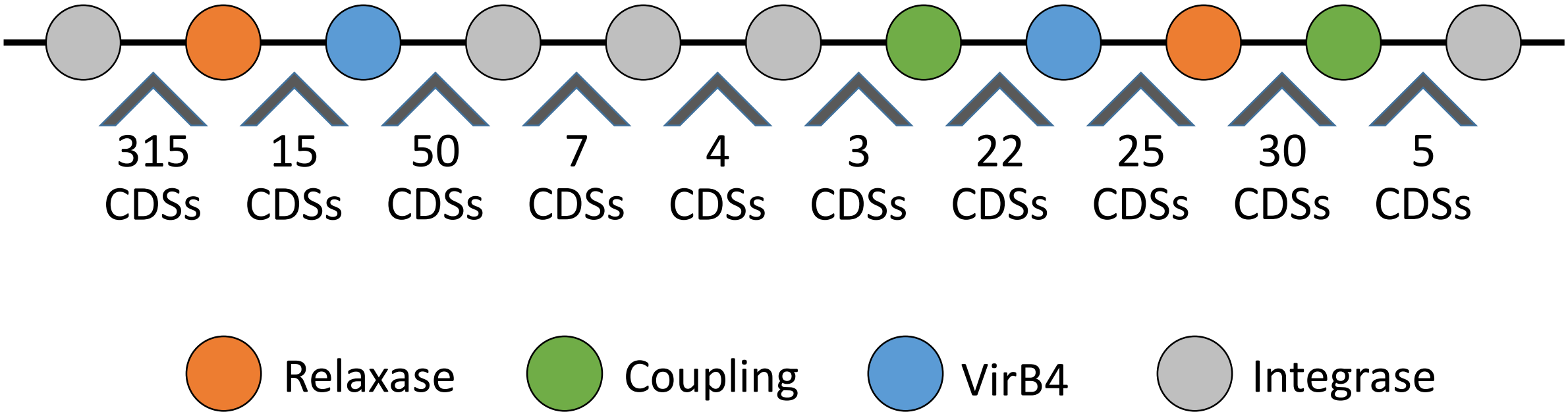
# Algorithm for the detection of ICEs/IMEs structures

1. Find anchors of signature proteins (SPs)
2. extends the anchors
3. eventually merge the anchors

Author : Thomas Lacroix

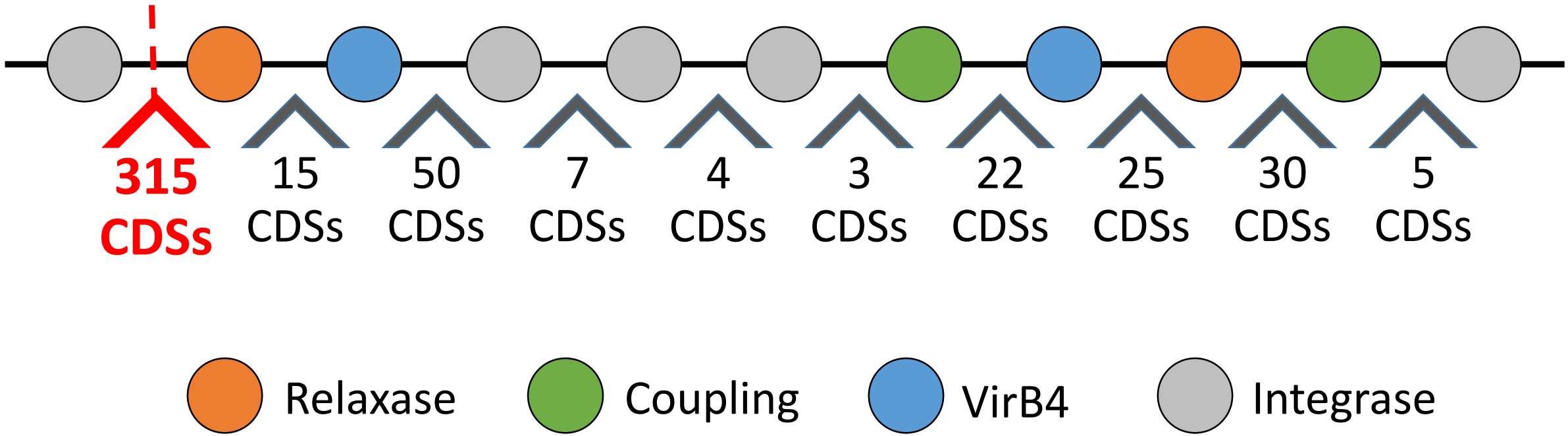
# Input data

- Sequence of signature proteins (SPs) ordered on the genome.



