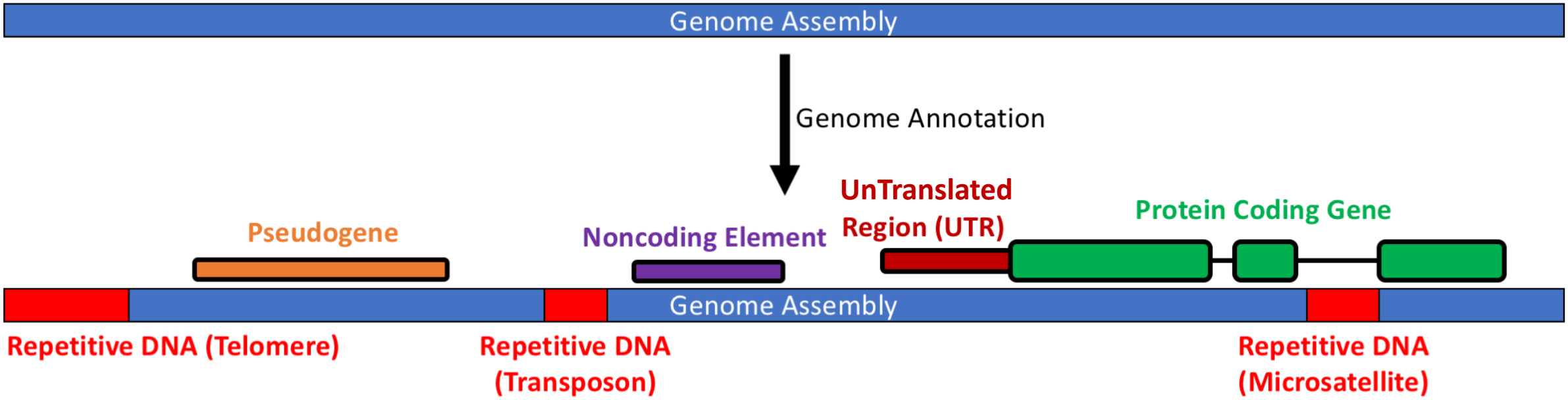
Alu-300bp

SVA-3kbp



Cordaux, R., & Batzer, M. A. *The impact of retrotransposons on human genome evolution.* (2009). *Nat Rev Genet*. PMID: 19763152.
Mills, R. E., Bennett, E. A., Iskow, R. C., et al. *Which transposable elements are active in the human genome?* (2007). *Trends Genet*. PMID: 17331616.

