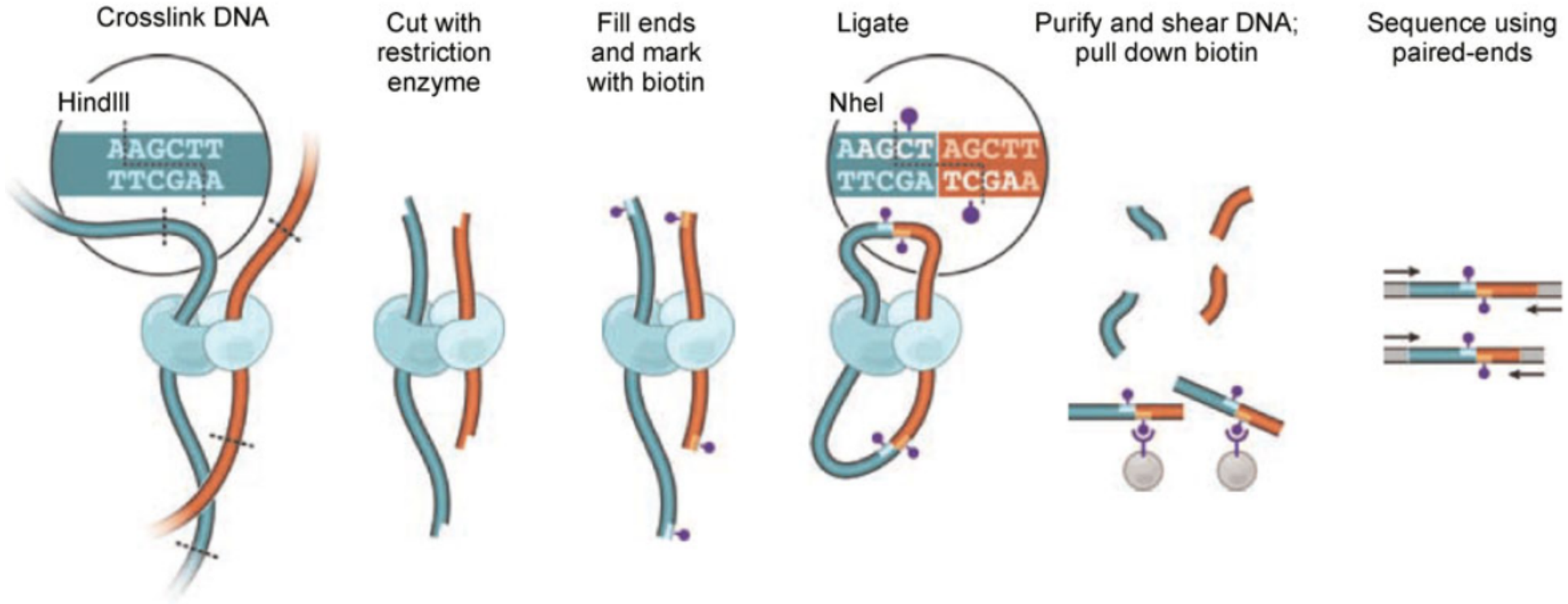
TADbit

Hi-C contact matrices: filtering

Marco Di Stefano, David Castillo & Marc A. Marti-Renom
Structural Genomics Group (CNAG-CRG)