# 1st step: ICEs / IMEs cannot be too large

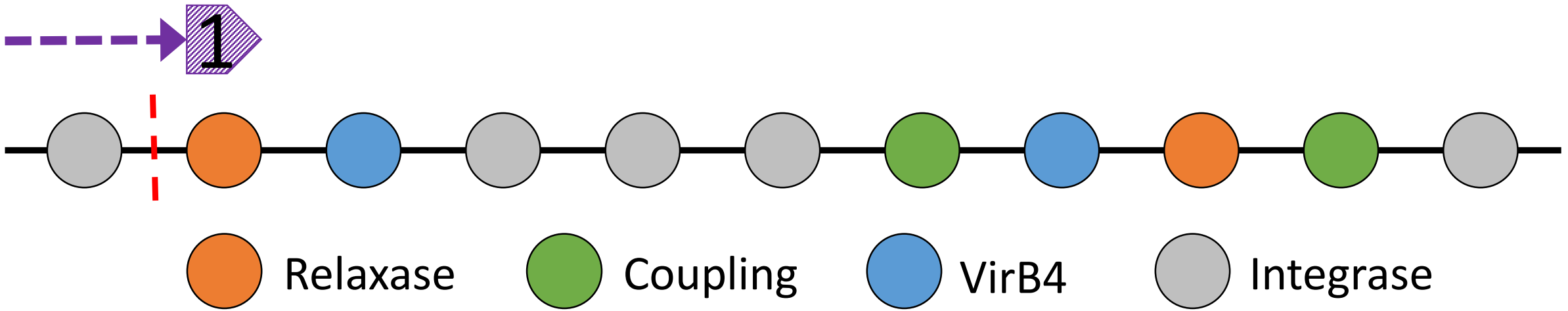
- The sequence is cut in segments if >100 CDSs between 2 successive SPs.



# 2nd step: rules for creating an anchor

The sequence is scanned from left to right (→). When either one of the 3 SPs relaxase, coupling, or virB4 is found → the anchor starts ( 1 ).

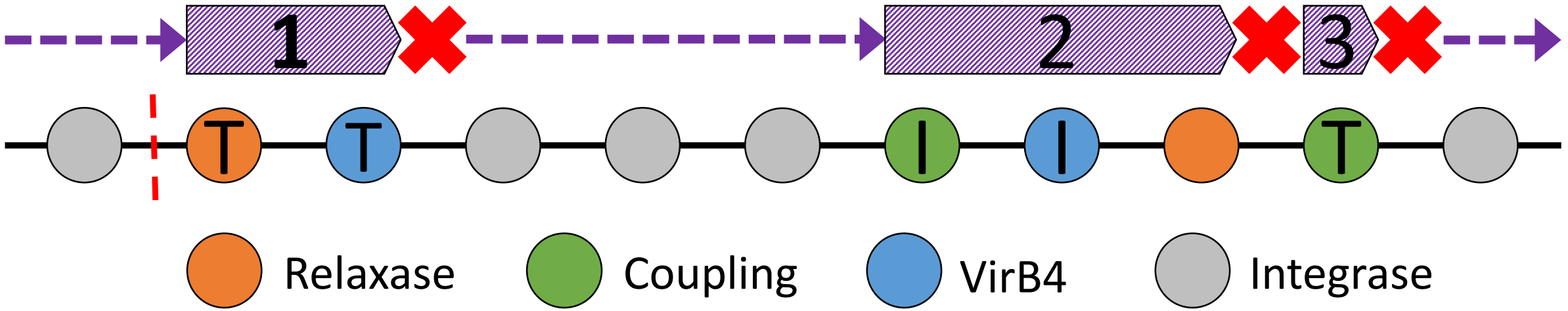
- Relaxase, coupling, and virB4 are quite specific of ICEs / IMEs structures if they are found in combination within a short genomic region.
- Integrases are less specific of ICEs / IMEs structures as they may also relate to other mobile elements (i.e. prophages for Tyr or Ser, transposons or IS for DDE). Integrases are always at the border of the mobile element.