Genome Annotation

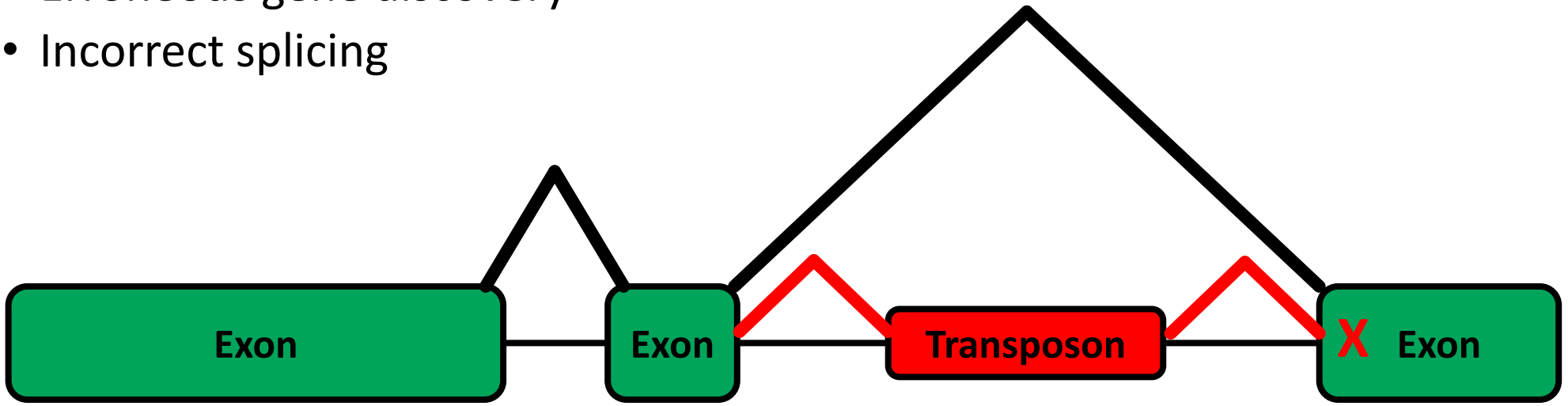


Genome Annotation – Steps

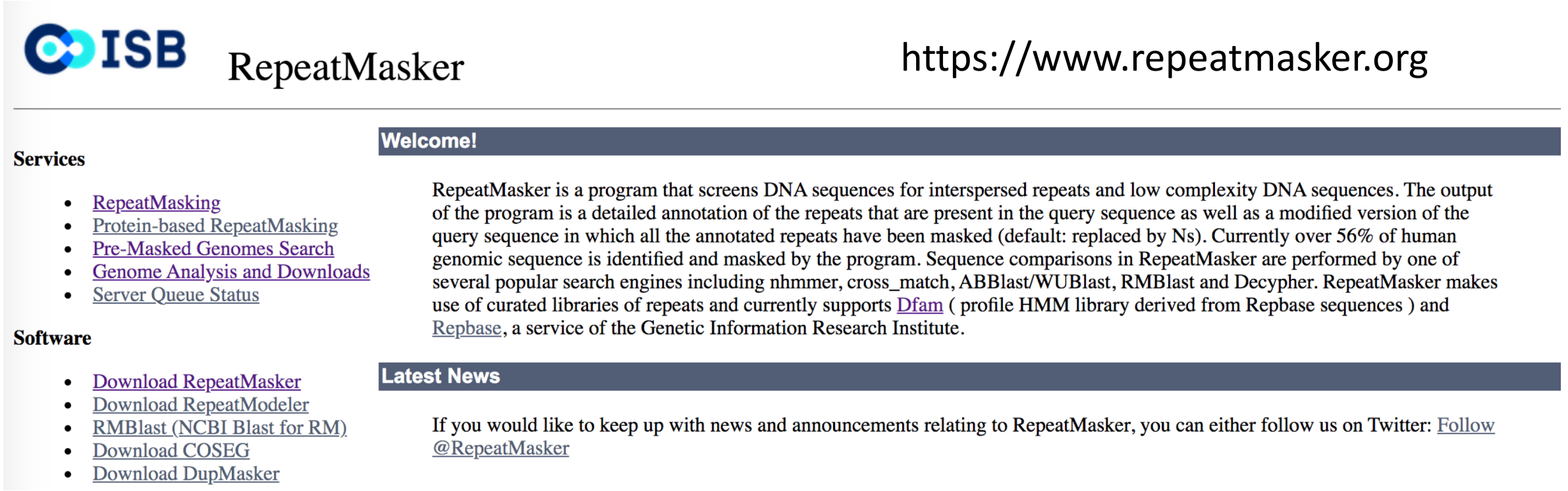
1. Identifying non-coding portions of the genome – ***Repeat Masking***
2. Discovery of functional elements – ***Discovery***
3. Attaching meaning to functional elements – ***Annotation***
4. Review of predictions – ***Manual Curation***

Repeat Masking – Why Are Repeats a Problem for Genome Annotation?

- Transposons can contain ORFs
  - Erroneous gene discovery
  - Incorrect splicing



Repeat Masking – RepeatMasker



- Developed by Arian Smit and colleagues in the 90s
- Used to identify repeat sequences in the human genome project
- Uses a Hidden Markov Model

Repeat Masking – RepeatMasker

TAAGCTTCGTACCCCGAACGATCACAGTTGAAGGCGTTACGGTAACTCTCTCTCTCTCTCTC
TCTCTCTCTCTCATGTTGACGACAATCGAGGGCGGCTATAAGATAAGGGTCATCTATCTA
GAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGGGCACTTGTGTGTGTAAGGCCGAGTC
ATCGATACTTGGTGTTTCGCGTGAGCACTGGAGTACTTGCACAATGCGTGAAGCGACACG
GCGACCATGCCTAGGTACGGCGAGCGCAGACGATATCAAAACGGGCGTTAGGGAACGCCG
TCTGAGACAGACCGGTTTTTTTTTTTTTTTTTTTAGACTTAGCGCTGAGGAACCATTGAC

Library of transposons, satellite sequences, repeats, etc. → **Hidden Markov Model**

