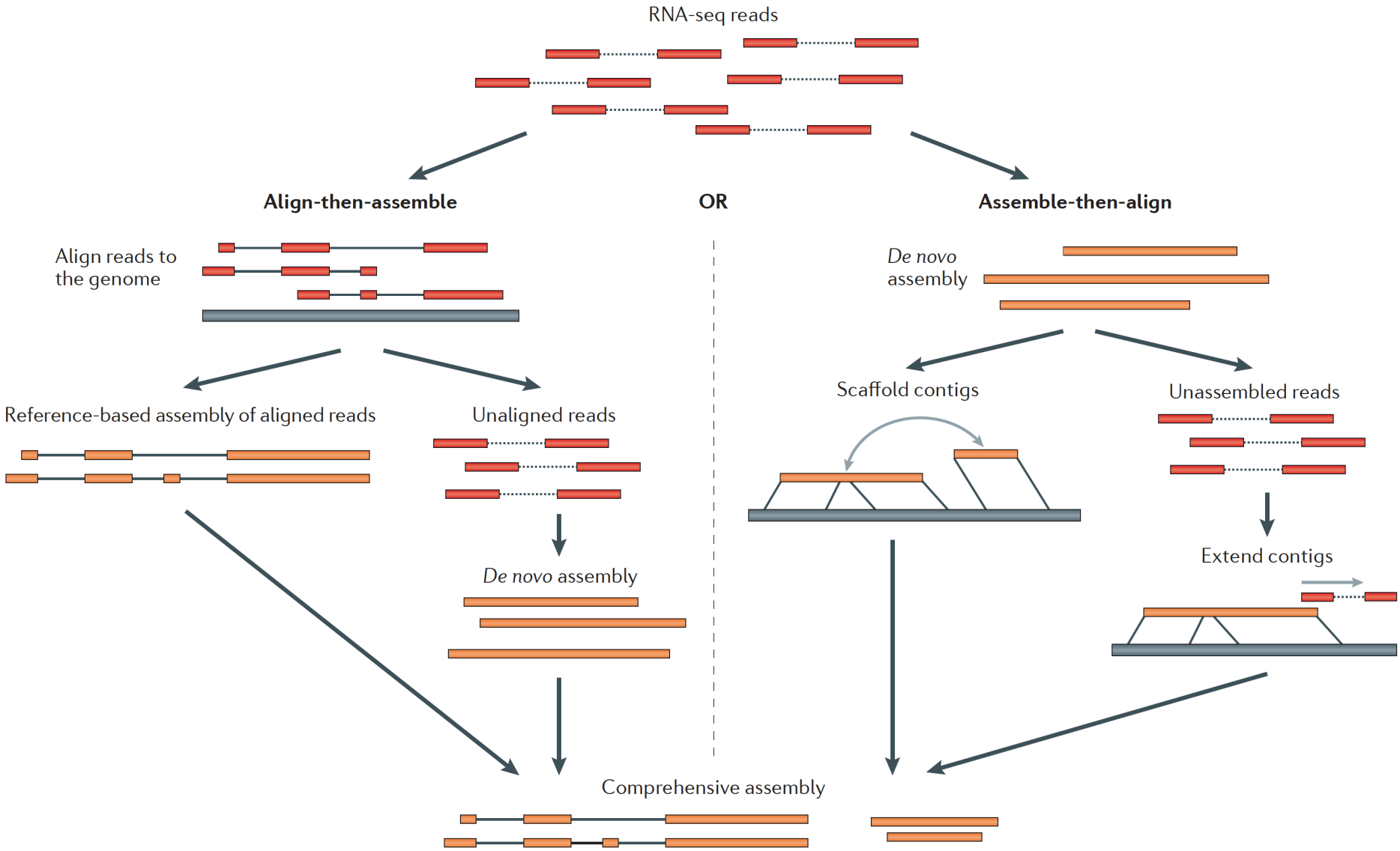
DE NOVO TRANSCRIPTOME RECONSTRUCTION

DE NOVO ASSEMBLY



DE NOVO ASSEMBLY

DE NOVO ASSEMBLY

DE NOVO ASSEMBLY

Keith Bradnam

DE NOVO ASSEMBLY

Keith Bradnam

DE NOVO ASSEMBLY

Repetitive regions are a big problem for genome assembly

DE NOVO ASSEMBLY

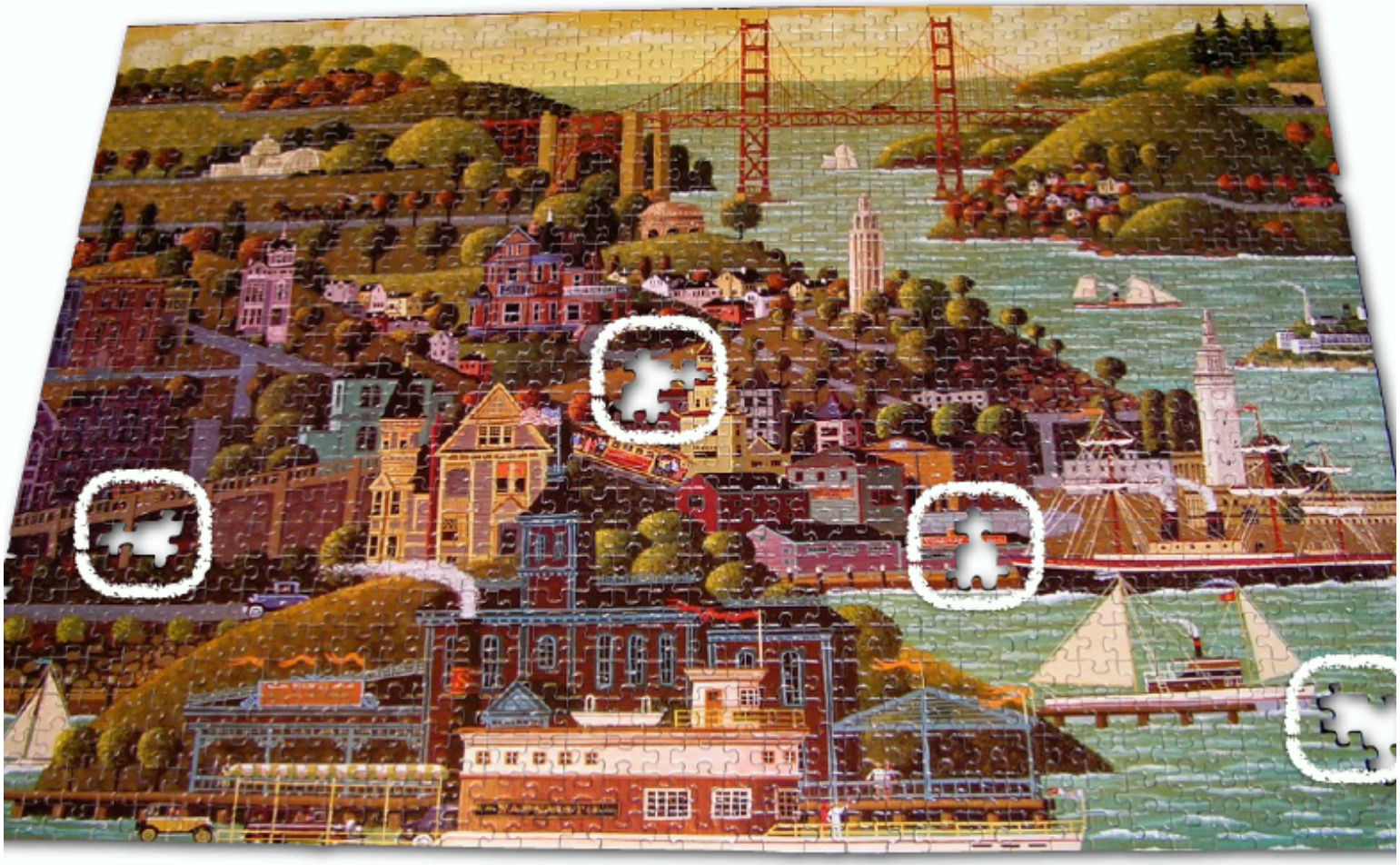
Certain information can help pair together regions

