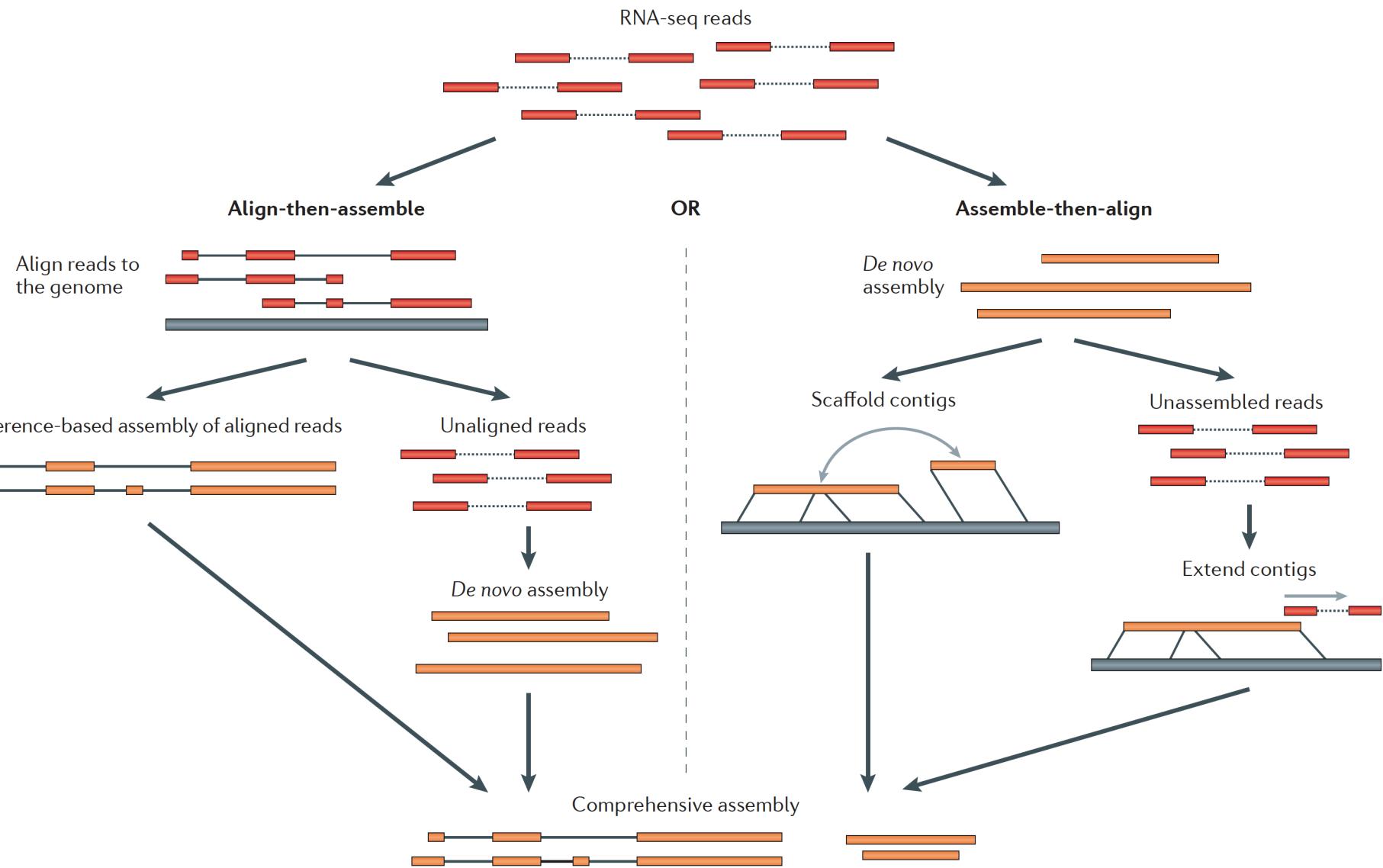


DE NOVO TRANSCRIPTOME RECONSTRUCTION

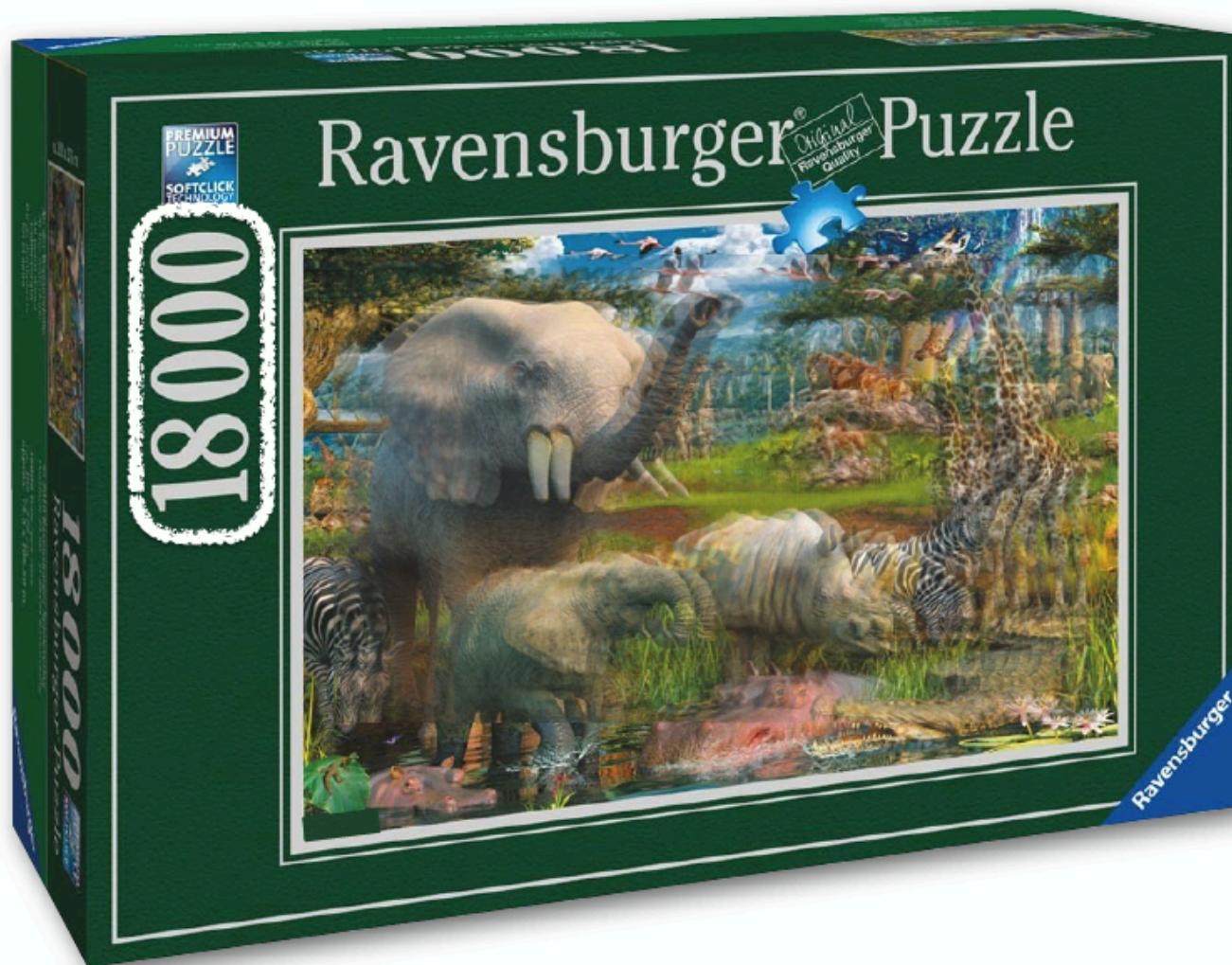
DE NOVO ASSEMBLY



DE NOVO ASSEMBLY



DE NOVO ASSEMBLY



DE NOVO ASSEMBLY