# 3rd step: rules for extending an anchor (1/2)

The sequence continue to be scanned from left to right. An ICE / IME anchor cannot contain (conditions for stopping the extension) :

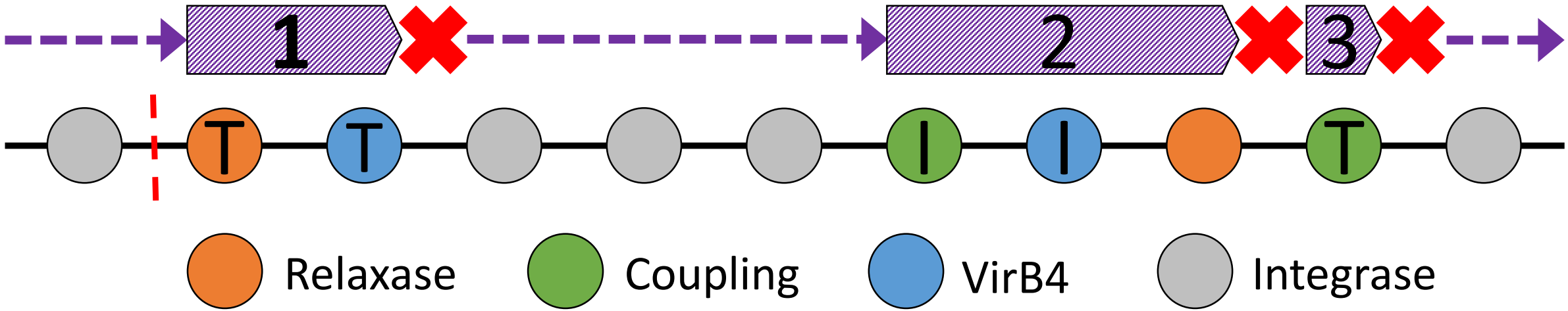
- 2 SPs separated from more than 100 CDSs (step 1).
- 2 virB4 or 2 coupling
- 2 relaxase unless they are adjacent on the genome or separated by one CDS.



# 3rd step: rules for extending an anchor (2/2)

An ICE / IME anchor cannot contain (conditions for stopping the extension) :

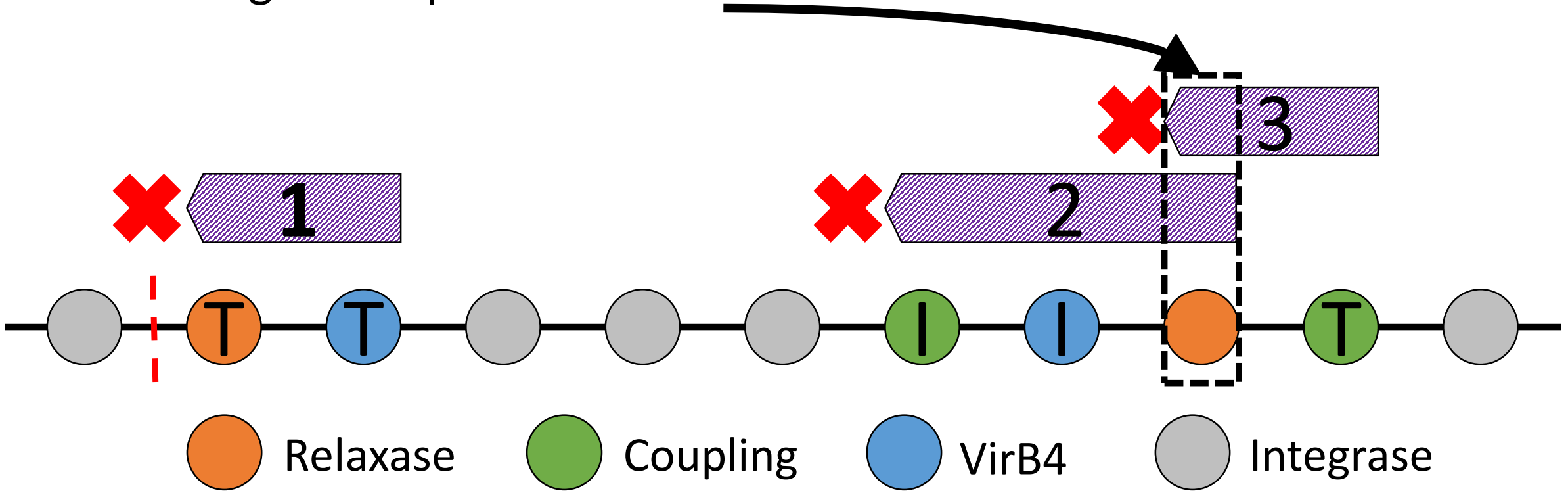
- SPs of different families (i.e. **I** = ICESt3, **T** = Tn916). Families of ICEs and IMEs are curated known elements in *Streptococcus*. BlastP hits of the same family are preferably grouped within an anchor while BlastP hits of different families are separated. SPs without any family information (i.e. HMM hits) can be added to an anchor regardless of the family criterion.
- Integrase (will be dealt with subsequently).



