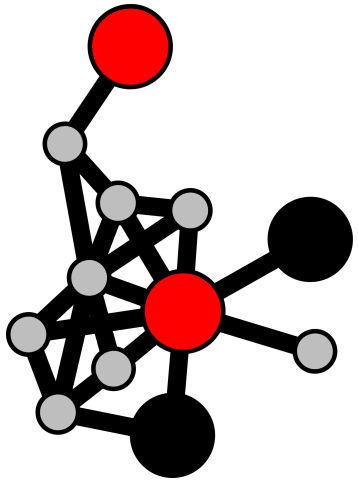
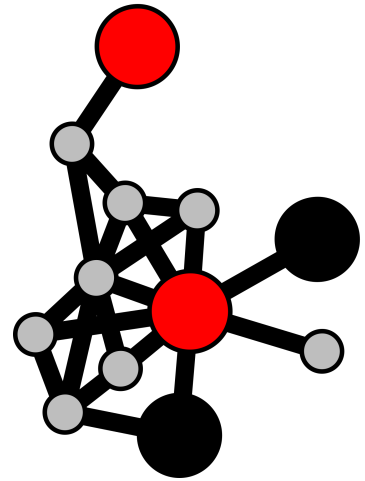
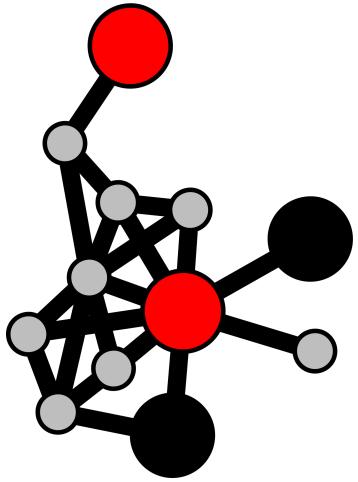


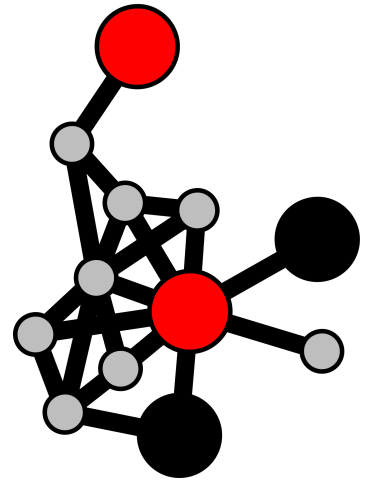


Functional Annotation

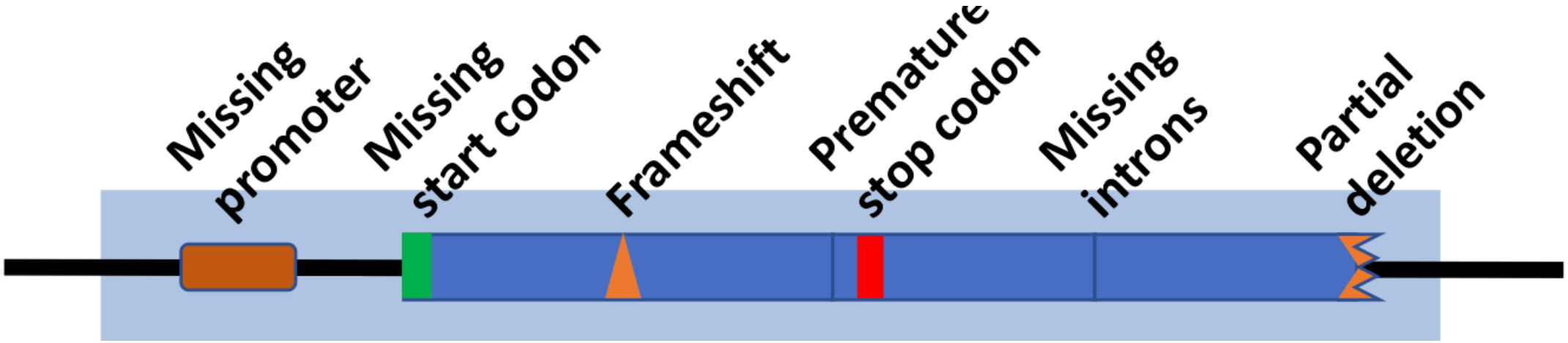
Lesson 8: Pseudogene prediction using Pseudo-Finder