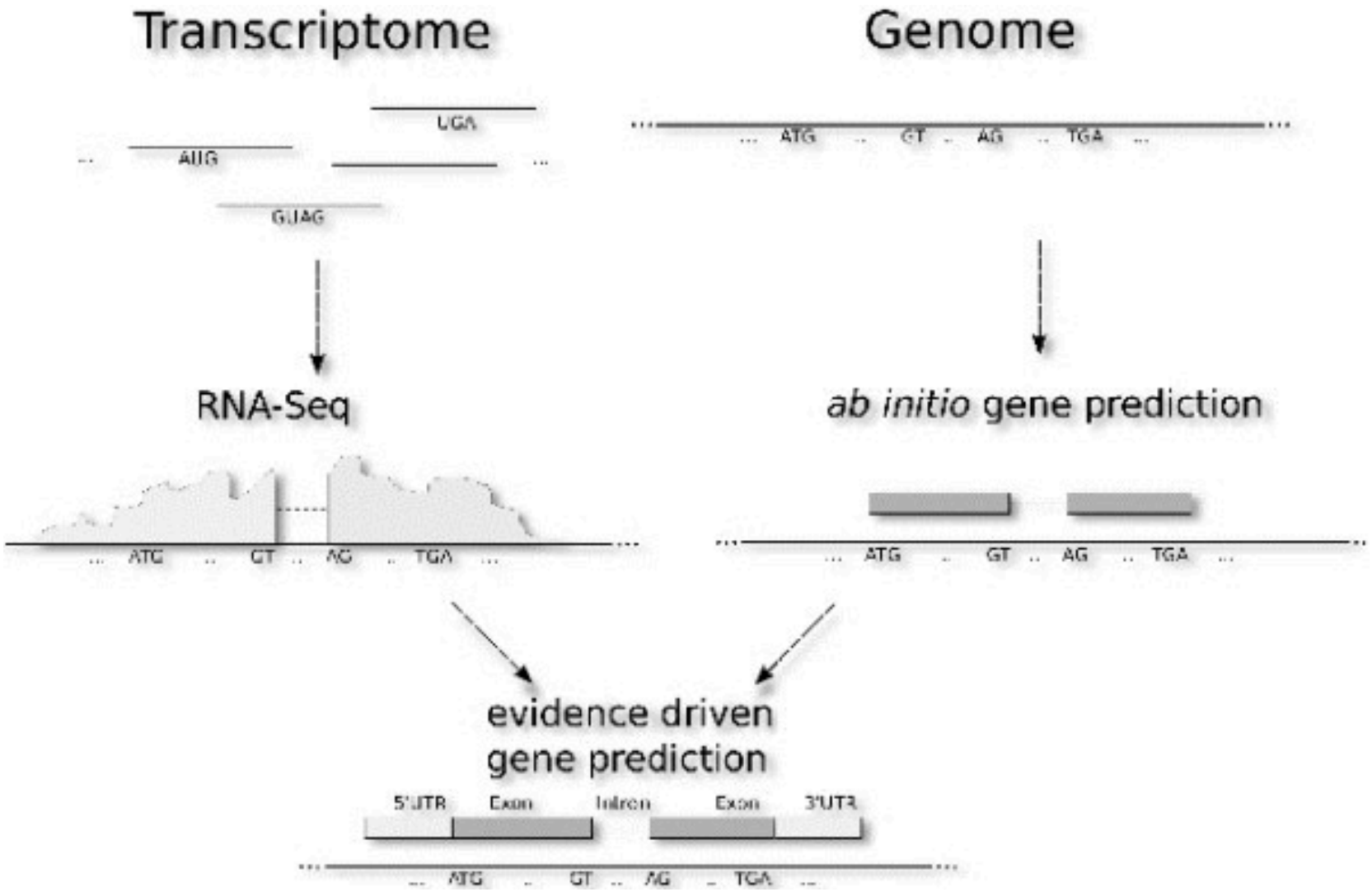
TAAGCTTCGTACCCCGAACGATCACAGTTGAAGGCGTTACGGTAActctctctctctctctc
tctctctctctcATGTTGACGACAATCGAGGGCGGCTATAAGATAAGGGTCATCTATCTA
gaagaagaagaagaagaagaagaagaagaagaaGGGCACTTGTGTGTGTAAGGCCGAGTC
ATCGATACTTGGTGTTTCGCGTGAGCACTGGAGTACTTGCACAATGCGTGAAGCGACACG
GCGACCATGCCTAGGTACGGCGAGCGCAGACGATATCAAAACGGGCGTTAGGGAACGCCG
TCTGAGACAGACCGGttttttttttttttttttAGACTTAGCGCTGAGGAACCATTGAC

Discovery - Types

- *Ab initio* – "from the beginning"
  - Identification of genes purely from the sequence of the genome
  - "Find me all possible open reading frames, sort them out later."