DE NOVO ASSEMBLY

DE NOVO ASSEMBLY

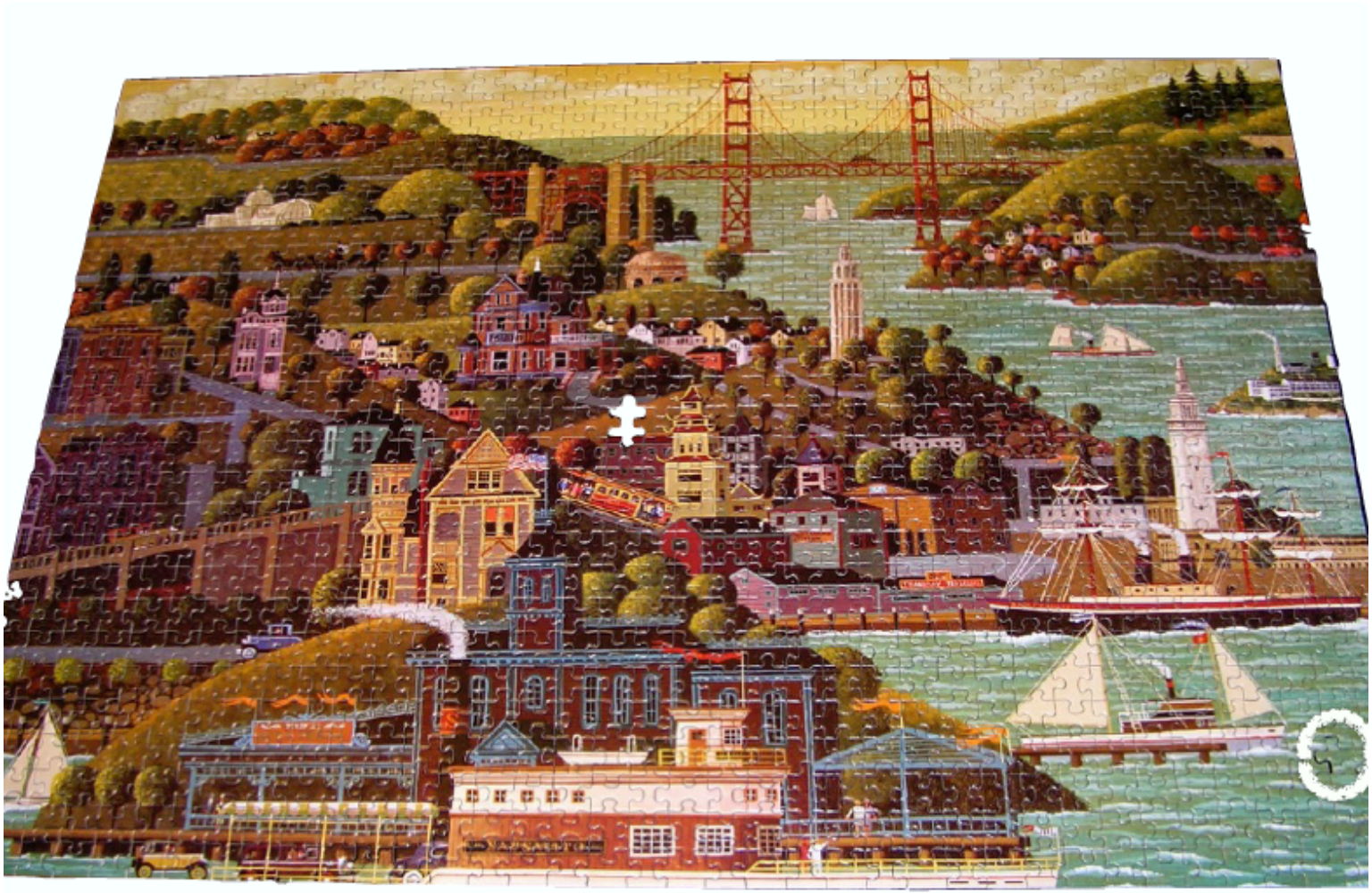
Is this good enough?

DE NOVO ASSEMBLY



We often end up with some missing pieces

DE NOVO ASSEMBLY

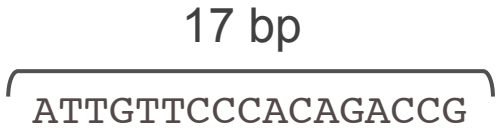


We often try to fit pieces in the wrong way

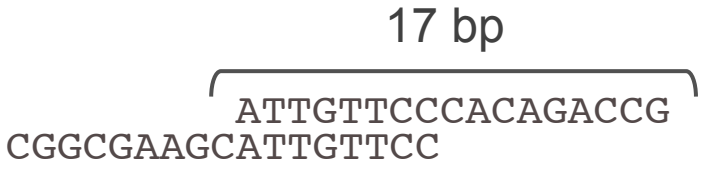
DE NOVO ASSEMBLY

We *never* get to this point with (eukaryotic) genome assembly!