Instructor: Arkadiy Garber



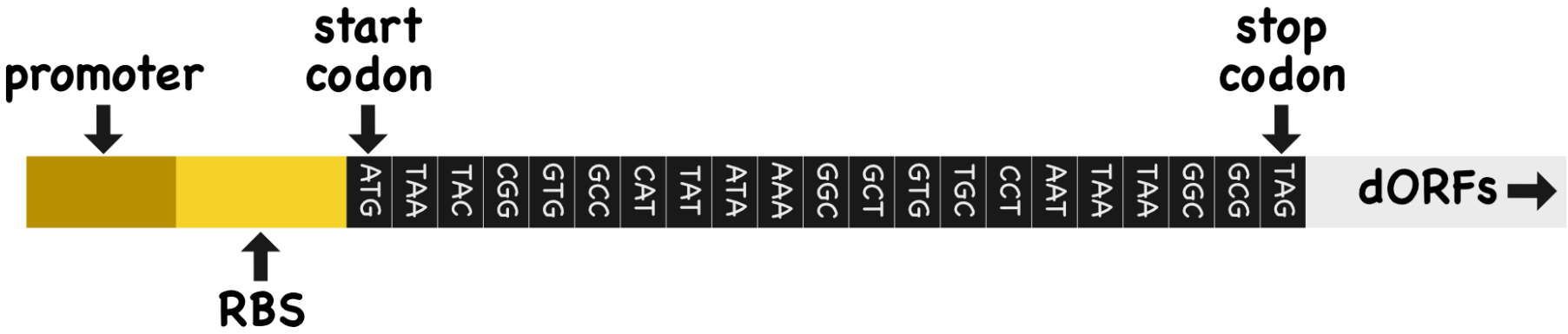


What is a pseudogene?

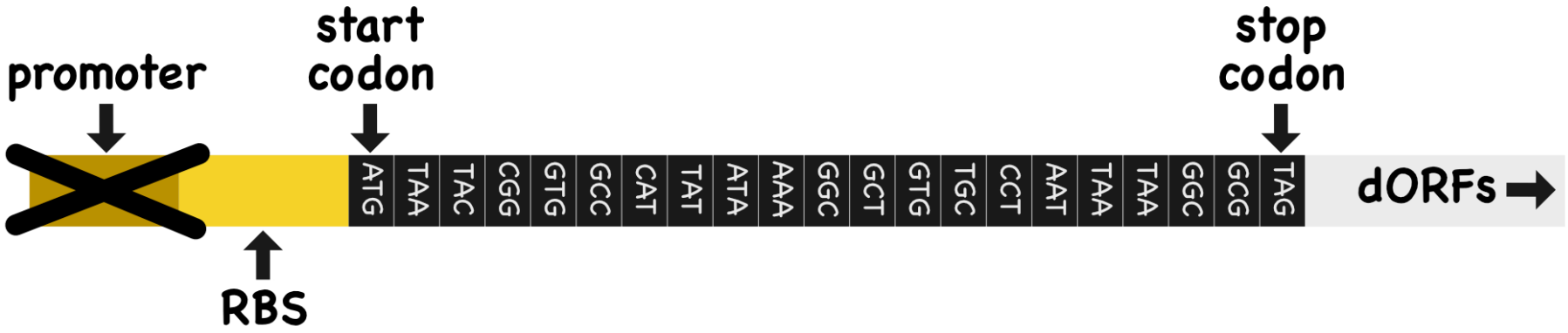


Rosierdfield (*wikipedia*)