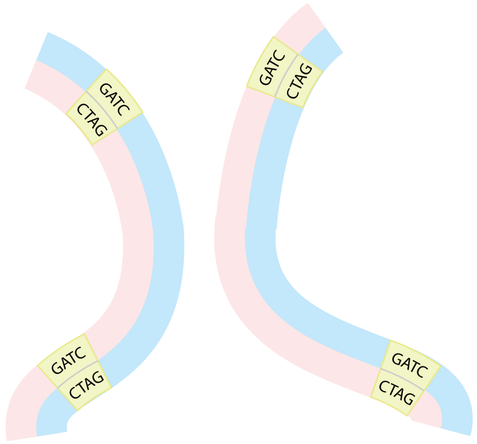
cnag CRG Centre for Genomic Regulation

Hi-C experiment

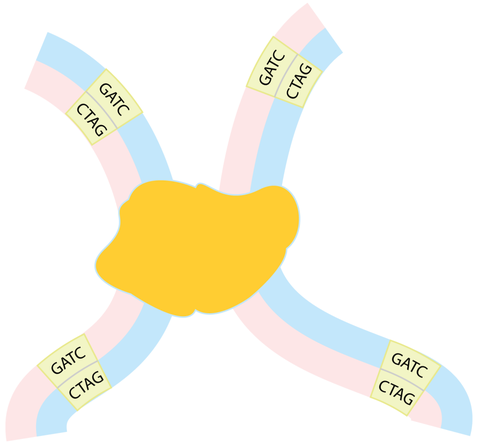


Lieberman-Aiden, E. … Dekker, J. (2009). Science, 326(5950),

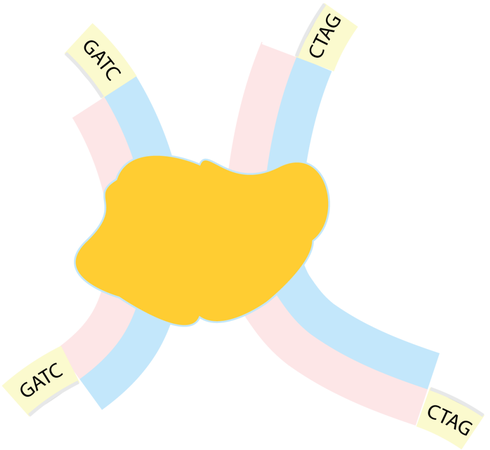
Hi-C experiment



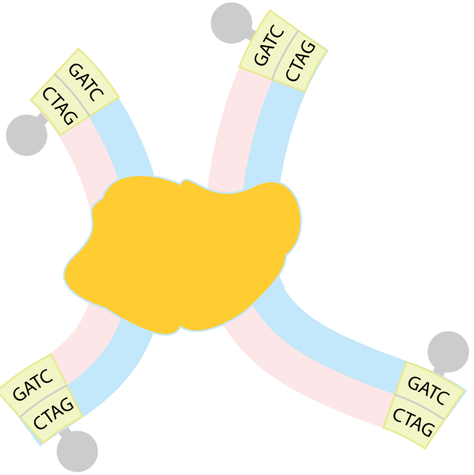
Two pieces of chromatin are close in 3D



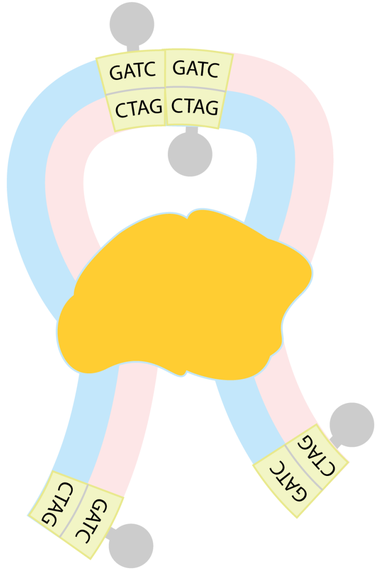
Cross-linking