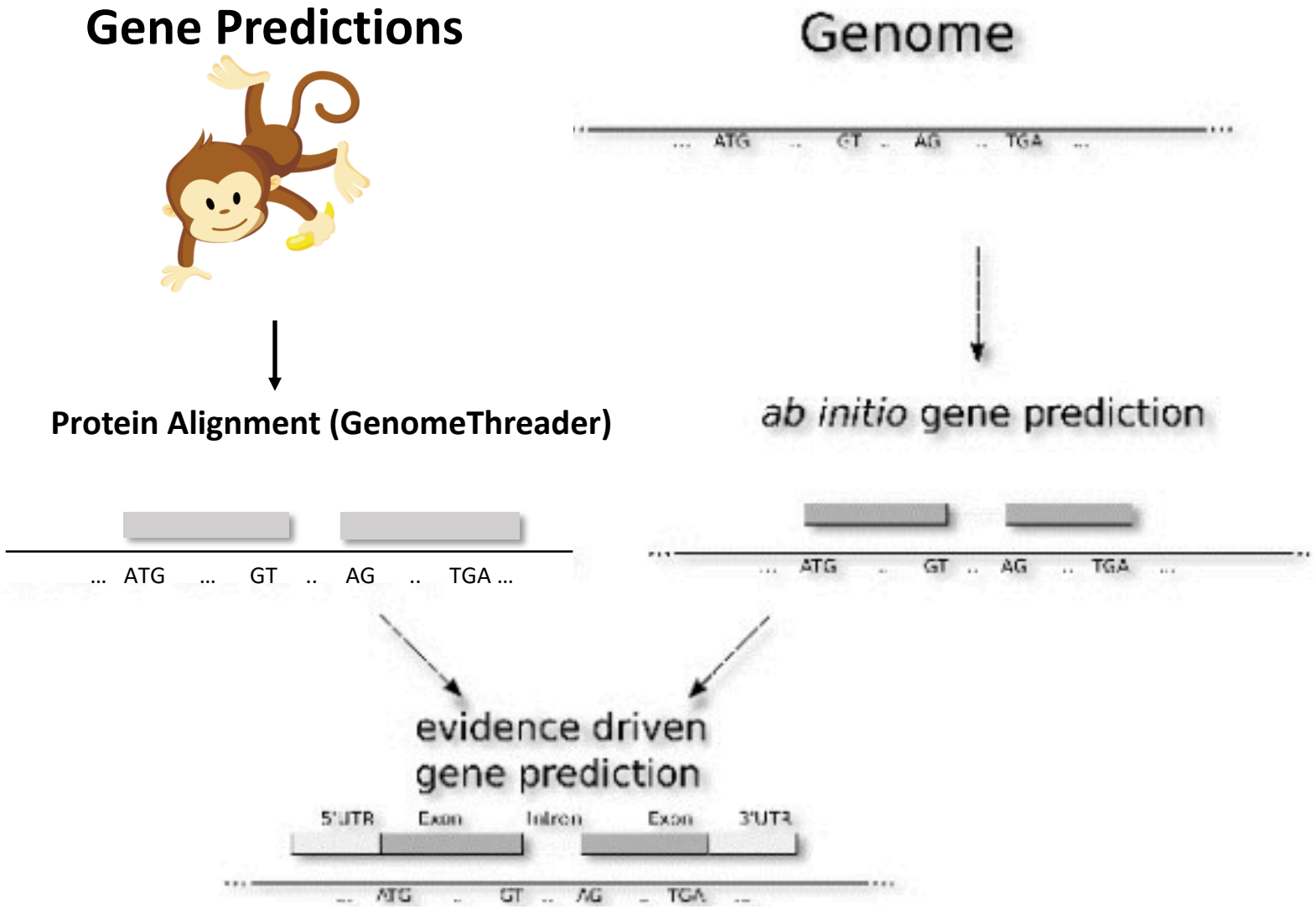
- Evidence based
  - Identification of genes based on some prior evidence (other species, RNA-seq data, cDNAs, etc.)
  - "Find me open reading frames supported by other evidence **AND** those that look very similar."

Discovery – Evidence Based



Müller et al. Comparison of transcriptome technologies in the pathogenic fungus *Aspergillus fumigatus* reveals novel insights into the genome and MpkA dependent gene expression**.** *BMC Genomics* (2012).

Discovery – Evidence Based

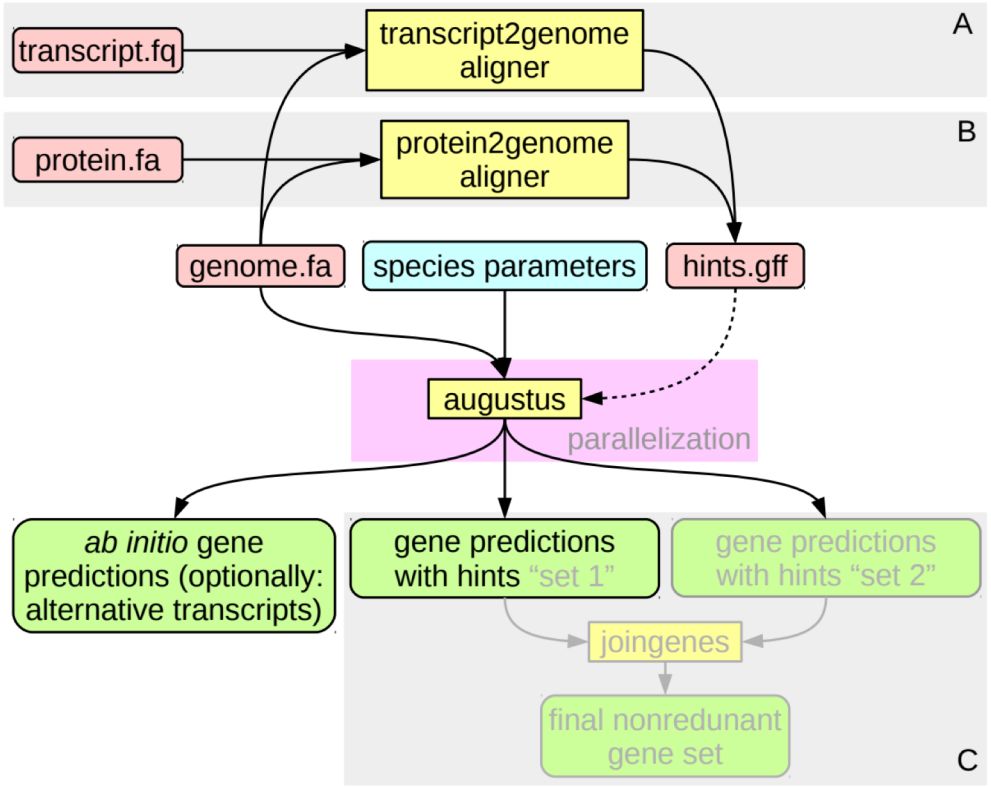


Müller et al. Comparison of transcriptome technologies in the pathogenic fungus *Aspergillus fumigatus* reveals novel insights into the genome and MpkA dependent gene expression**.** *BMC Genomics* (2012).