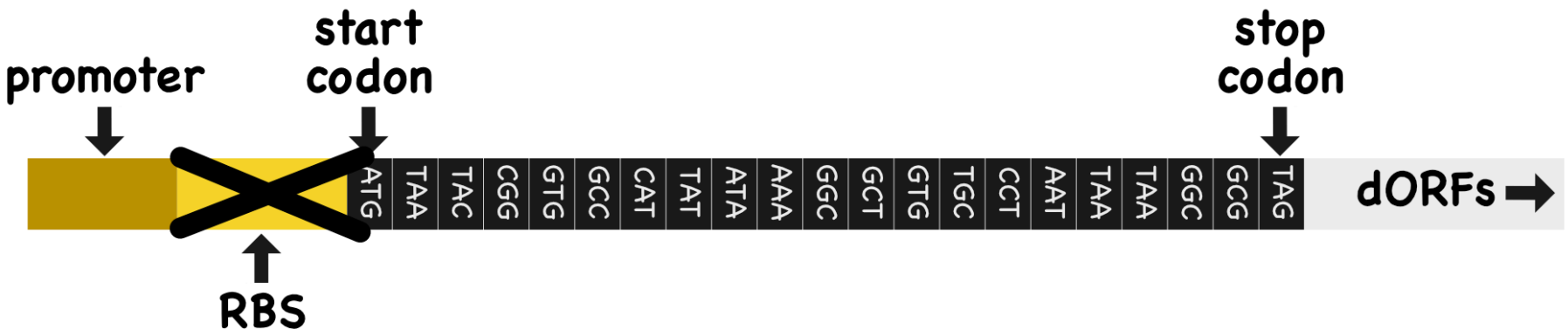
There are several ways a gene can be inactivated, some obvious, and some subtle