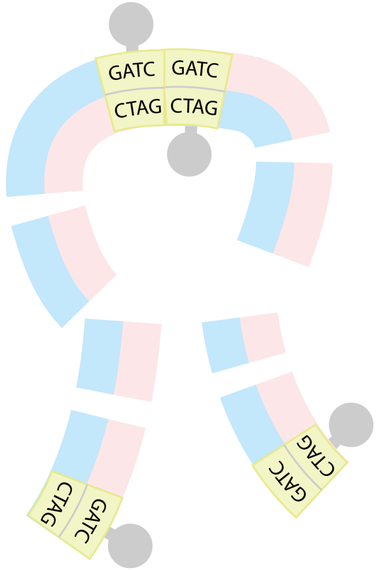
Digestion



Biotinylation

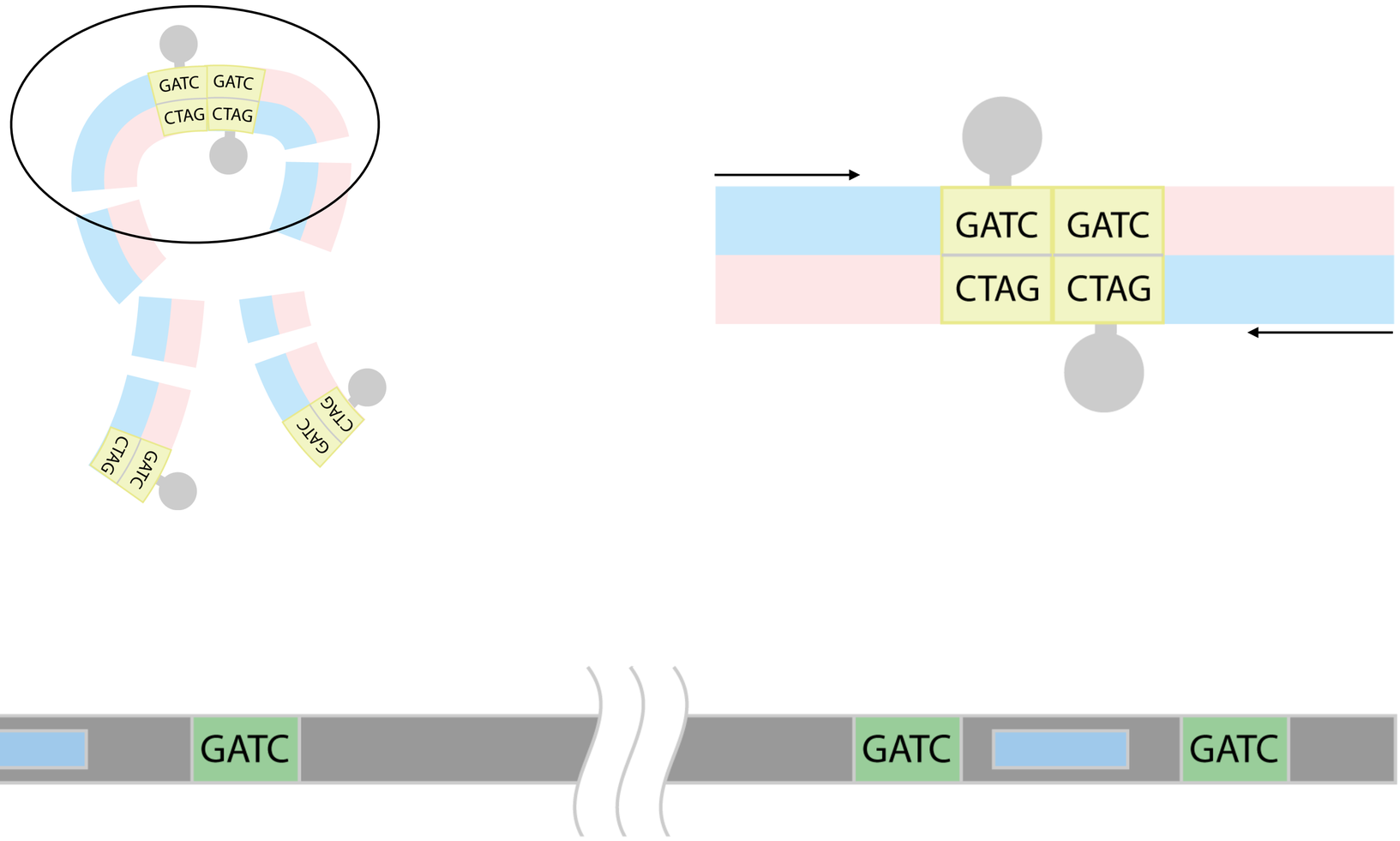


Ligation

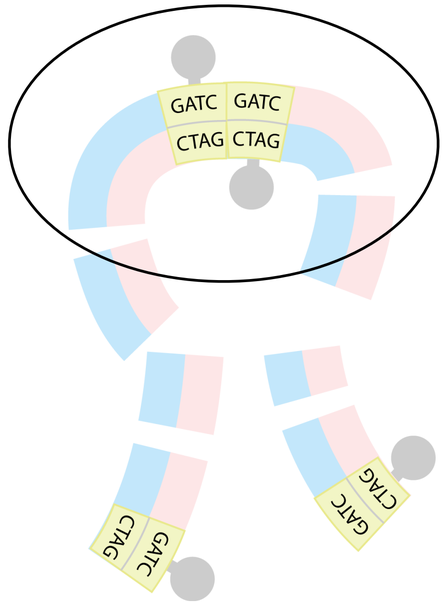


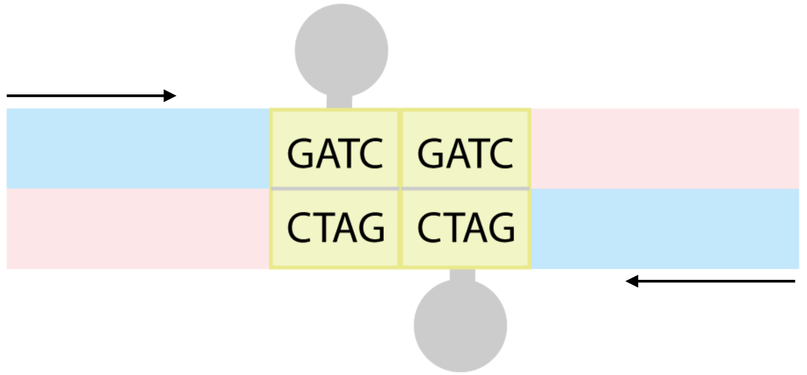
Shearing

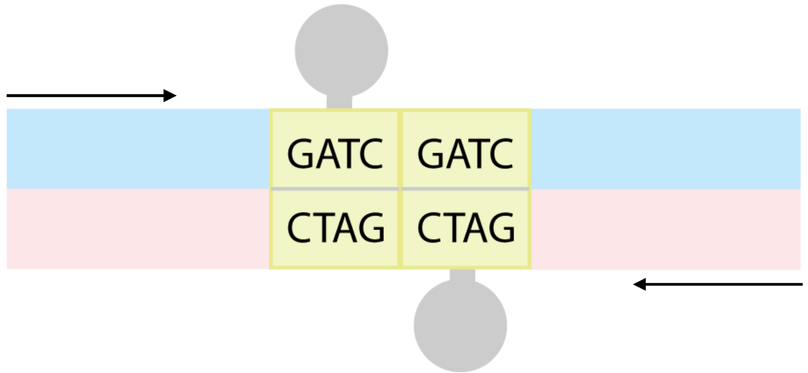
Valid pair


GATC
GATC
CTAG
CTAG
GATC
CTAG
CTAG
GATC
GATC
CTAG
CTAG
GATC
GATC
GATC

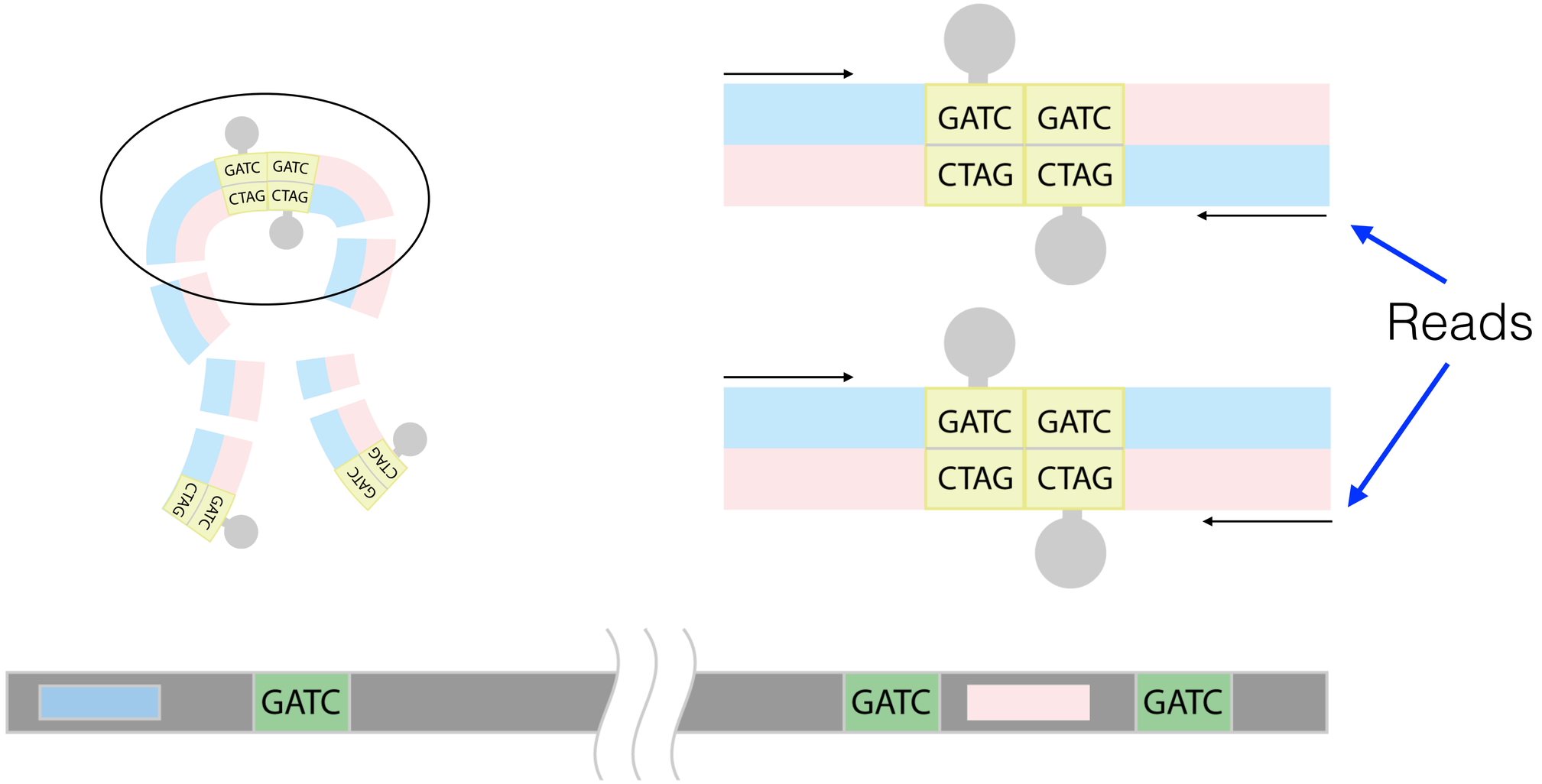
Valid pair (~36% of the mapped read-pairs)