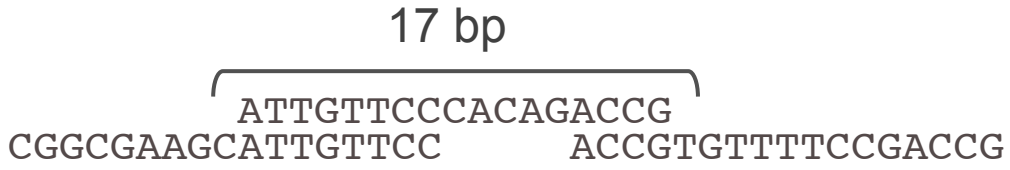
DE NOVO ASSEMBLY



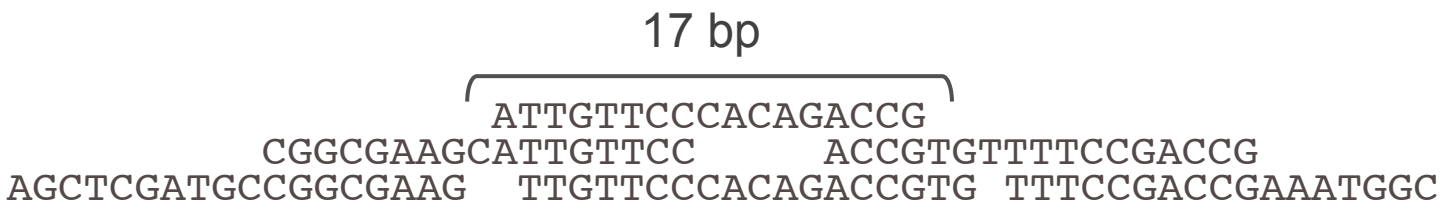
DE NOVO ASSEMBLY



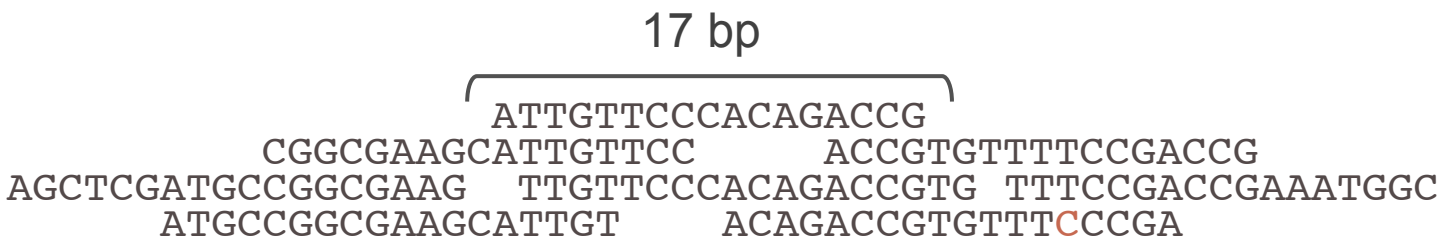
DE NOVO ASSEMBLY



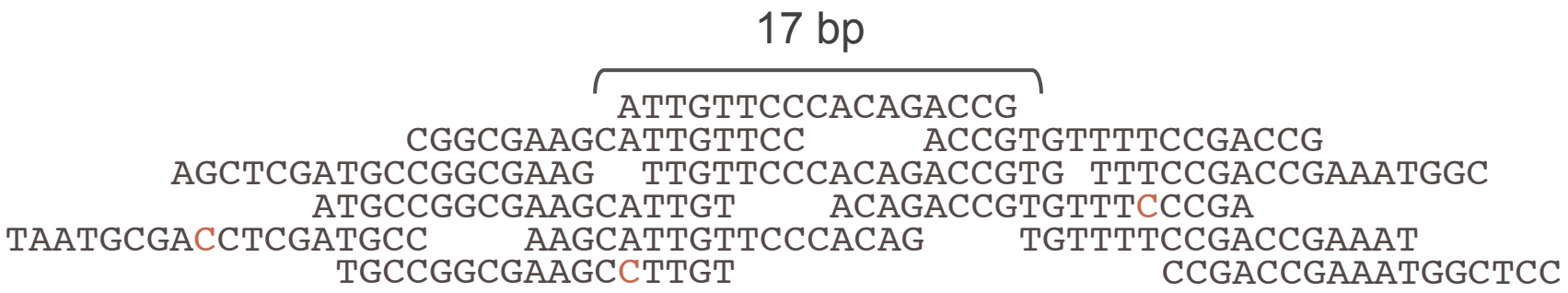
DE NOVO ASSEMBLY



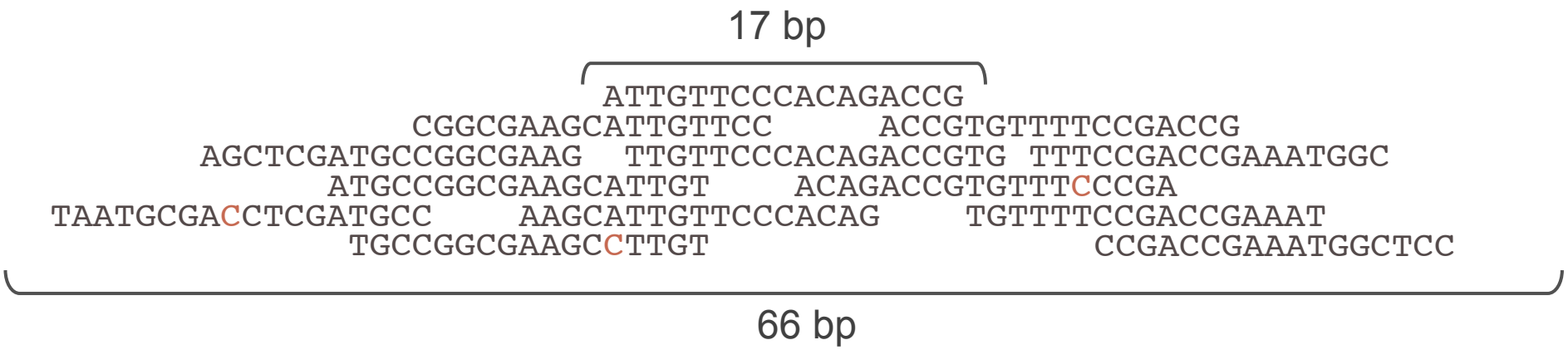
DE NOVO ASSEMBLY



DE NOVO ASSEMBLY



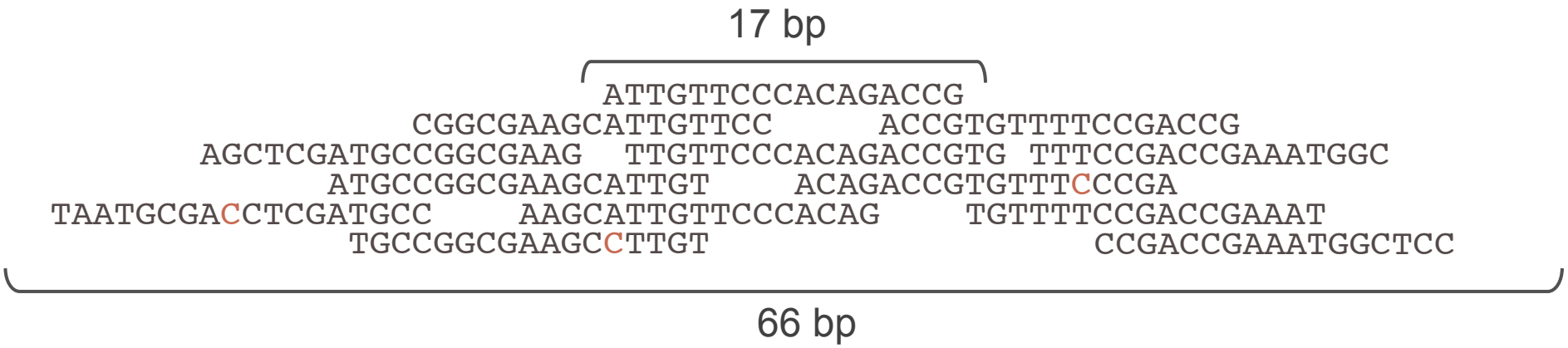
DE NOVO ASSEMBLY



DE NOVO ASSEMBLY