There are several ways a gene can be inactivated: loss of promoter

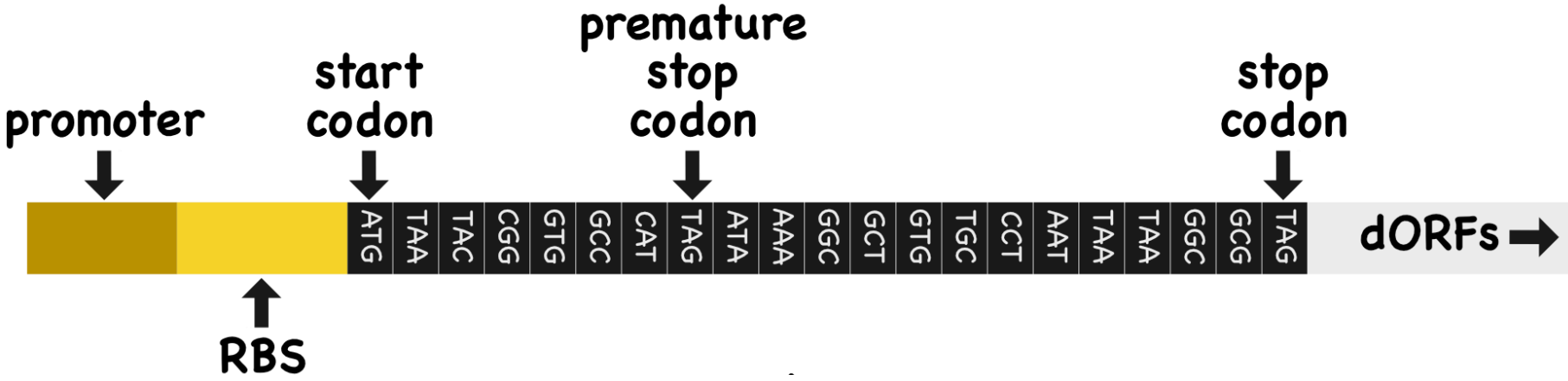


There are several ways a gene can be inactivated: loss of ribosomal binding site

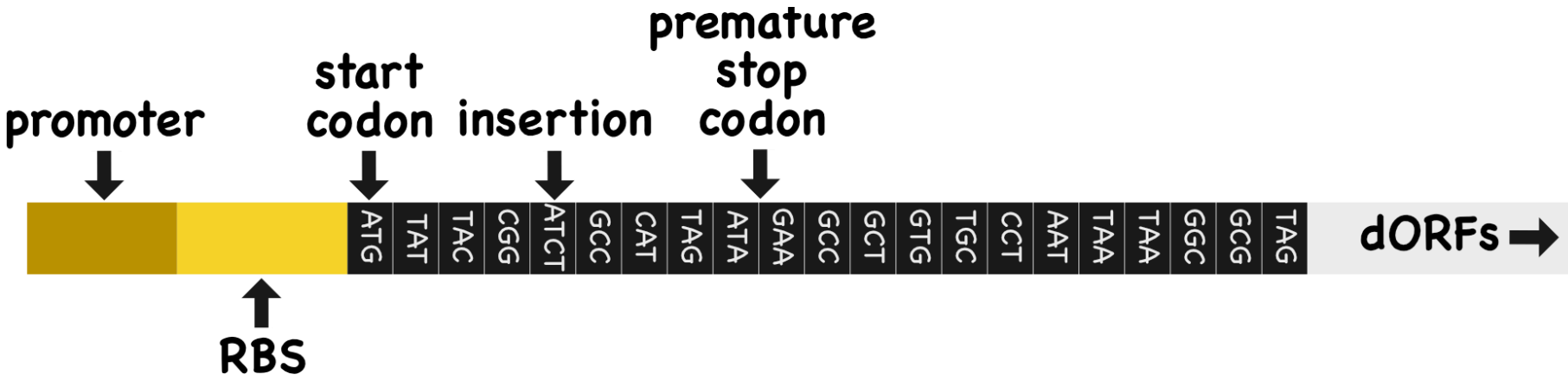


There are several ways a gene can be inactivated: altered start codon