DE NOVO ASSEMBLY

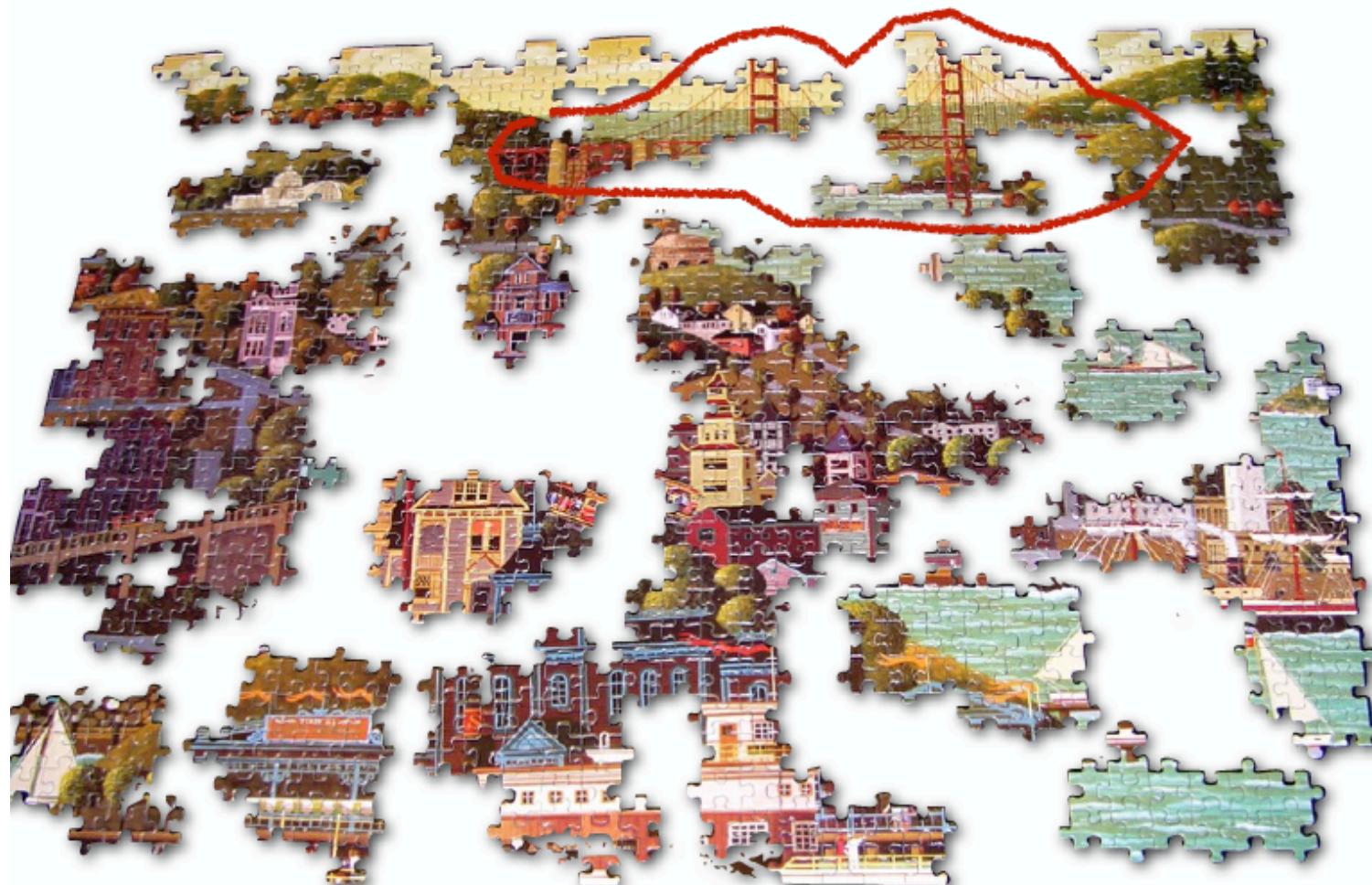


DE NOVO ASSEMBLY



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

17 bp

ATTGTTCCCCACAGACCG

DE NOVO ASSEMBLY

17 bp

```
ATTGTTCCCACAGACCG
CGGCGAAGCATTGTTCC
```

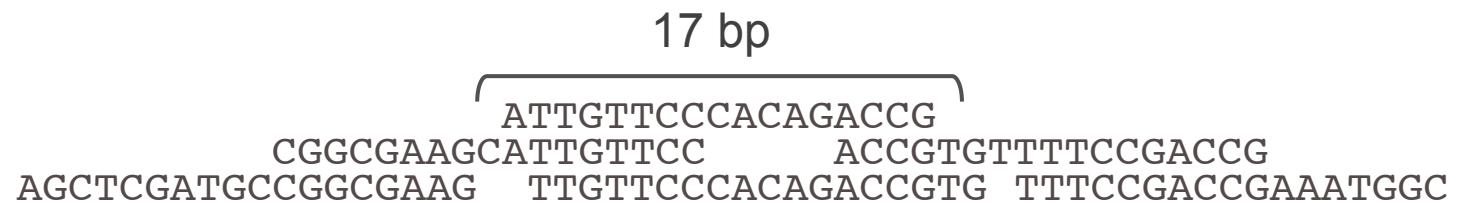
DE NOVO ASSEMBLY

17 bp

```
ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC      ACCGTGTTTCCGACCG
```