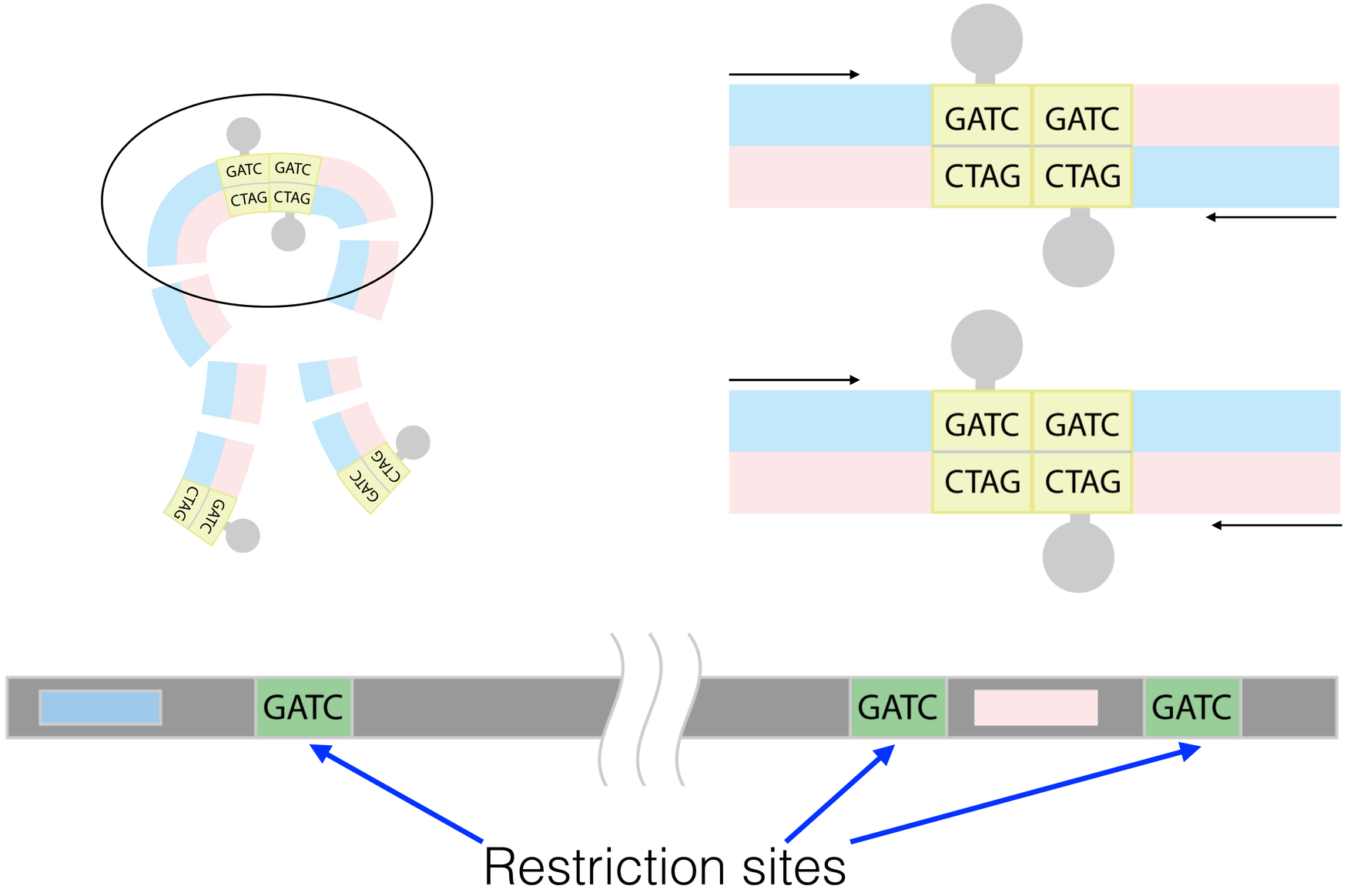




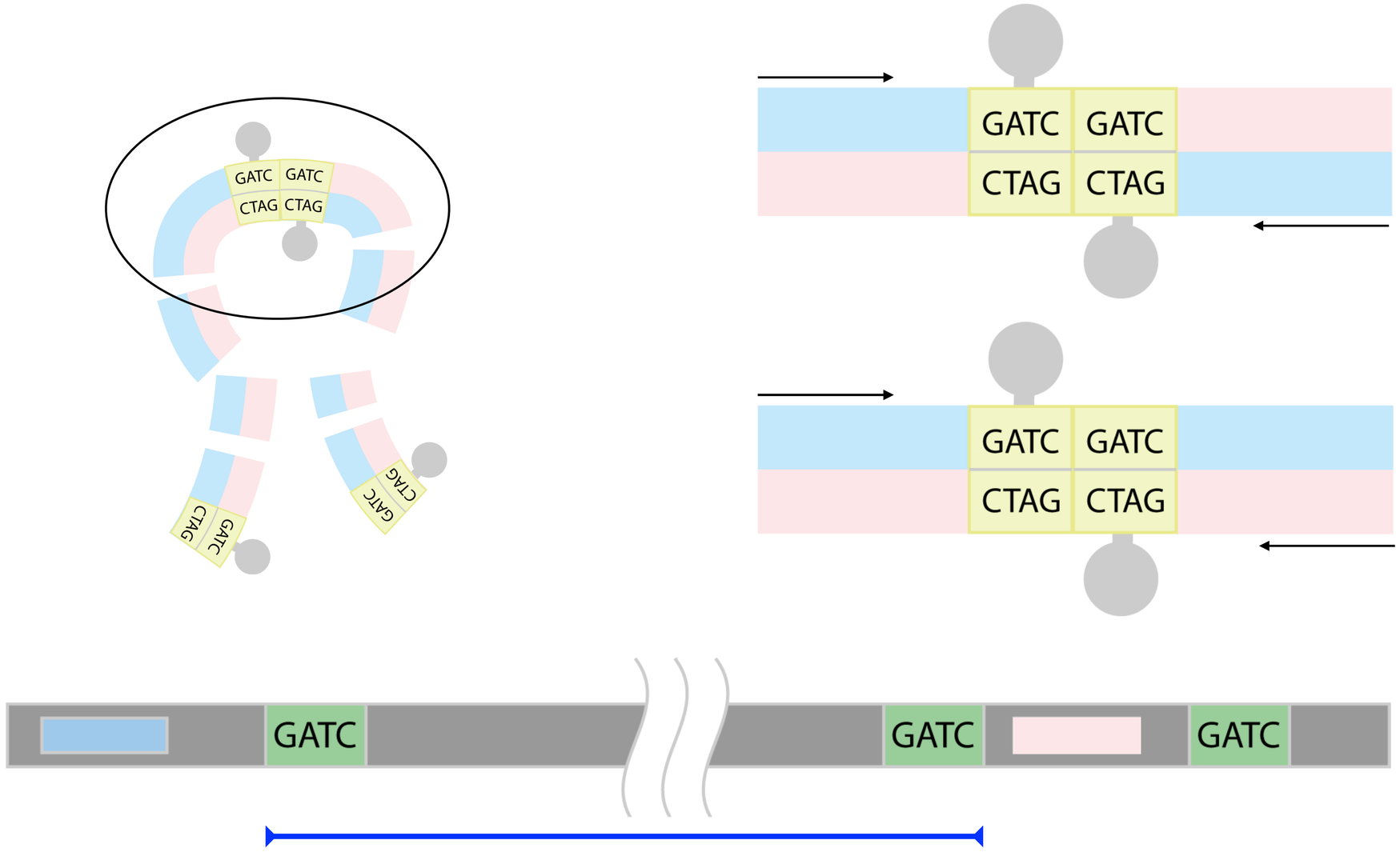
Valid pair (~36% of the mapped read-pairs)



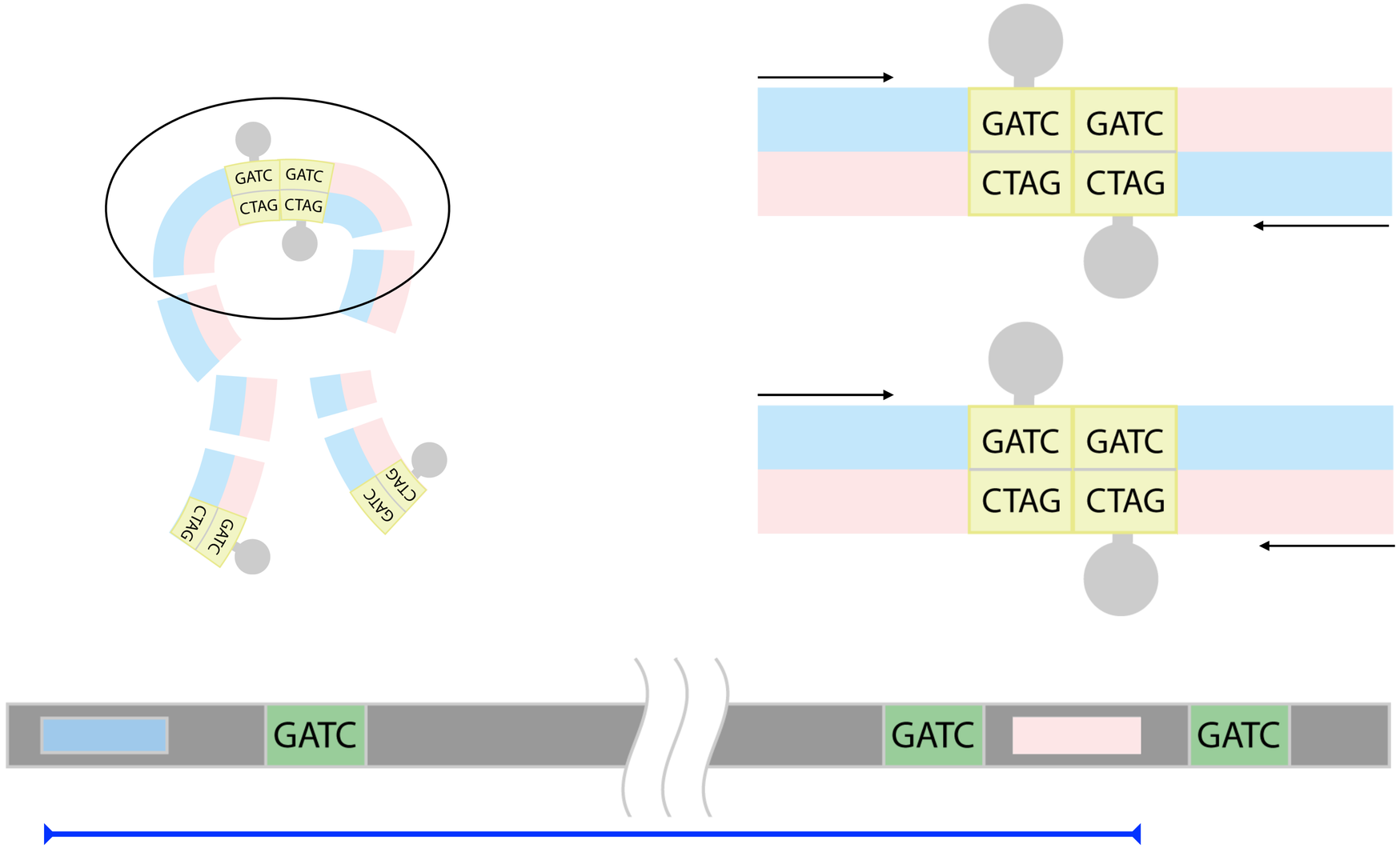
Valid pair (~36% of the mapped read-pairs)