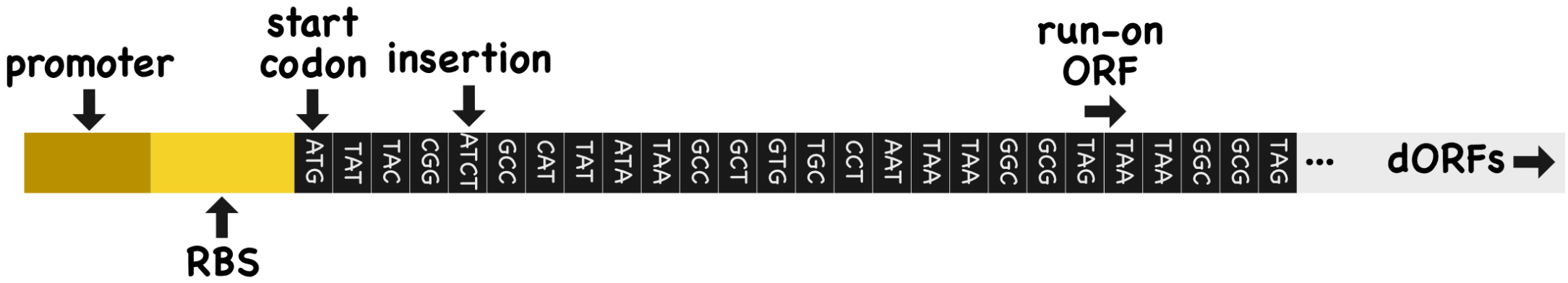
There are several ways a gene can be inactivated: premature stop codon due to missense mutation



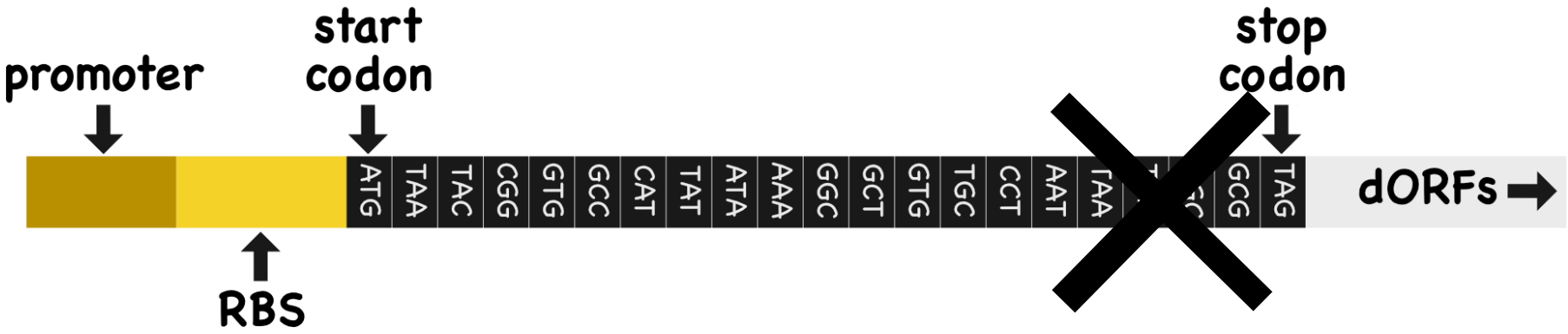
There are several ways a gene can be inactivated: premature stop codon due to indel/frameshift