Discovery – AUGUSTUS + GeneMark-ET + BRAKER

Augustus [gene prediction]

- Performs *ab initio* gene discovery with annotation-based training.
- Actually a combination of two tools:
  - AUGUSTUS: *ab initio* gene discovery
  - GeneMark-ET: RNA-seq for training gene discovery
  - Can be run together with BRAKER OR Web-based AUGUSTUS tool
  - BRAKER2: Automates AUGUSTUS + GeneMark-ET
- Input:
  - .fa file
  - RNA-seq alignment
- Algorithm:
  - Identify likely open-reading frames *ab initio*
  - Evaluate predictions with RNA-seq data
  - Refine prediction parameters
  - Perform evidence-based Discovery
- Output:
  - .gff3 format file of likely genes



Hoff and Stanke. *Predicting Genes in Single Genomes with AUGUSTUS*. Current Prot. Bioinf. (2018).

Discovery – GFF3 Format

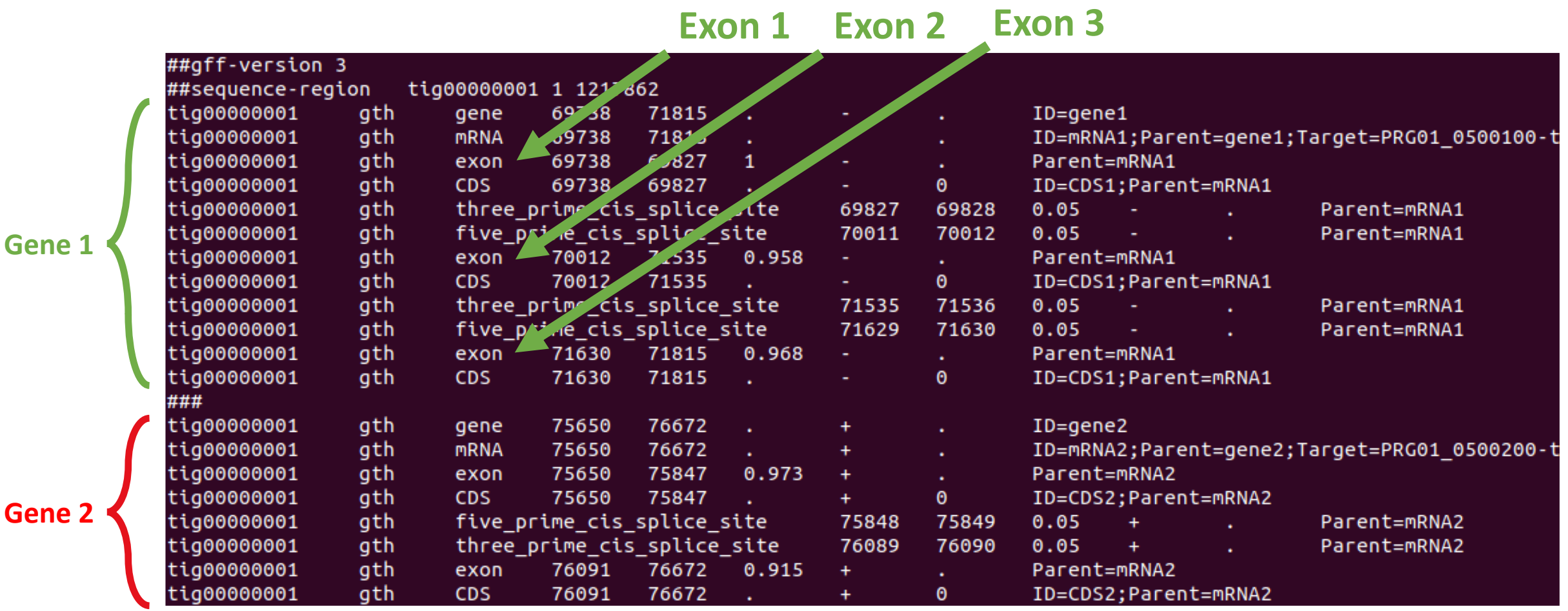
Gene 1

Gene 2

```
##gff-version 3
##sequence-region   tig00000001 1 1217862
tig00000001     gth     gene    69738   71815   .       -       .       ID=gene1
tig00000001     gth     mRNA    69738   71815   .       -       .       ID=mRNA1;Parent=gene1;Target=PRG01_0500100-t
tig00000001     gth     exon    69738   69827   1       -       .       Parent=mRNA1
tig00000001     gth     CDS     69738   69827   .       -       0       ID=CDS1;Parent=mRNA1
tig00000001     gth     three_prime_cis_splice_site     69827   69828   0.05    -       .       Parent=mRNA1
tig00000001     gth     five_prime_cis_splice_site      70011   70012   0.05    -       .       Parent=mRNA1
tig00000001     gth     exon    70012   71535   0.958   -       .       Parent=mRNA1
tig00000001     gth     CDS     70012   71535   .       -       0       ID=CDS1;Parent=mRNA1
tig00000001     gth     three_prime_cis_splice_site     71535   71536   0.05    -       .       Parent=mRNA1
tig00000001     gth     five_prime_cis_splice_site      71629   71630   0.05    -       .       Parent=mRNA1
tig00000001     gth     exon    71630   71815   0.968   -       .       Parent=mRNA1
tig00000001     gth     CDS     71630   71815   .       -       0       ID=CDS1;Parent=mRNA1