Consensus sequence

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

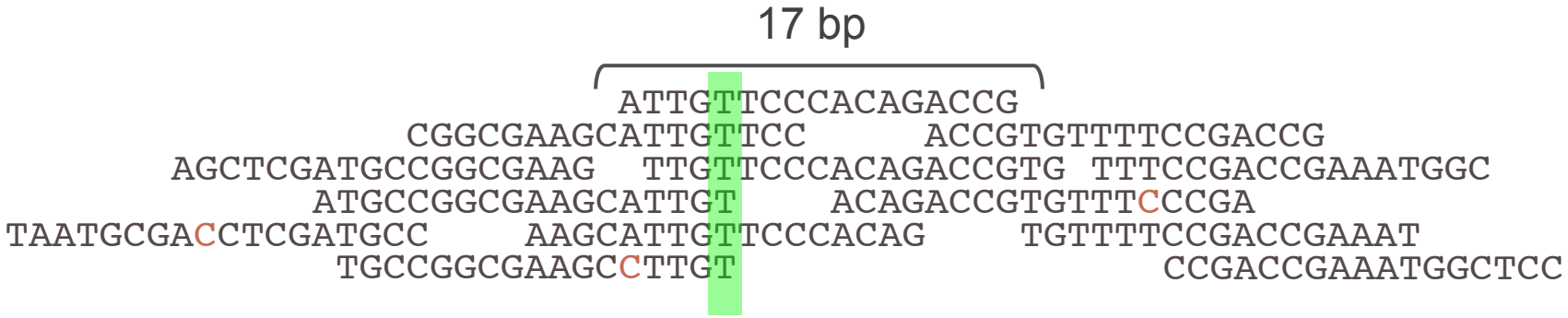


Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



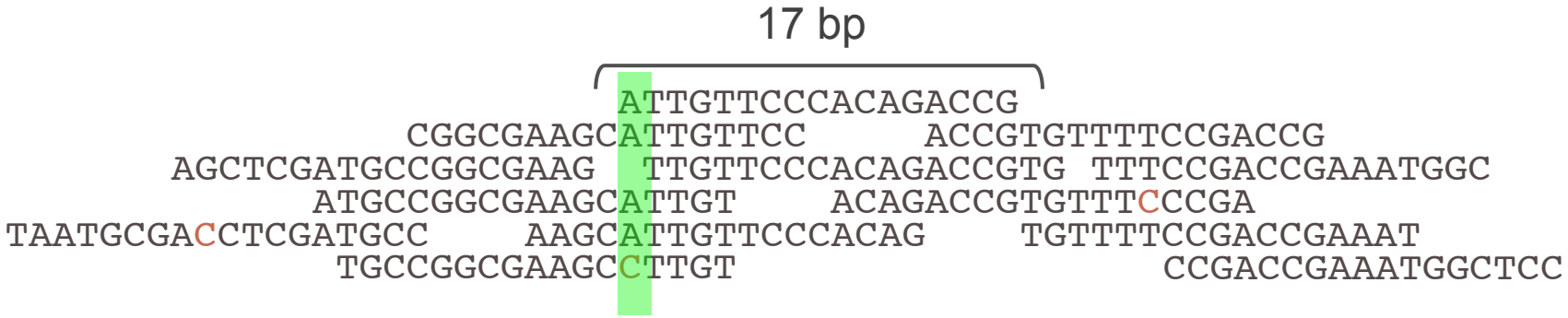
6x coverage
100% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



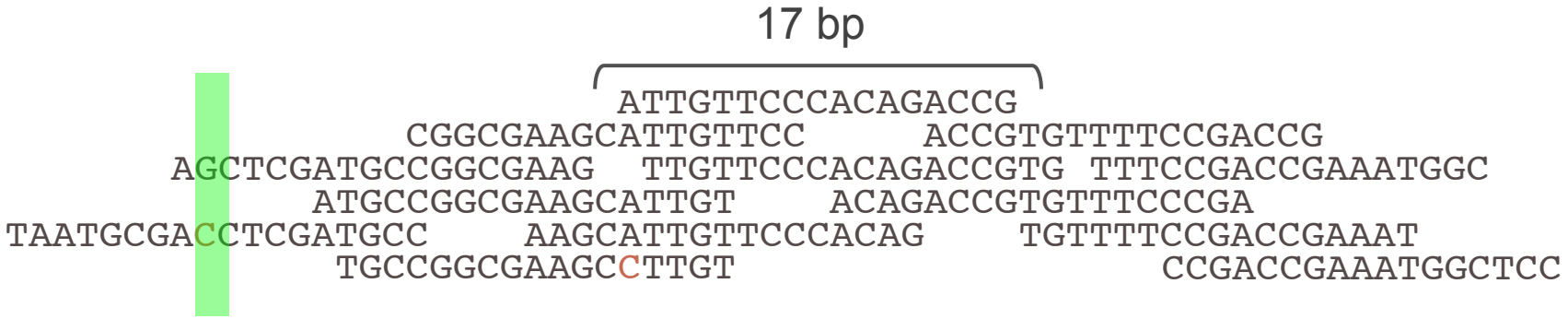
5x coverage
80% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC



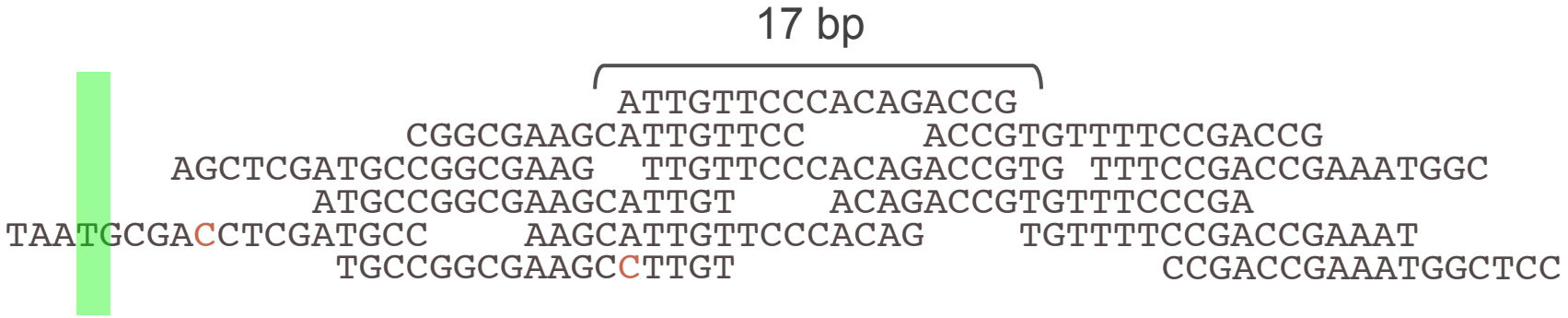
2x coverage
50% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

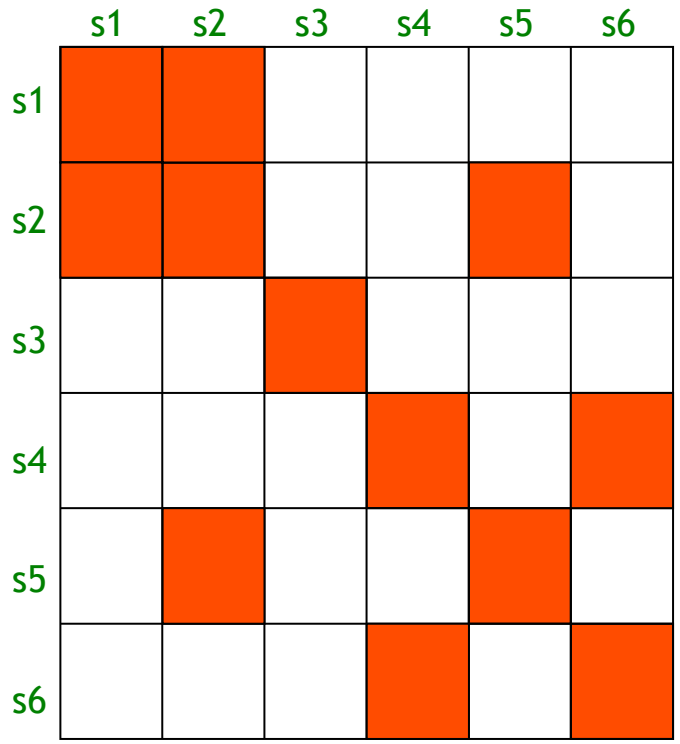


1x coverage

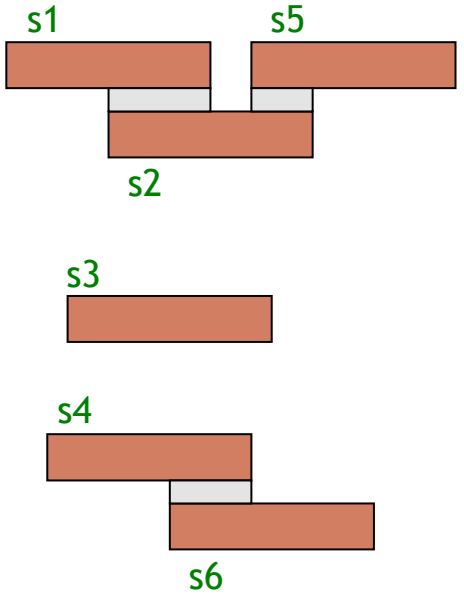
Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