There are several ways a gene can be inactivated: loss of stop codon due to indel/frameshift



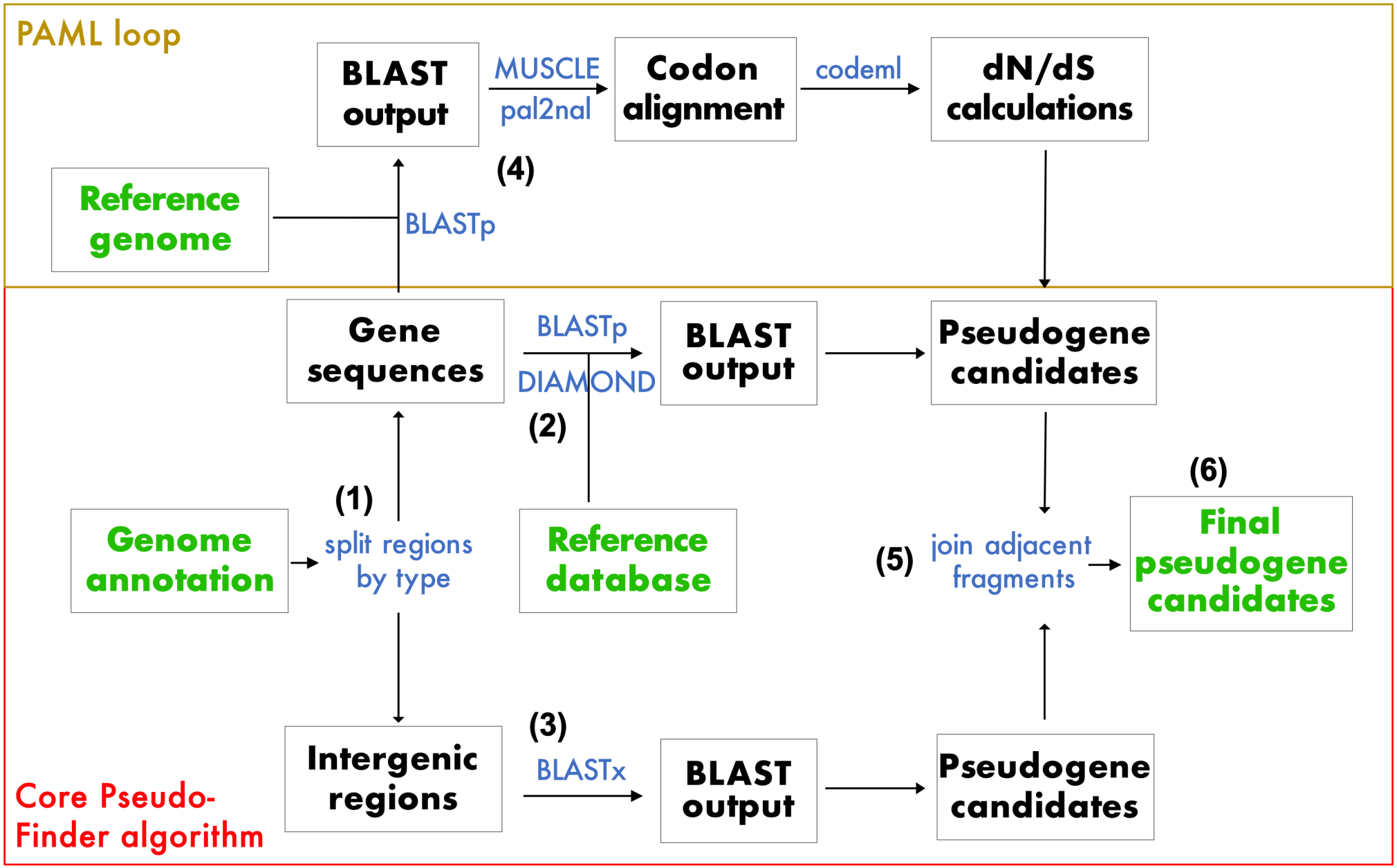
There are several ways a gene can be inactivated: partial deletion