# 4th step: extending anchors from right to left

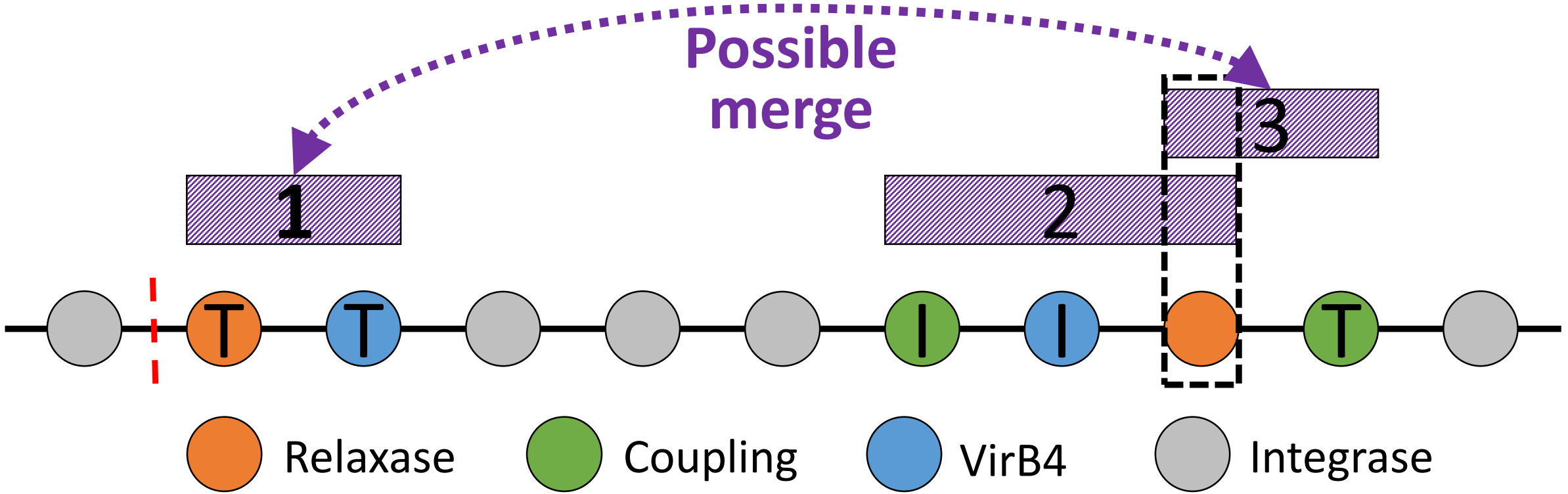
After the 3rd step (creation and extension of anchors from left to right), each anchor is extended from right to left (same stopping conditions).

- ICEs / IMEs have no direction.
- Algorithm consistent and independent of the choice of scanning direction.
- Possible signature protein in "conflict" attached to 2 different anchors.