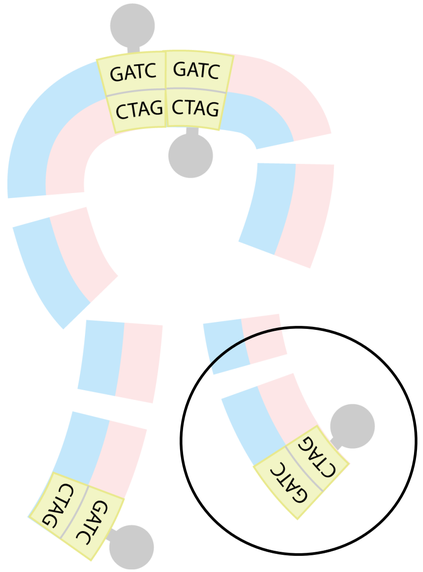
Valid pair (~36% of the mapped read-pairs)

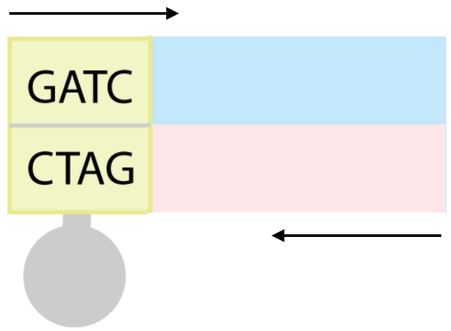


Valid pair (~36% of the mapped read-pairs)

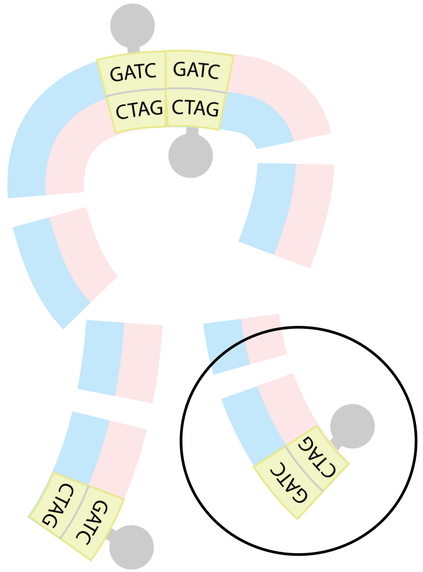


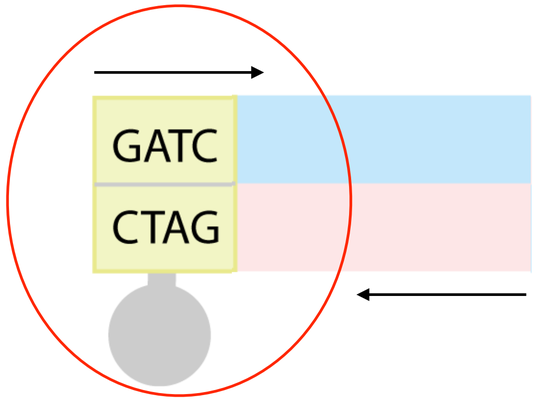
Dangling-end (~15%)



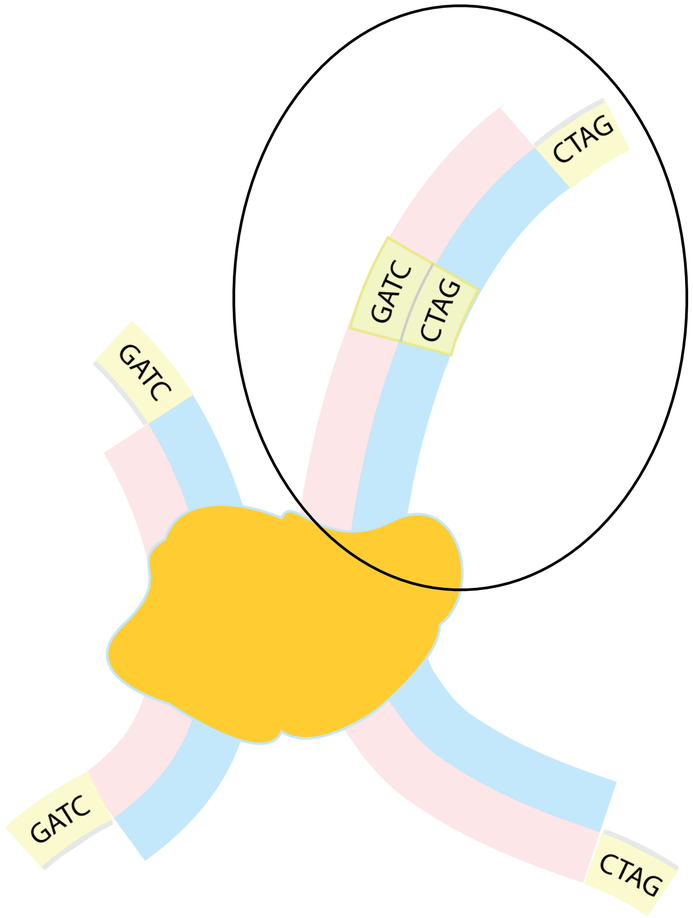


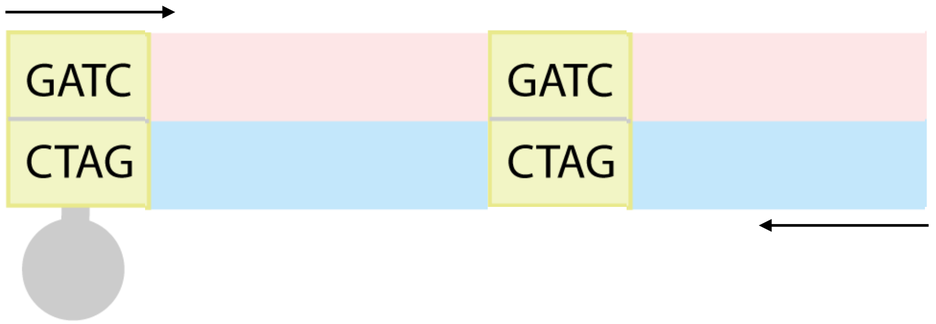
Dangling-end (~15%)