Pseudo-Finder: Reference-based identification of pseudogenes

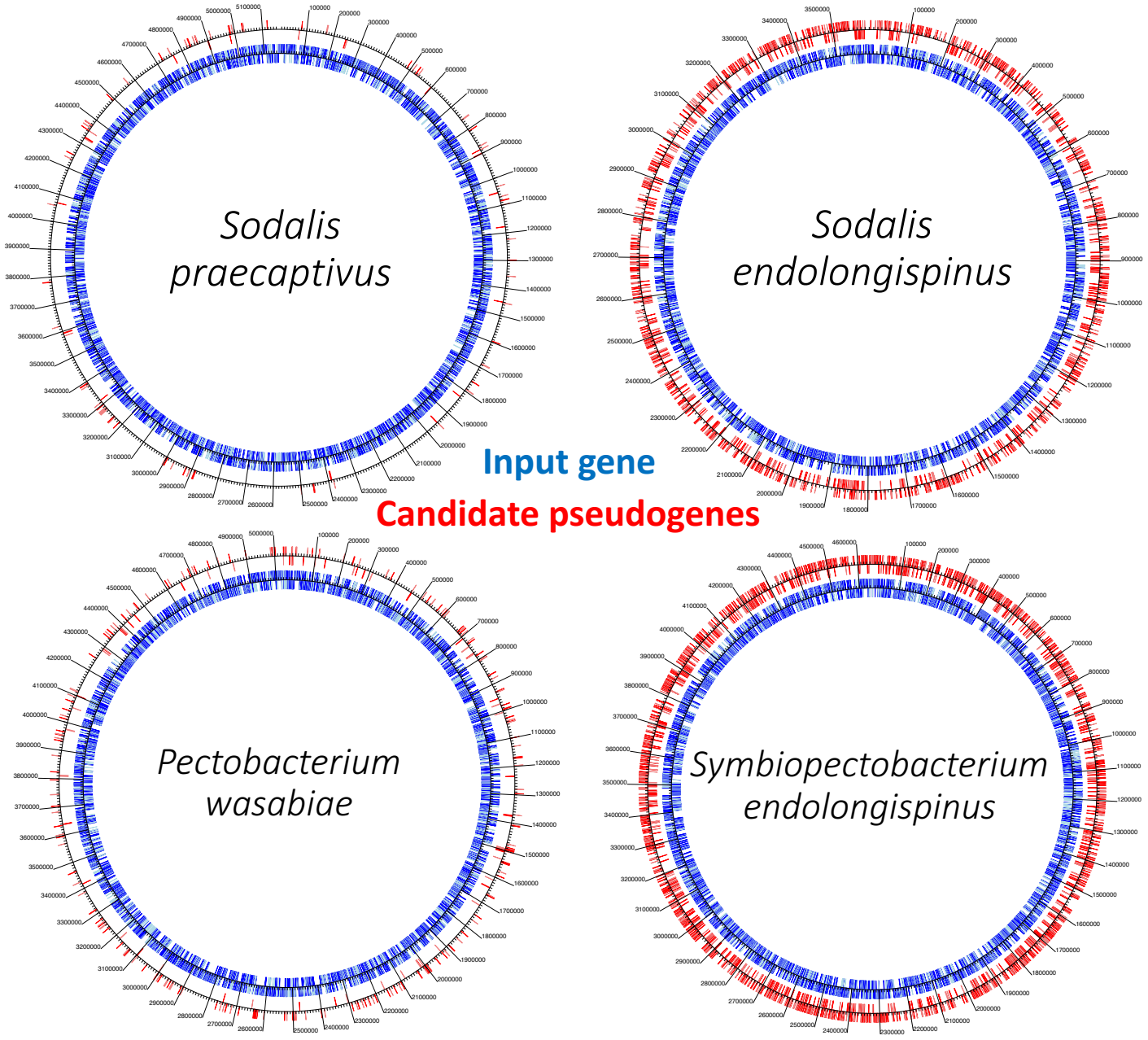
- Length relative to closest homologs from free-living relatives
- Genes fragmented by premature stop codons
- Indels causing frameshifts
- Elevated dN/dS values
  - dS = rate of synonymous mutations/synonymous site (divergence time)
  - dN = rate of nonsynonymous mutations/nonsynonymous site (sequence constraint)
  - Higher dN/dS → neutral or adaptive selection
  - Low dN/dS → purifying selection
- Factors to consider when choosing reference genomes and databases