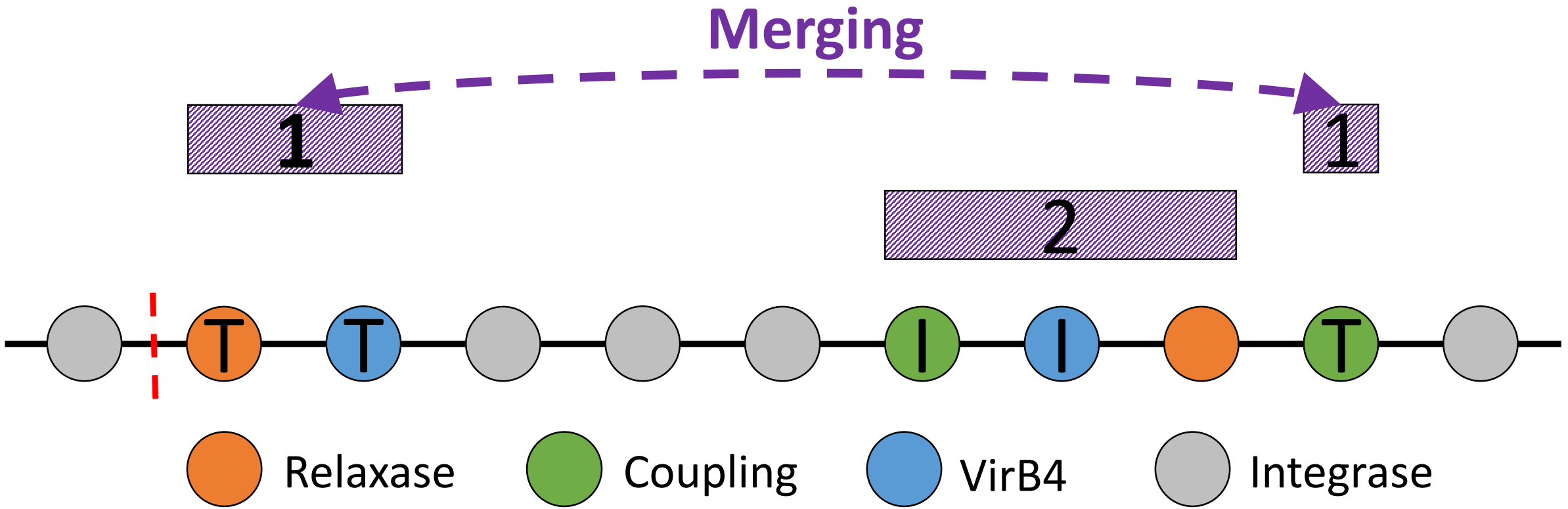
# 5th step: merging of anchors (1/2)

- Exhaustive: all combinations of merging are tested. The priority is given to the merging of the nearest anchors if there is multiple possibilities.
- Recursive: detection of cases with multiple levels of nesting and/or when the ICEs / IMEs are "split apart" in more than 2 pieces (rare case).
- The rules for merging are identical to the rules for extending an anchor.



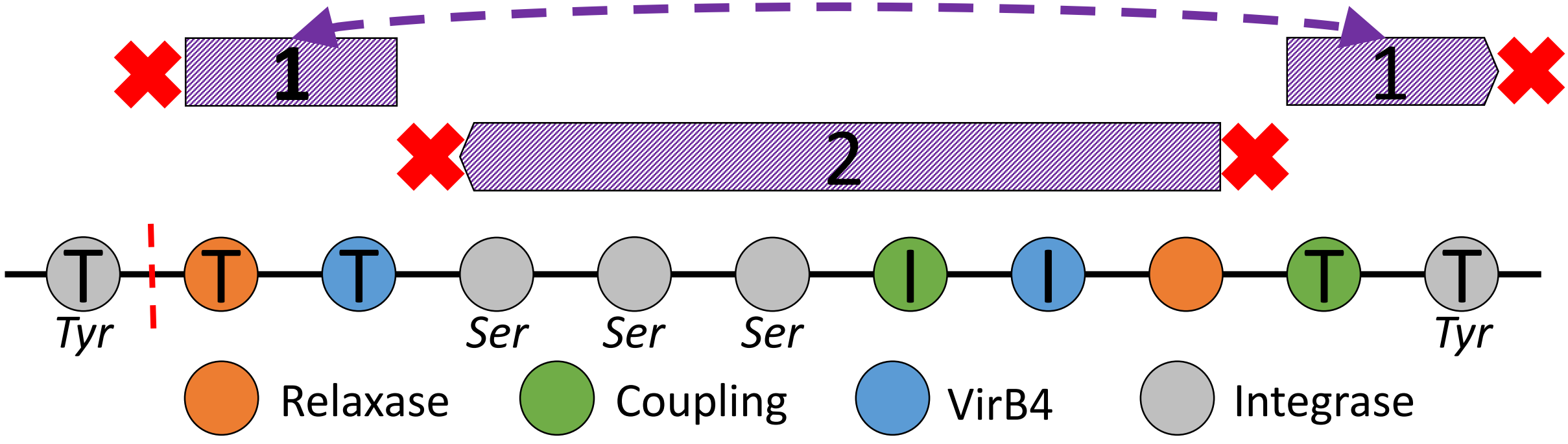
# 5th step: merging of anchors (2/2)

- This step can help resolve SPs in “conflict” (attached to 2 different anchors).