DE NOVO ASSEMBLY

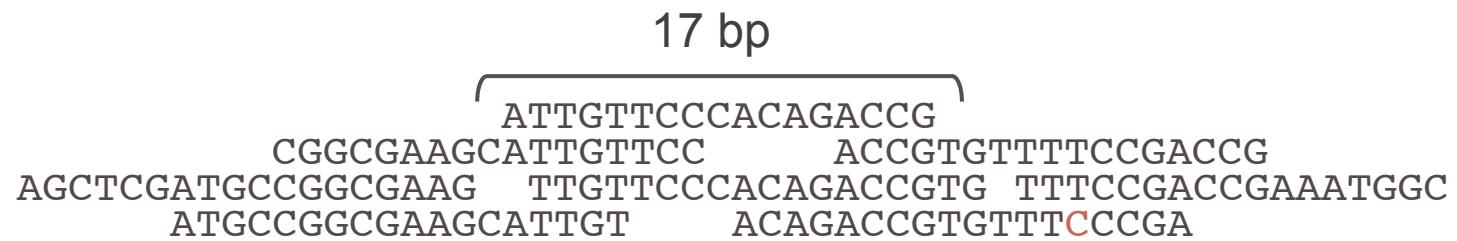
17 bp



```
ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC ACCGTGTTTCCGACCG
AGCTCGATGCCGGCGAAG TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
```

DE NOVO ASSEMBLY

17 bp



ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC ACCGTGTTTCCGACCG
AGCTCGATGCCGGCGAAG TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
ATGCCGGCGAAGCATTGT ACAGACCGTGTTC**CCGA**

DE NOVO ASSEMBLY

17 bp

The diagram illustrates a sequence assembly. A horizontal bracket above the sequence indicates a length of "17 bp". The sequence itself consists of two rows of DNA bases. The top row contains: ATTGTTCCCACAGACCG, CGCGAAGCATTGTTCC, ACCGTGTTTCCGACCG, AGCTCGATGCCGGCGAAG, TTGTTCCCACAGACCGTG, TTTCCGACCGAAATGGC, ATGCCGGCGAAGCATTGT, ACAGACCGTGTTC, and CCGACCGAAATGGCTCC. The bottom row contains: TAATGCGAC, CTCGATGCC, AAGCAT, TGTTTCCGACCGAAAT, and TGCCGGCGAAGC, followed by a red "CTTGT". The red highlighted segments "CTCGATGCC" and "TGCCGGCGAAGC" align with the "17 bp" bracket.

ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC ACCGTGTTTCCGACCG
AGCTCGATGCCGGCGAAG TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
ATGCCGGCGAAGCATTGT ACAGACCGTGTTC CCGACCGAAATGGCTCC
TAATGCGAC CTCGATGCC AAGCAT TGTTTCCGACCGAAAT
TGCCGGCGAAGC CTTGT CCGACCGAAATGGCTCC

DE NOVO ASSEMBLY

17 bp

ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC ACCGTGTTTCCGACCG
AGCTCGATGCCGGCGAAG TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
ATGCCGGCGAAGCATTGT ACAGACCGTGTTT**C**CCGA
TAATGCGAC**C**TCGATGCC AAGCATTGTCCCCACAG TGTTTCCGACCGAAAT
TGCCGGCGAAGC**CT**TGT CCGACCGAAATGGCTCC

66 bp

DE NOVO ASSEMBLY

Consensus sequence

TAATGCGACCTCGATGCCGGCGAACGCATTGTTCCCACAGACCGTGTTCGACCGAAATGGCTCC

17 bp

ATTGTTCCCACAGACCG
CGCGAAGCATTGTTCC ACCGTGTTTCCGACCG
AGCTCGATGCCGGCGAACG TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
ATGCCGGCGAACGCATTGT ACAGACCGTGTTC CCGA
TAATGCGACCTCGATGCC AAGCATTGTTCCCACAG TGTTTCCGACCGAAAT
TGCCGGCGAACGC TTGT CCGACCGAAATGGCTCC