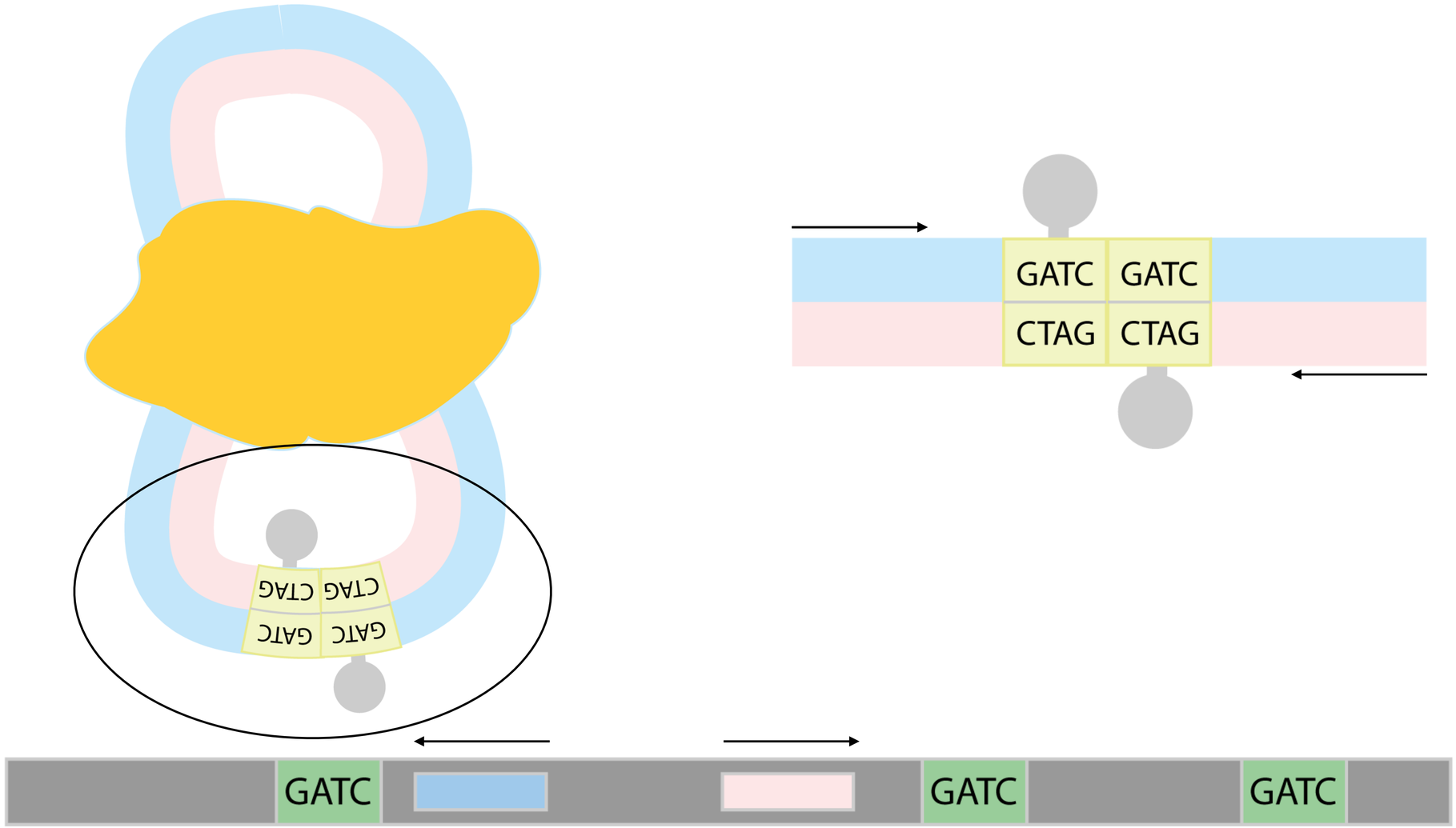


Extra dangling-end (~5%)