###
tig00000001     gth     gene    75650   76672   .       +       .       ID=gene2
tig00000001     gth     mRNA    75650   76672   .       +       .       ID=mRNA2;Parent=gene2;Target=PRG01_0500200-t
tig00000001     gth     exon    75650   75847   0.973   +       .       Parent=mRNA2
tig00000001     gth     CDS     75650   75847   .       +       0       ID=CDS2;Parent=mRNA2
tig00000001     gth     five_prime_cis_splice_site      75848   75849   0.05    +       .       Parent=mRNA2
tig00000001     gth     three_prime_cis_splice_site     76089   76090   0.05    +       .       Parent=mRNA2
tig00000001     gth     exon    76091   76672   0.915   +       .       Parent=mRNA2
tig00000001     gth     CDS     76091   76672   .       +       0       ID=CDS2;Parent=mRNA2
```

Contrast to .bed format which has 1 record per-line

Discovery – GFF3 Format



Discovery – BED12 Format

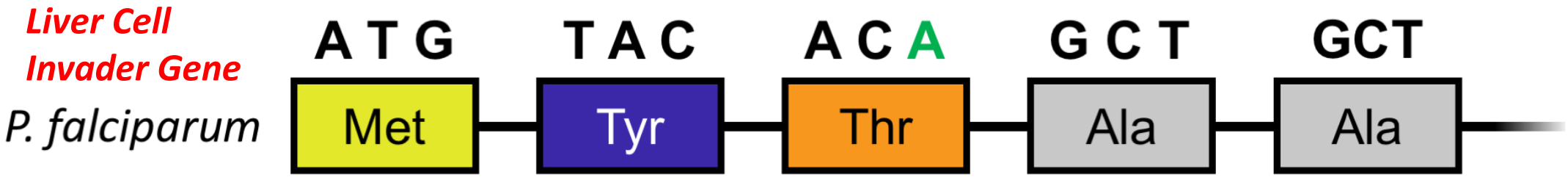
	Gene Start	Gene Stop								Exon sizes	Exon starts	
tig00000001	1215597	1216599	mRNA191	0	+	1215597	1216599	0	2	747,18,	0,984,	Gene 1
tig00000001	1210454	1211108	mRNA190	0	-	1210454	1211108	0	2	28,521,	0,133,	Gene 2
tig00000001	1206223	1208380	mRNA189	0	-	1206223	1208380	0	1	2157,	0,	
tig00000001	1199892	1201503	mRNA188	0	-	1199892	1201503	0	1	1611,	0,	
tig00000001	1197725	1198337	mRNA187	0	-	1197725	1198337	0	1	612,	0,	
tig00000001	1196455	1196770	mRNA186	0	-	1196455	1196770	0	1	315,	0,	
tig00000001	1194409	1196548	mRNA185	0	+	1194409	1196548	0	10	88,113,87,62,213,67,93,94,97,7,	0,199,408,6	
tig00000001	1191202	1193404	mRNA184	0	+	1191202	1193404	0	8	117,276,65,415,95,73,79,98,	0,299,700,9	

Contrast to .gff format where one gene is represented on multiple lines

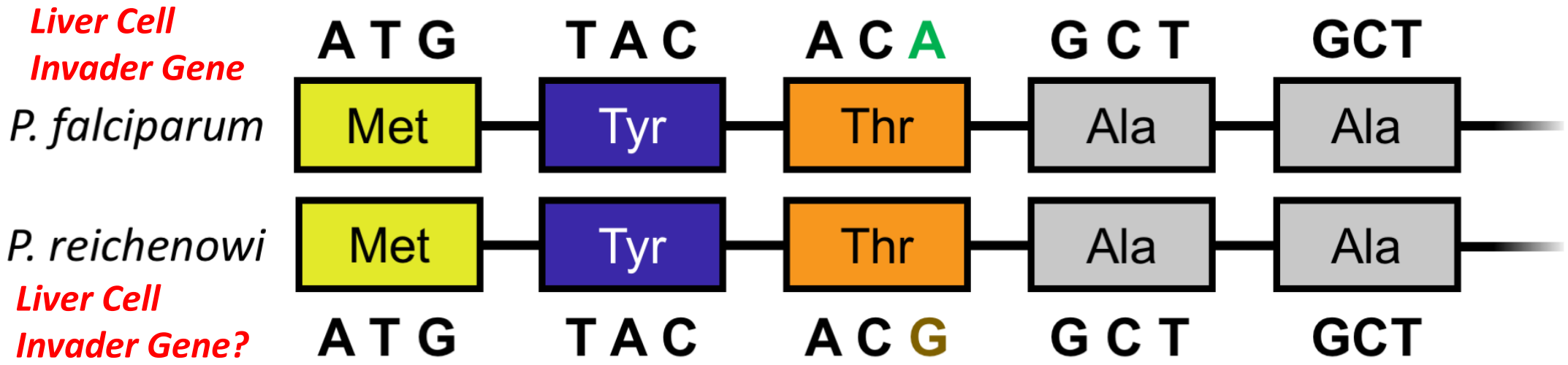
Discovery – Other Annotations

- Pseudogenes – standard gene discovery, screen for degraded ORFs, comparative alignment
- Noncoding RNAs – specialized RNA-sequencing protocols
- Transcription Factor Binding Sites (Promoters/Enhancers/Silencers) – ChIP-seq