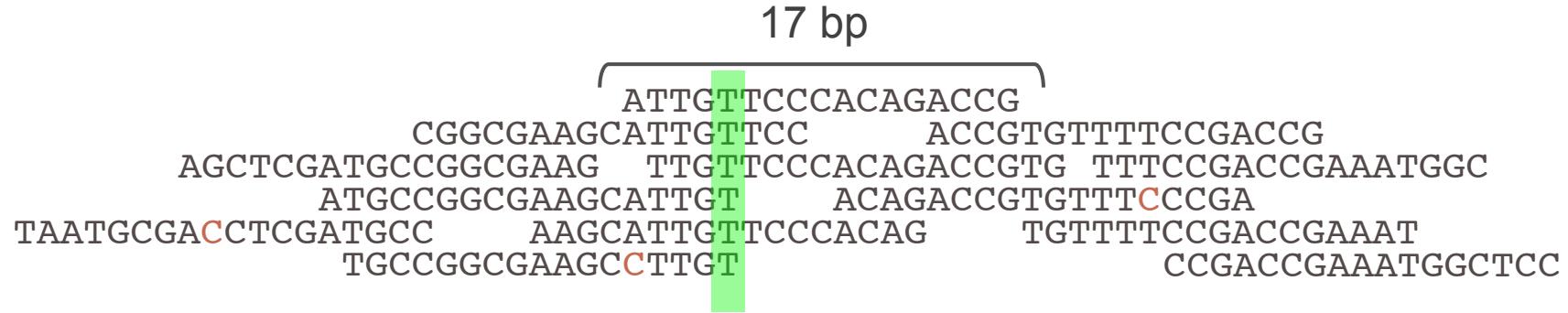
66 bp

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAACGCATTGTTCCCACAGACCGTGTTCGACCGAAATGGCTCC



6x coverage
100% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAACGCATTGTTCCCACAGACCGTGTTCGACCGAAATGGCTCC

17 bp

A sequence alignment showing multiple reads of the same 17 bp consensus sequence. The consensus sequence is highlighted in green. A green box encloses the first 17 bp of each read, which are identical to the consensus. The last two reads show some variation at the 18th position.

CGGCGAAGC	ATTGTTCC	ACCGTGTTCGACCG
AGCTCGATGCCGGCGAAG	TTGTTCCCACAGACCGTG	TTTCCGACCGAAATGGC
ATGCCGGCGAACG	ATTGT	ACAGACCGTGTTC
TAATGCGAC	CTCGATGCC	CCGA
TGCCGGCGAACG	CTTGT	TGTTTCCGACCGAAAT
		CCGACCGAAATGGCTCC

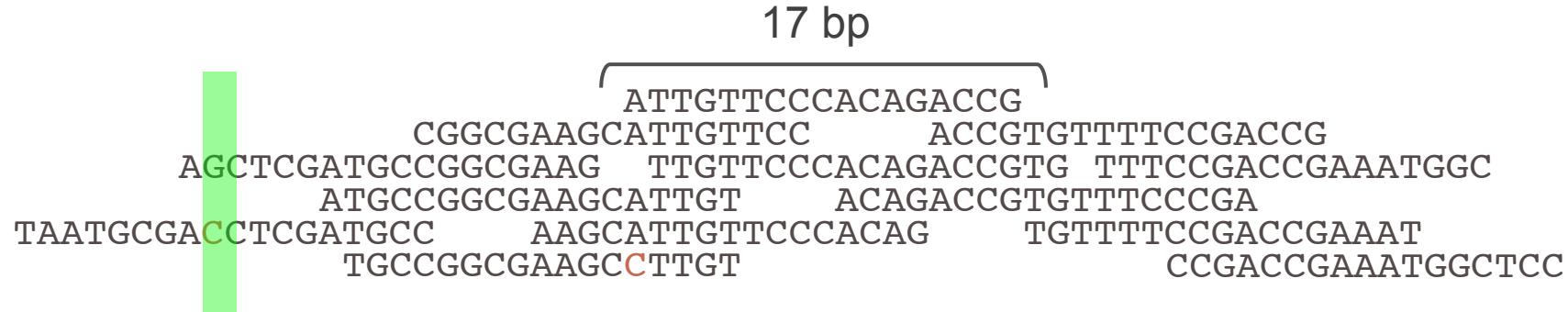
5x coverage
80% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAACGCATTGTTCCCACAGACCGTGTTCGACCGAAATGGCTCC