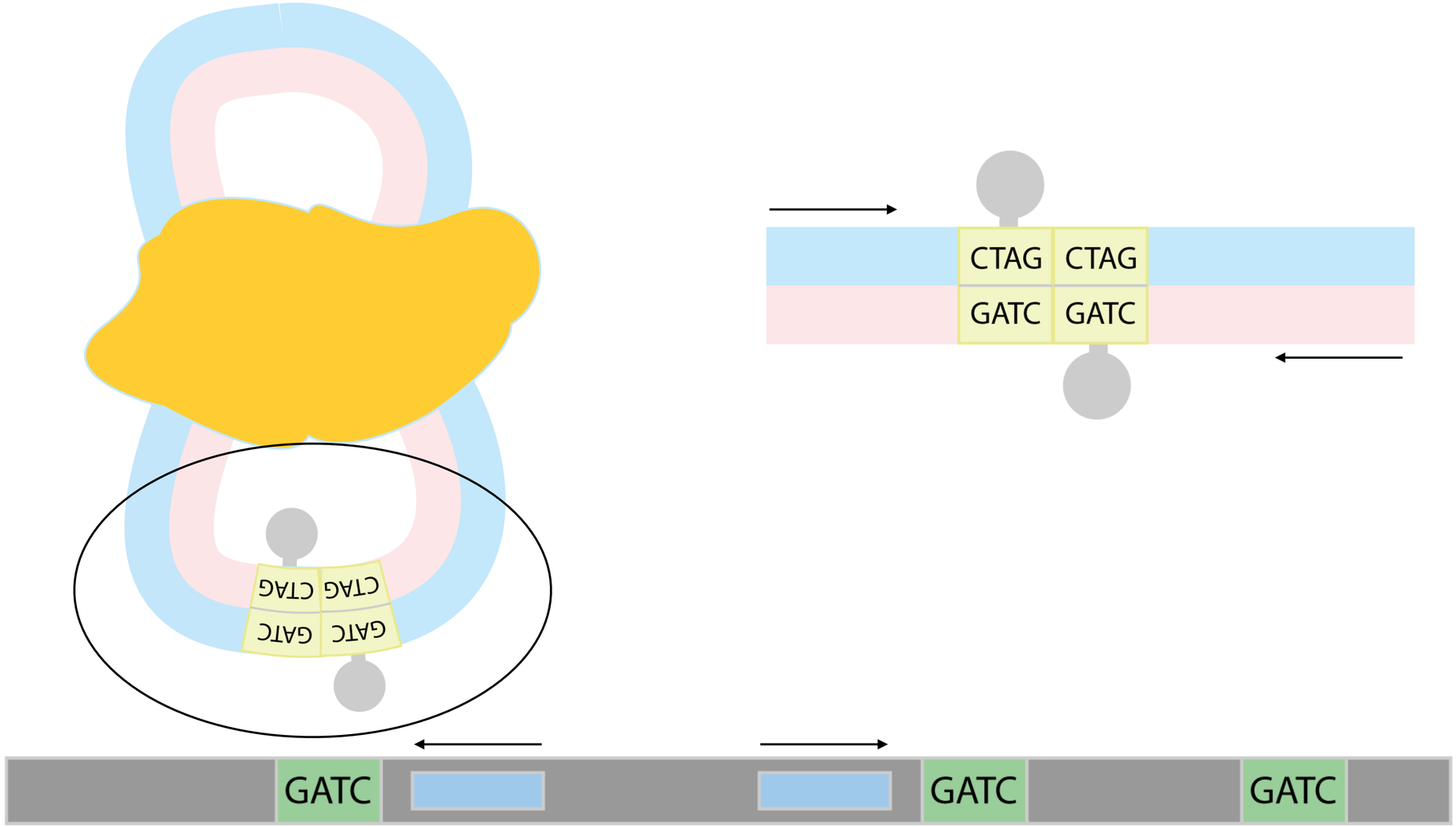




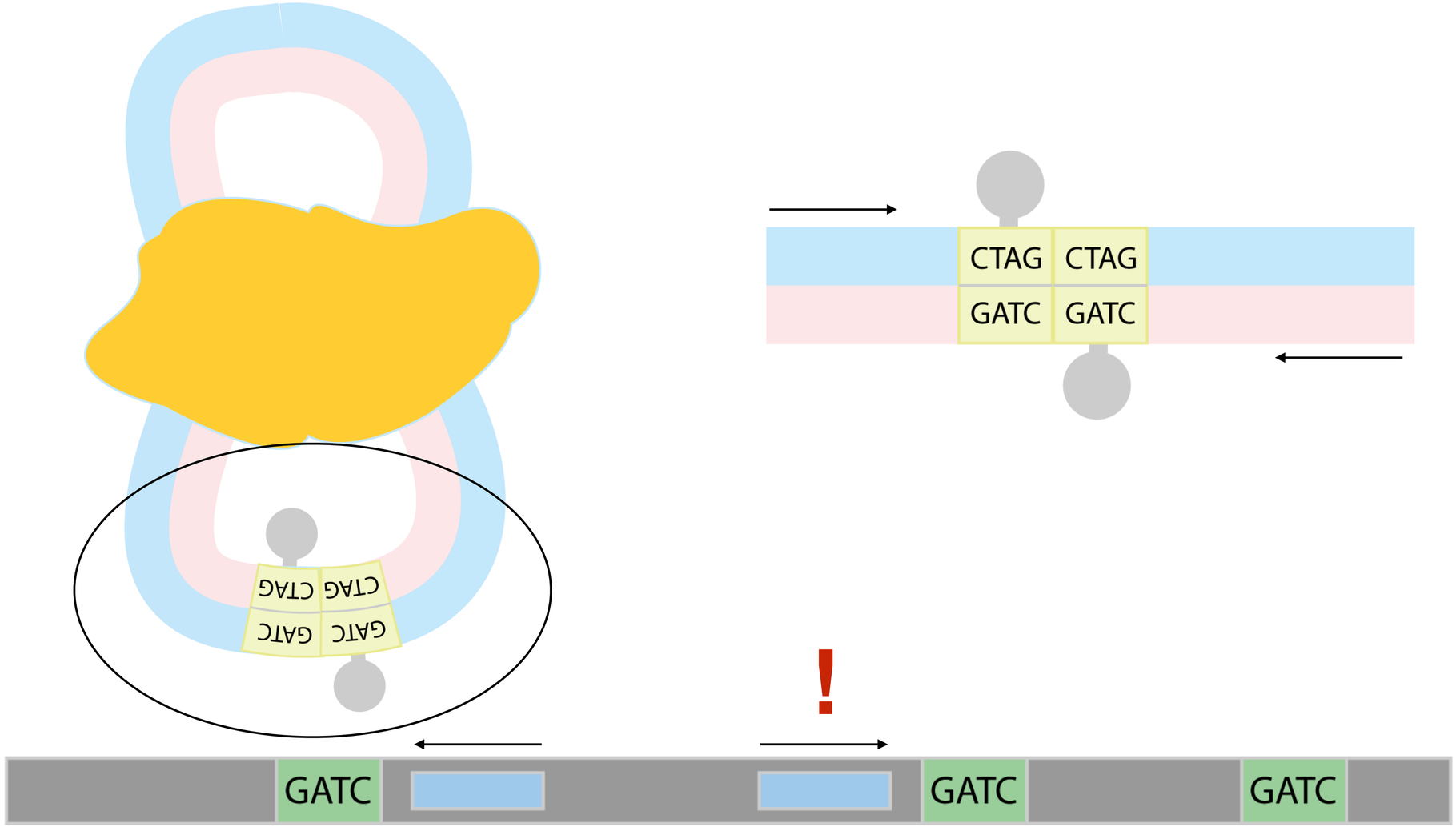
Self-circle (~10%)



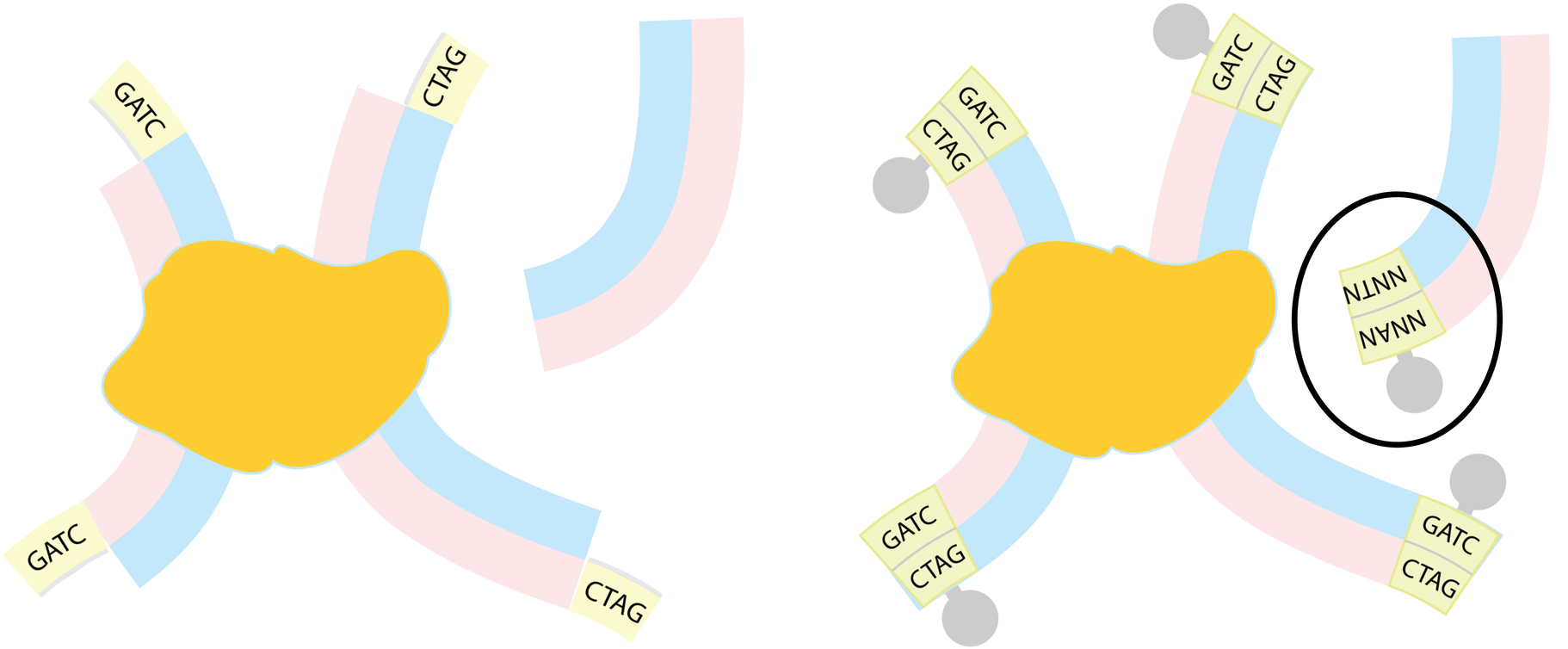
Error (<2%)