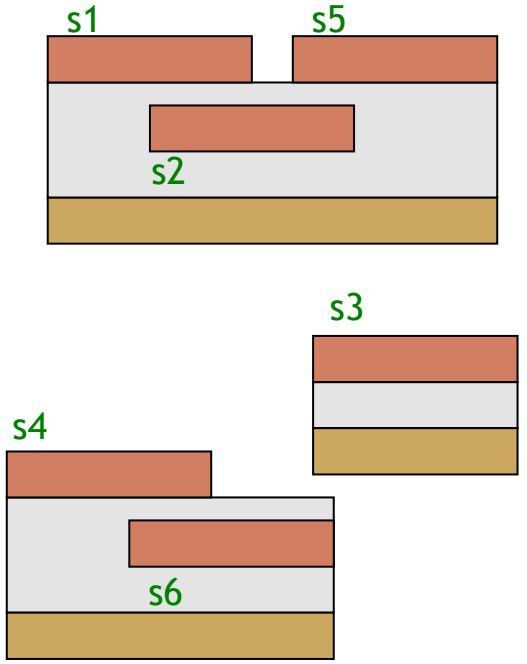
overlap



layout

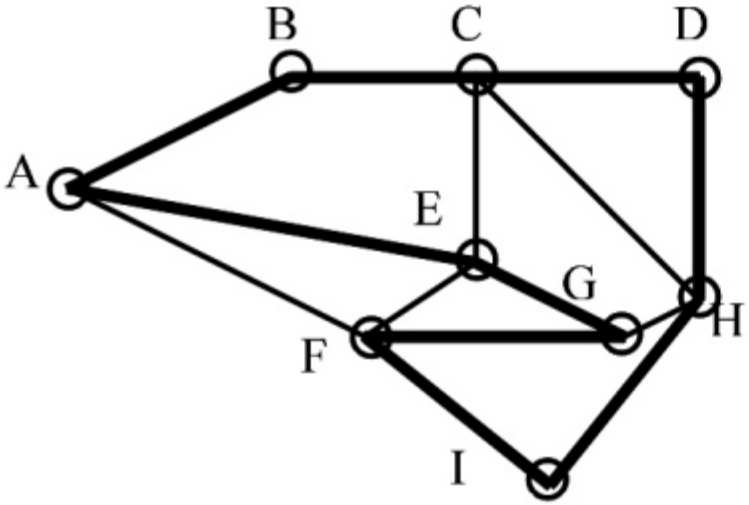


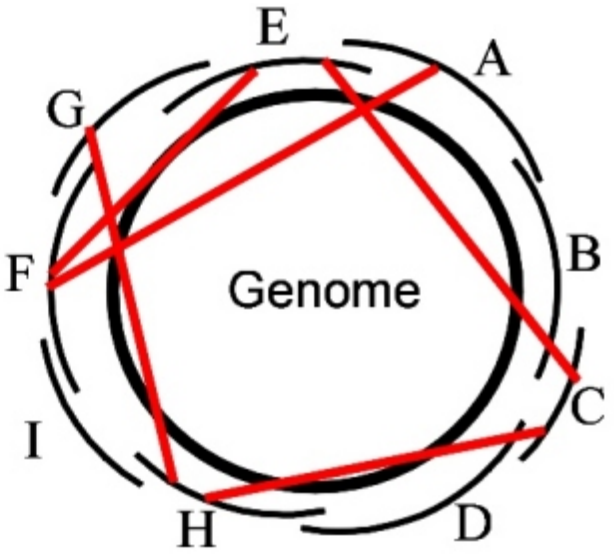
consensus



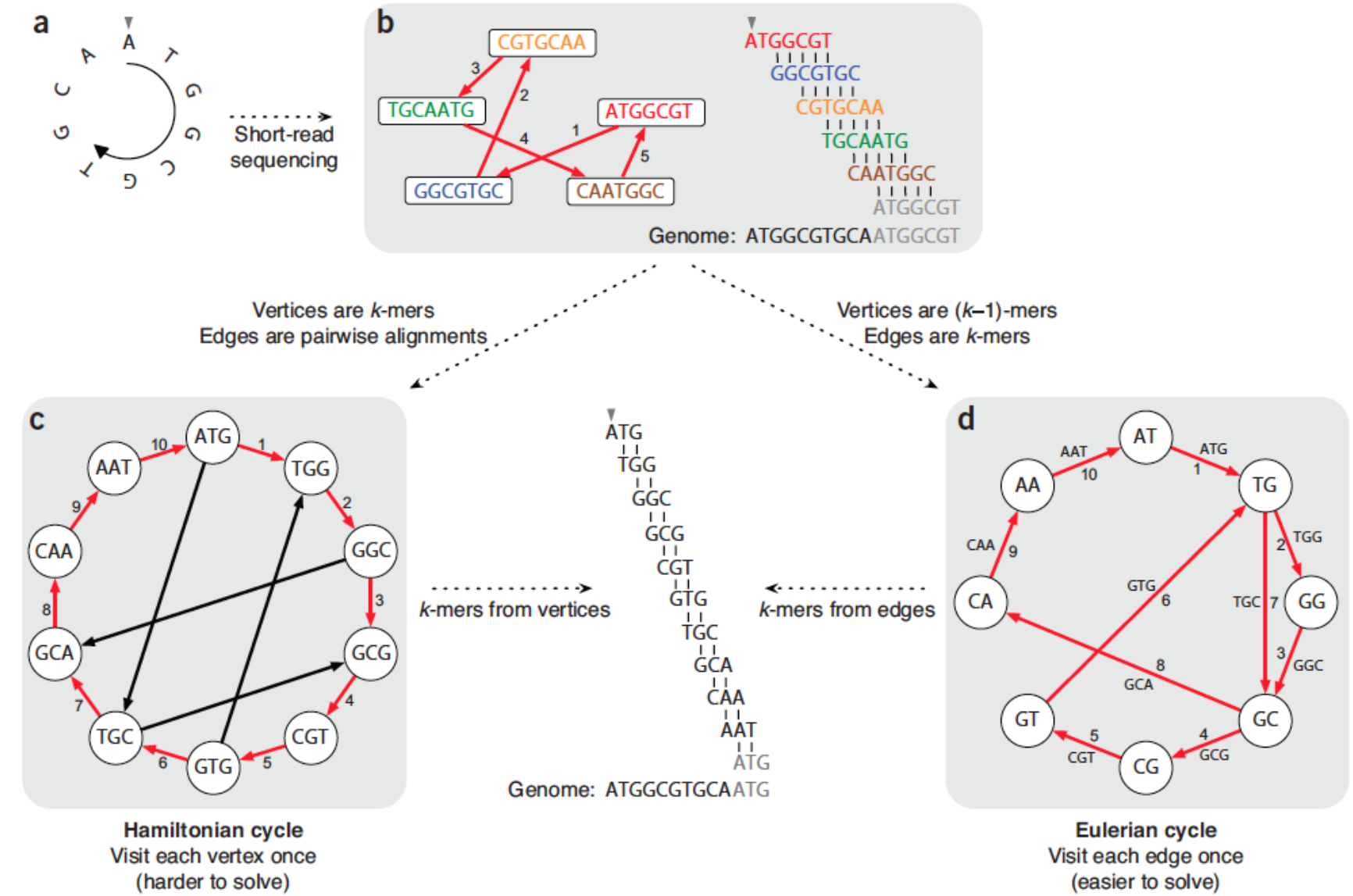
DE NOVO ASSEMBLY

Overlap graph: graph in which nodes represent reads, and edges connect overlapping reads (i.e. reads sharing some similarity). In this representation, the assembly problem can be solved by looking at the shortest path in the graph passing for each node exactly once.





DE NOVO ASSEMBLY



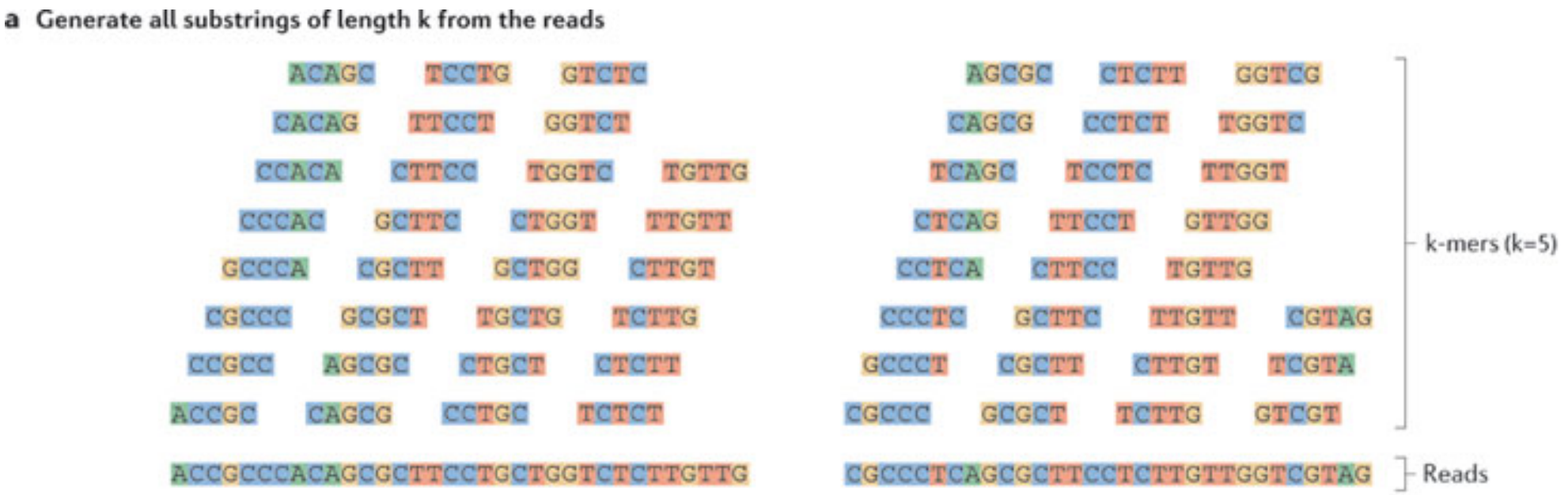
[Compeau et al., 2011]