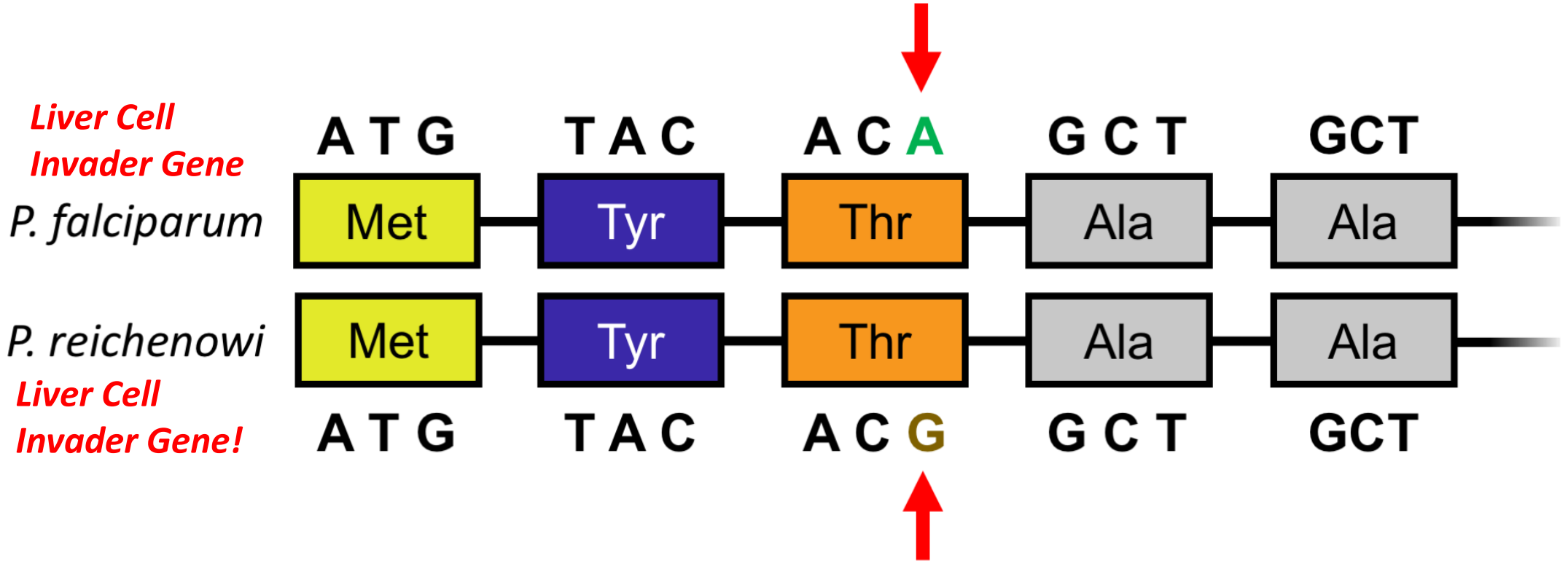
Annotation – Sequence Homology



Annotation – Sequence Homology



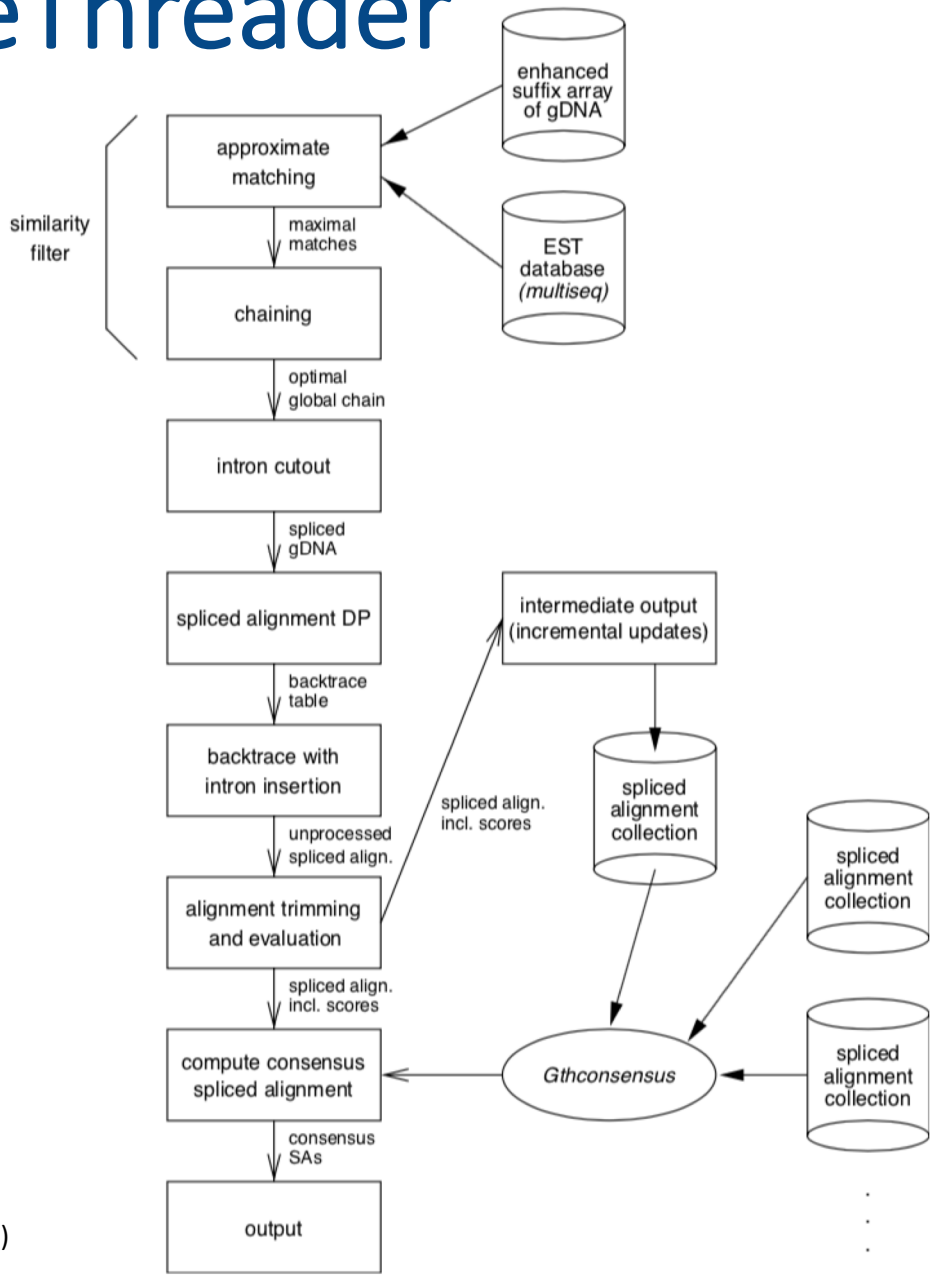
Annotation – Sequence Homology



Alignment of protein sequences while taking into account redundancy in the amino acid code

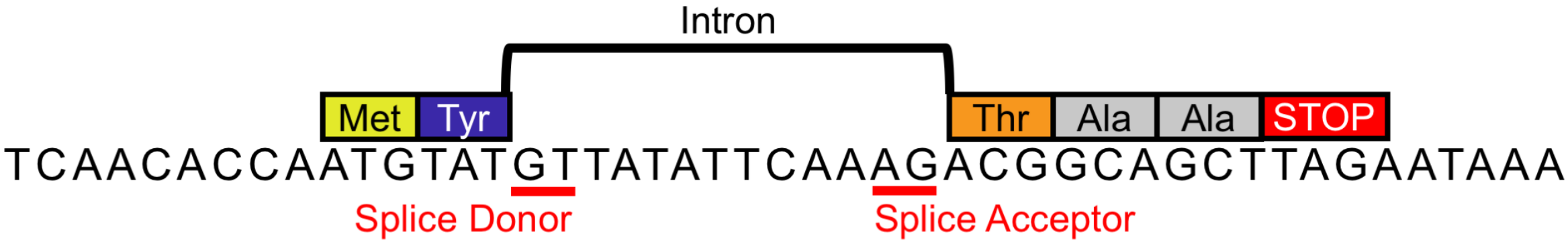
Annotation – GenomeThreader

- Aligns protein sequence to DNA
- Takes into account wobble position
- Input:
  - .fa file of organism being annotation
  - .fa file of protein sequence of already annotated organism
- Algorithm:
  - Approximately match protein → .fa in size N chunks
  - Merge chunks via spliced alignments
- Output:
  - .gff3 format file of likely genes



Gremme et al. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Tech.* (2005)

Annotation – GenomeThreader



Manual Curation

- Visualisation using tools like IGV alongside evidence (RNA-seq)
- Identification of poly-adenylation sites / transcription start sites
- Sequence alignment to identify gene function
- Wet lab work to identify gene function (knockouts, etc.)

havana
human and vertebrate analysis and annotation

How Genome Annotation Works at Scale – *It's Automated*

Darwin TREE of LIFE

wellcome sanger institute

NATURAL HISTORY MUSEUM

EMBL-EBI

Royal Botanic Gardens Kew

Earlham Institute

UNIVERSITY OF CAMBRIDGE

Royal Botanic Garden Edinburgh

THE MARINE BIOLOGICAL ASSOCIATION

UNIVERSITY OF OXFORD

THE UNIVERSITY of EDINBURGH

wellcome sanger institute

Genome Sequencing/Assembly

Genome Annotation

e! Ensembl

Practical Assignment

- Genome Annotation for the *Plasmodium falciparum* chromosome you assembled during the last practical assignment
  - Repeat Elements (masking)
  - Gene Discovery
  - Comparative Genomics