https://github.com/filip-husnik/pseudo-finder

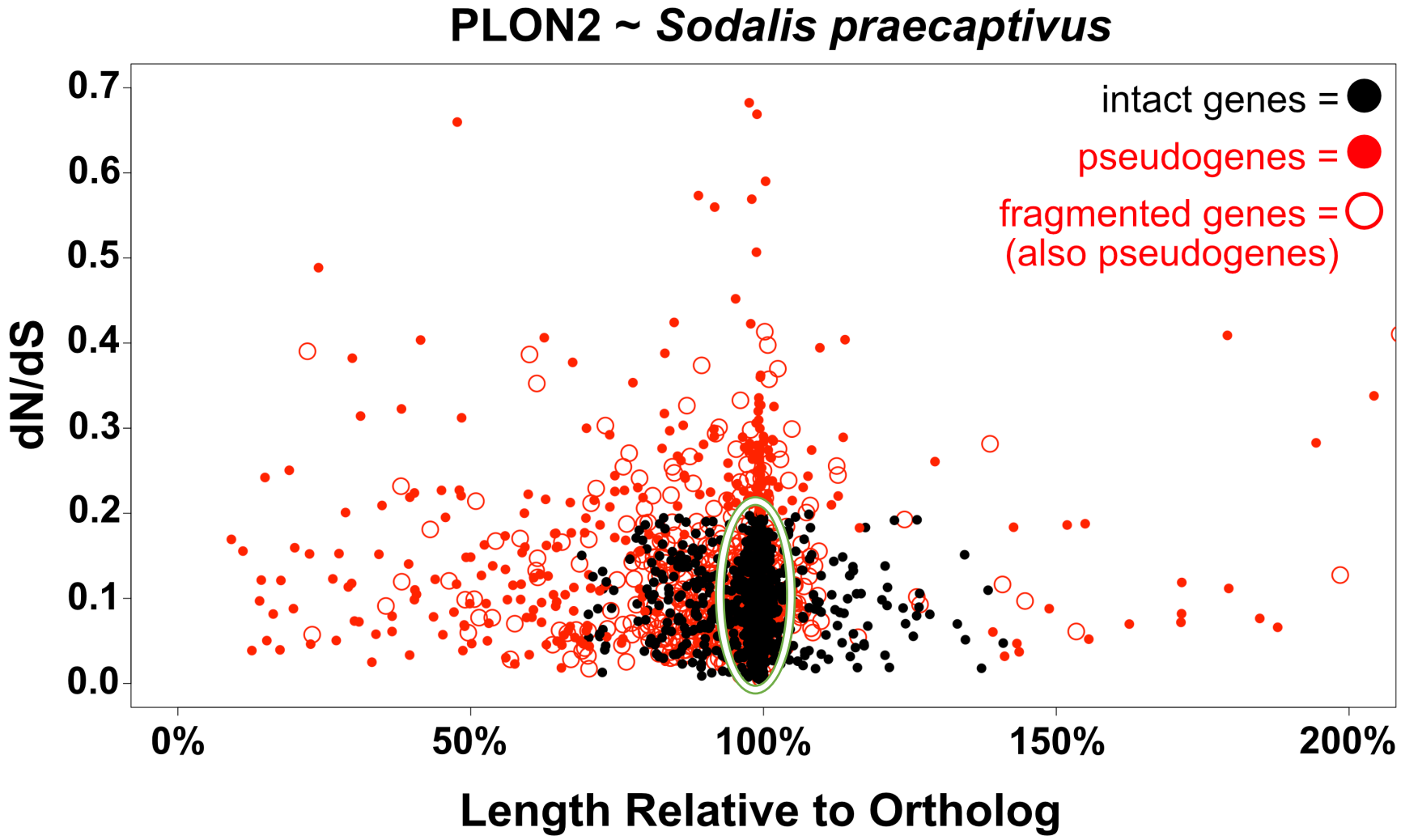

PAML loop
BLAST output
MUSCLE
pal2nal
Codon alignment
codeml
dN/dS calculations
(4)
Reference genome
BLASTp
Gene sequences
BLASTp
DIAMOND
BLAST output
Pseudogene candidates
(2)
(6)
(1)
Genome annotation
split regions by type
Reference database
(5)
join adjacent fragments
Final pseudogene candidates
(3)
Intergenic regions
BLASTx
BLAST output
Pseudogene candidates
Core Pseudo-Finder algorithm
User Input/Output
Intermediate Files
External tools and custom scripts

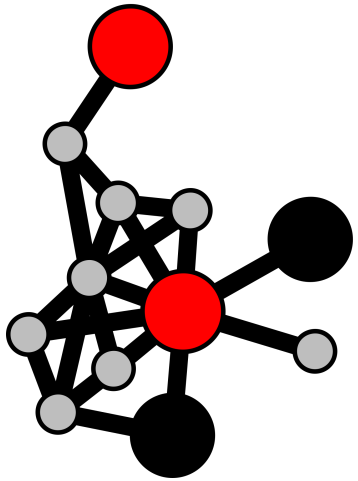
File	Description
[prefix]_functional.gff	Functional genes in GFF3 format.
[prefix]_functional.faa	Functional genes in fasta format.
[prefix]_intergenic.fasta	Intergenic regions in fasta format.
[prefix]_blastX_output.tsv	Tab-delimited output of BLASTX run on intergenic regions.
[prefix]_log.txt	Summary of all inputs, outputs, parameters and results.
[prefix]_map.pdf	Concatenated chromosome map. Input genes appear on the inner track in blue, and candidate pseudogenes are shown in red on the outer track.
[prefix]_proteome.faa	All protein sequences in fasta format.
[prefix]_blastP_output.tsv	Tab-delimited output of BLASTP run on proteome.
[prefix]_pseudos.gff	Candidate pseudogenes in GFF3 format.
[prefix]_pseudos.fasta	Candidate pseudogenes in fasta format.
[prefix]_dnds	Directory containing output from the dnds module: BLAST results, dN/dS summary file, and a folder containing the nucleotide, amino acids, and codon alignments that were used to calculate dN and dS values.

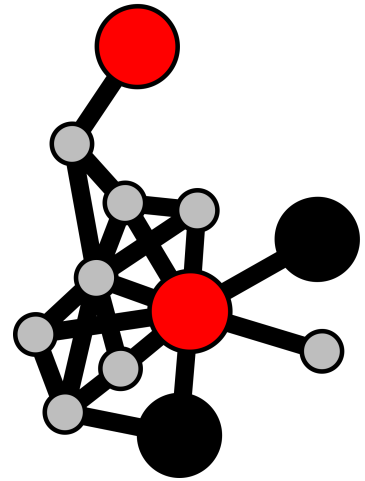
Pseudo-Finder results



PLON1 and PLON2 are both recently-acquired endosymbionts, each showing early and already-extreme genome disruption







Onto the Jupyter Binder Tutorial: Pseudo-Finder