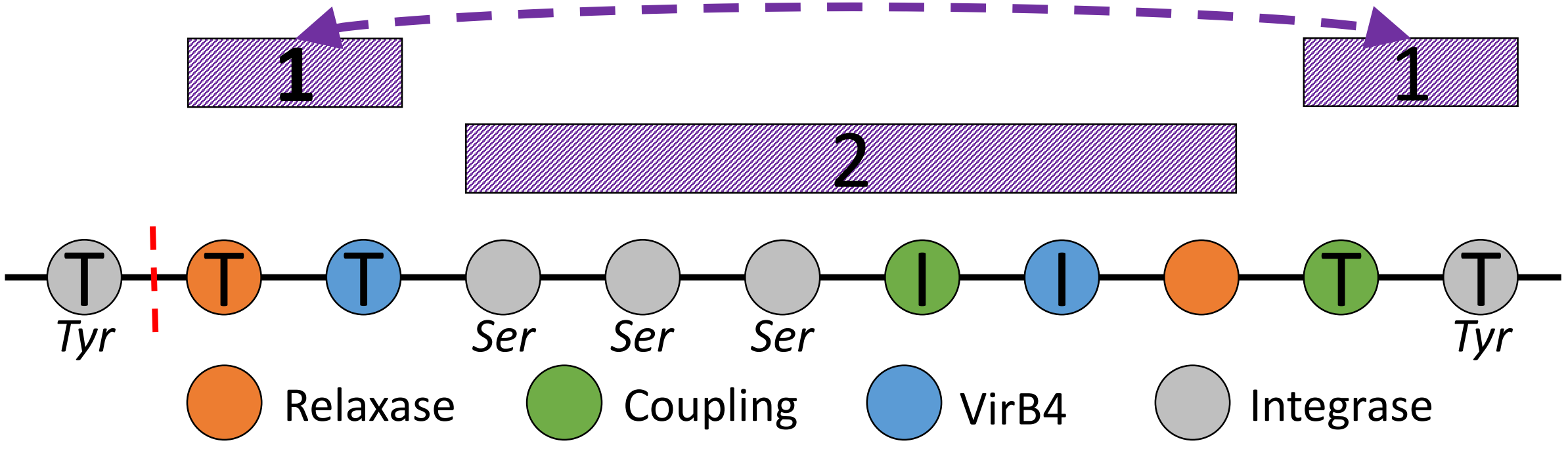
# 6th step: rules for adding an integrase to an anchor

- Integrase can be up or downstream within the 100 CDS limit (step 1).
- Integrase adjacent to the anchor makes good candidate but there can be distant integrase if nested ICEs/IMEs
- Integrase previously seen associated to an ICE/IME family (i.e. I = ICESt3, T = Tn916) makes good candidate.



# Special cases regarding the integrase

- Adjacent trio or duo of integrases Ser (may be separated by a CDS).
- Upstream ICE → integrase strand - ; downstream ICE → strand +.
- The algorithm may not be able to choose between an upstream or a downstream integrase, or between a distant integrase previously seen associated to an ICE/IME family and an adjacent one.



# 7th step: classification of different types of ICEs / IMEs

Complete, partial, to be verified experimentally, nested, etc. :

- Complete ICE: R+C+V+I
- Conjugation module: R+C+V
- Partial ICE: V + other signature proteins
- Complete IME: R+I or R+C+I with distance < 10 CDS
- Mobilizable element: R+C with distance < 10 CDS
- Other partial element: R+C>10 CDS, R+V, V+C

# Test sets of 89 ICEs / IMEs

Manually adapted from real cases to test the algorithm on a variety of complex cases:

- Signatures proteins : 356
- Complete ICEs: 23
- Conjugation modules: 8
- Partial ICEs: 11
- Complete IMEs: 37
- Mobilizable elements (R+C < 10 CDS) : 3
- Other partial elements (R+C>10 CDS, R+V, V+C) : 7
- Nested elements: 47