DE NOVO ASSEMBLY

- The problem of transcriptome assembly is further complicated by the fact that you are not trying to assemble only one sequence, but each gene is a unit that must be treated separately, and one must try to identify which reads originates from the same gene and then assemble them
- Moreover, a gene can encoded for a number of different splicing variants, therefore from the same group of reads one must be able to identify all possible reconstructions, each corresponding to a different splicing variant

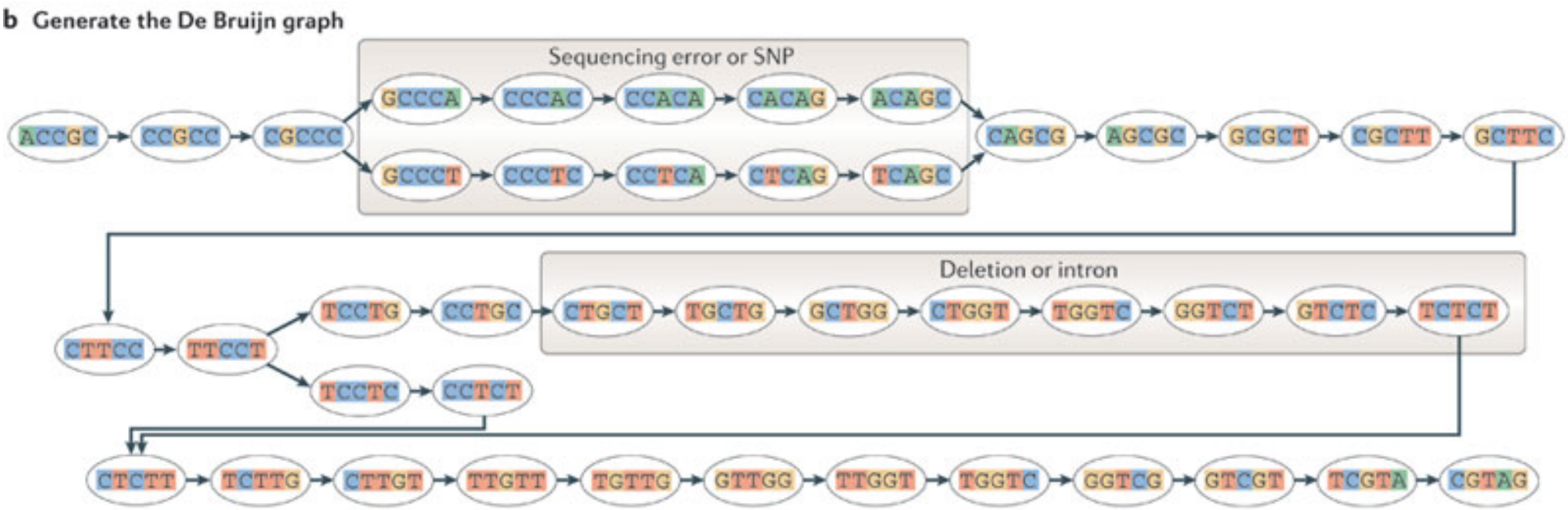
DE NOVO ASSEMBLY

De novo assembly (De Bruijn graph construction)