Error (<2%)



Random break (~20%)



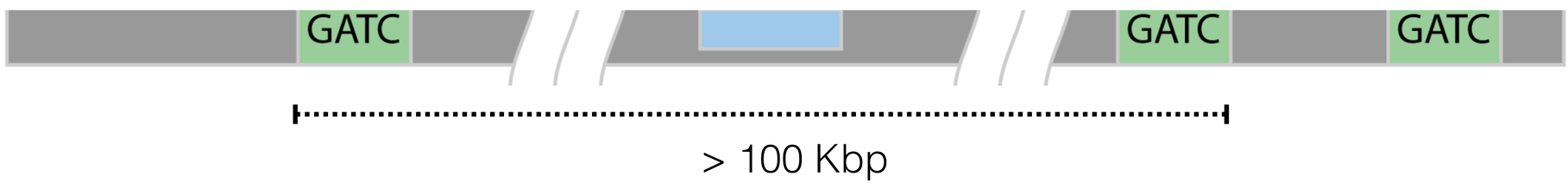
> min_distance_to_RE

> min_distance_to_RE

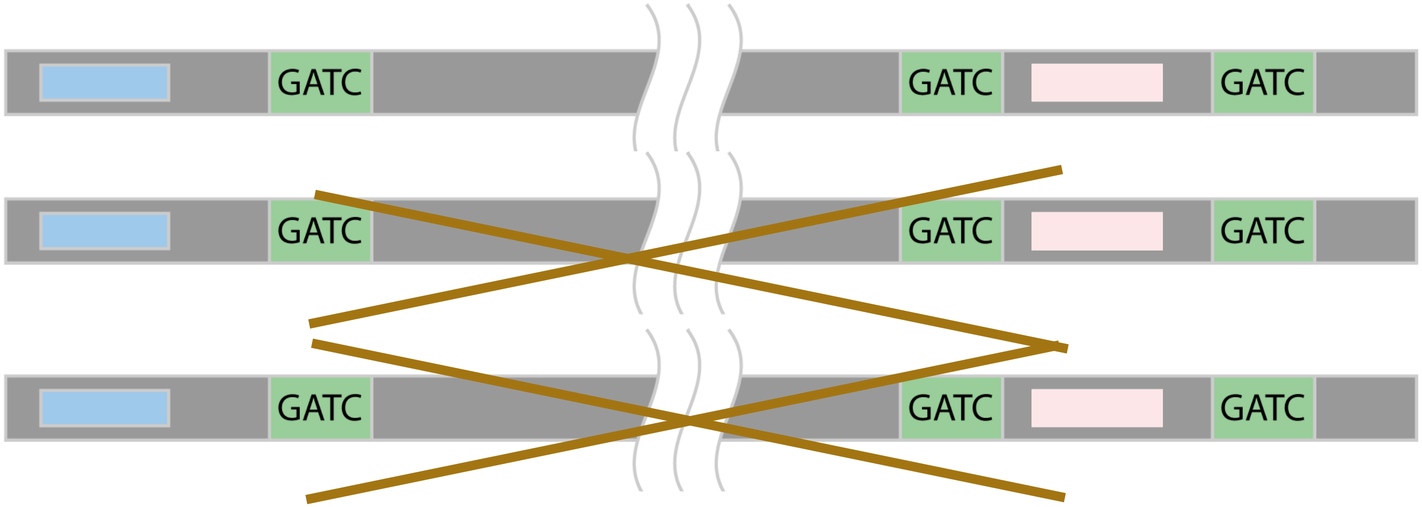
Too close from RE sites (~2%)

Too short (<1%)

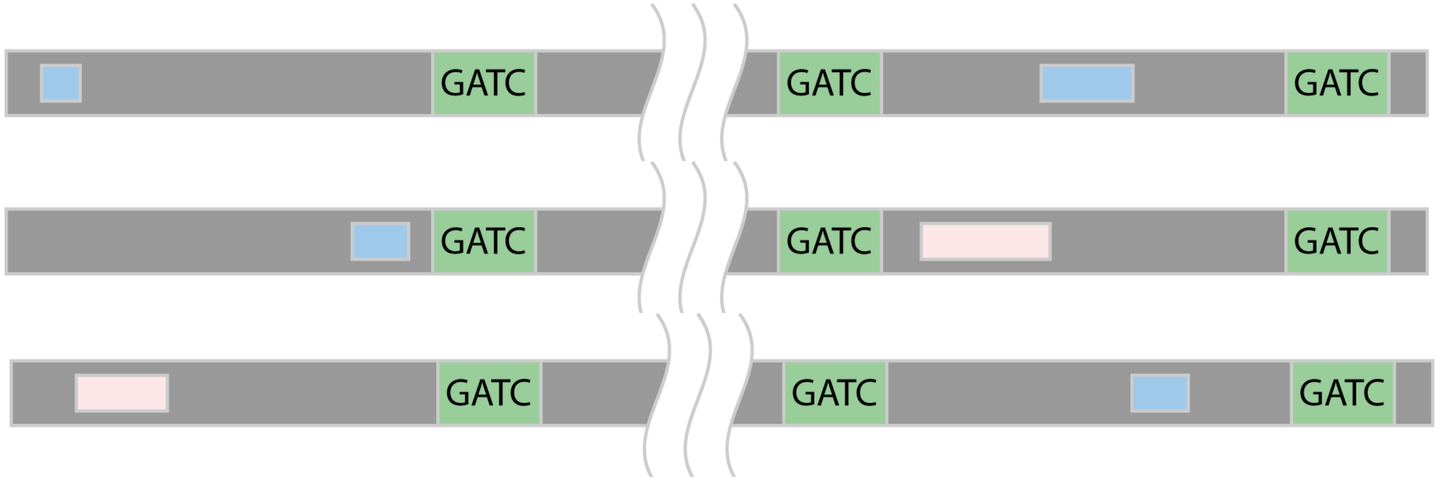
Too large (<1%)



Duplicated (~20%)



Over represented (<1%)



How much do we map?

- Valid reads: 80-90% each end => 64-81% intersection => 60% of the intersection, 36% of the total
- 1% multiple contacts
- many of the reads will be lost in the filtering…