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

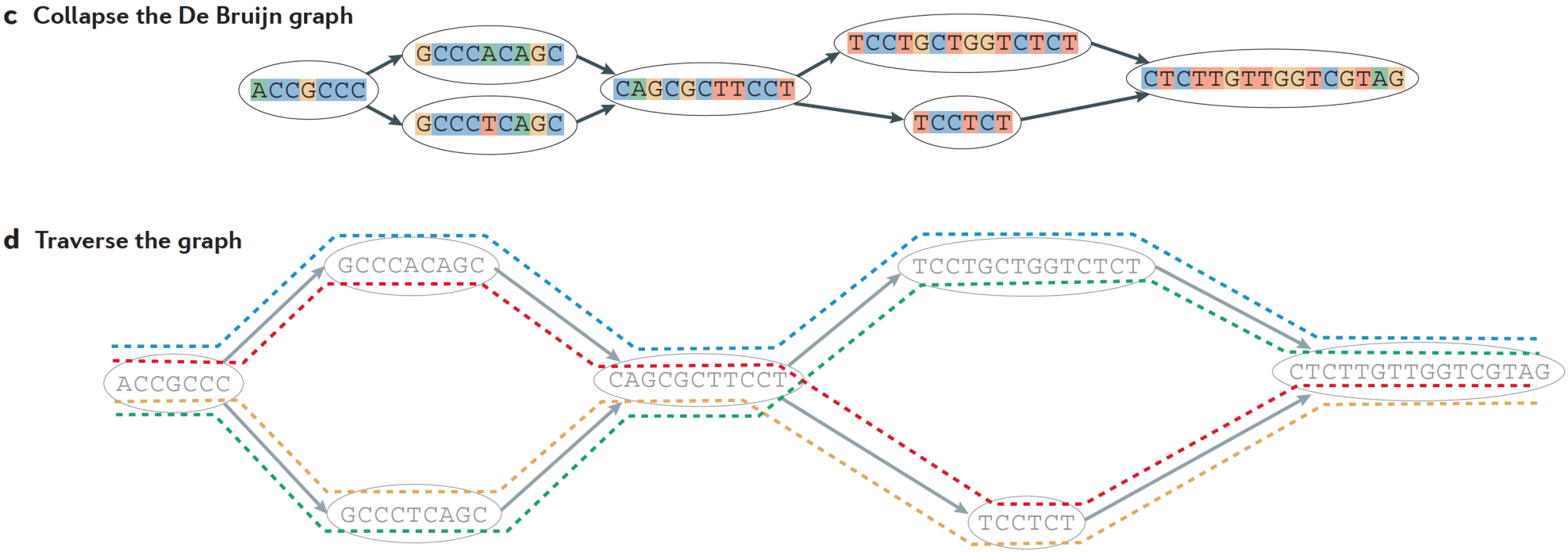
DE NOVO ASSEMBLY



Branching paths can represent different splicing variants encoded by the gene

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

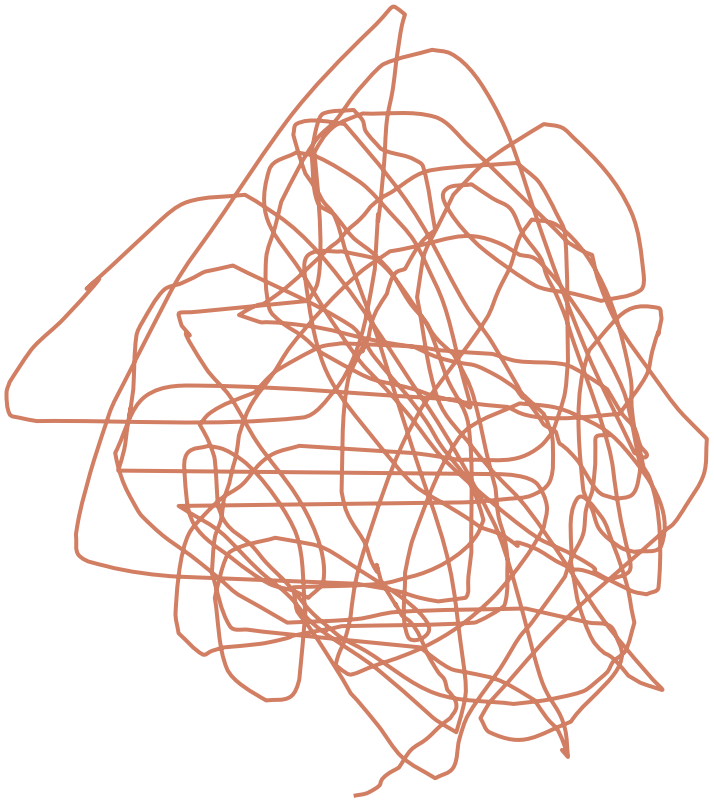
DE NOVO ASSEMBLY



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

DE NOVO ASSEMBLY

Genome Assembly
Single Massive Graph



Entire chromosomes represented

Trinity Transcriptome Assembly
Many Thousands of Small Graphs

Ideally, one graph per expressed gene

DE NOVO ASSEMBLY

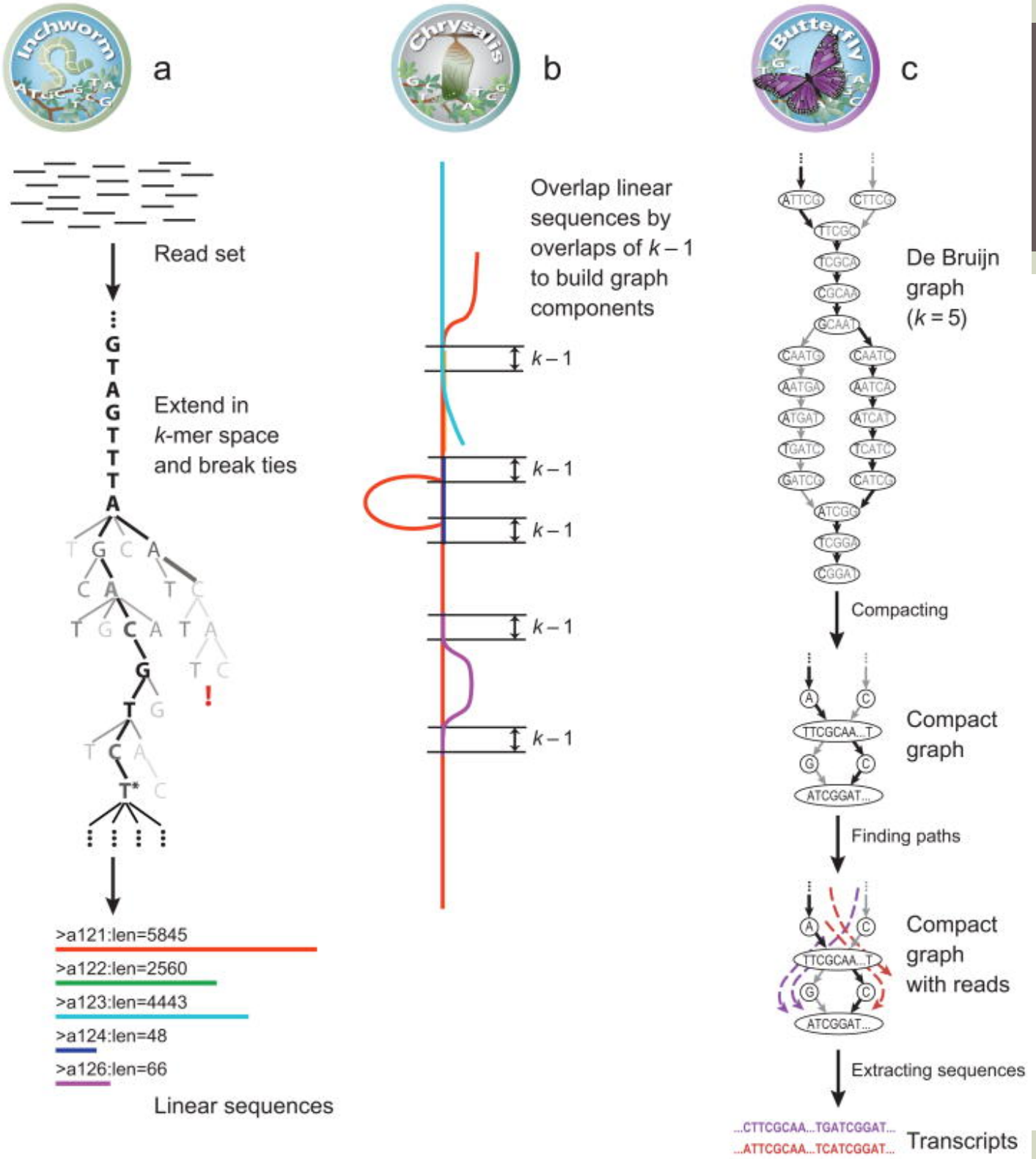
Trinity – How it works:

RNA-Seq reads

Linear contigs

de-Bruijn graphs

Transcripts + Isoforms


Inchworm
a
Chrysalis
b
Butterfly
c
Read set
Extend in
k-mer space
and break ties
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a126:len=66
Linear sequences
Overlap linear
sequences by
overlaps of k – 1
to build graph
components
k – 1
De Bruijn
graph
(k = 5)
Compacting
Compact
graph
Finding paths
Compact
graph
with reads
Extracting sequences
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...
Transcripts