2x coverage
50% identity

Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

Sequenza consenso:

TAATGCGACCTCGATGCCGGCGAACGCATTGTTCCCACAGACCGTGTTCGACCGAAATGGCTCC

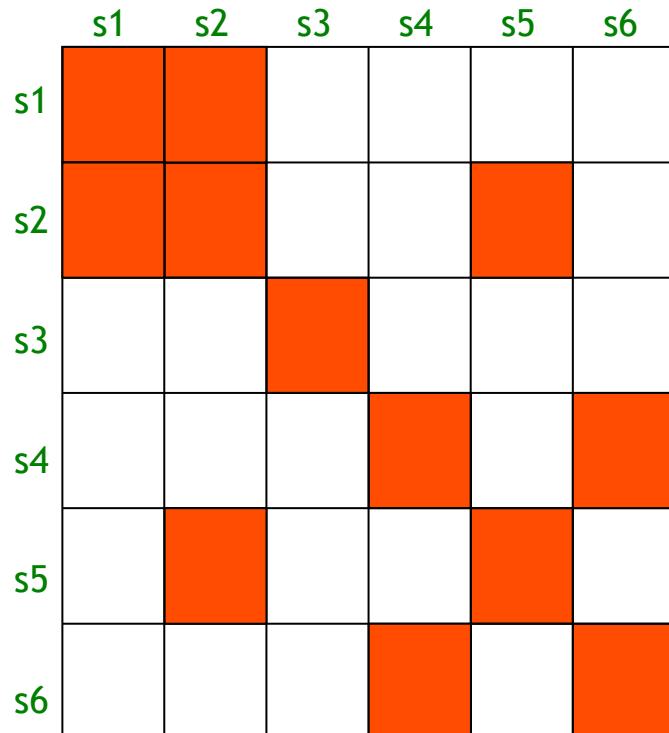
17 bp



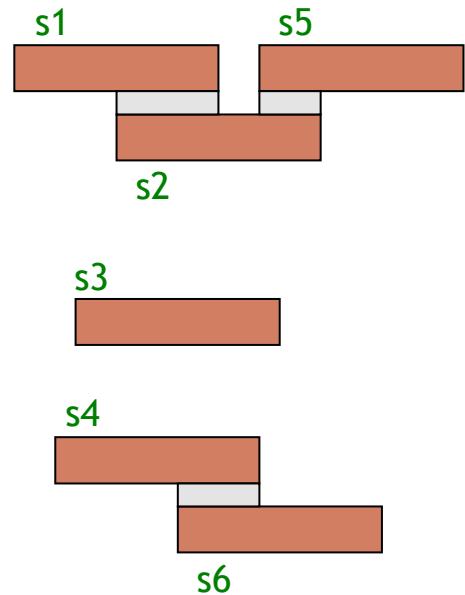
Coverage: # of reads supporting the consensus

DE NOVO ASSEMBLY

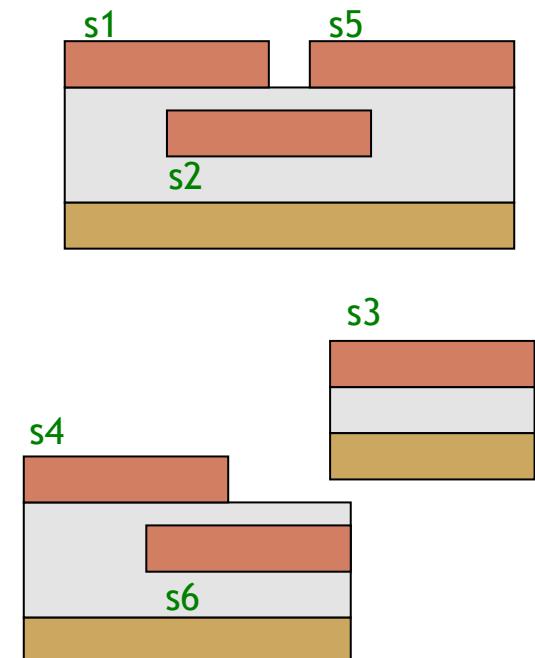
overlap



layout

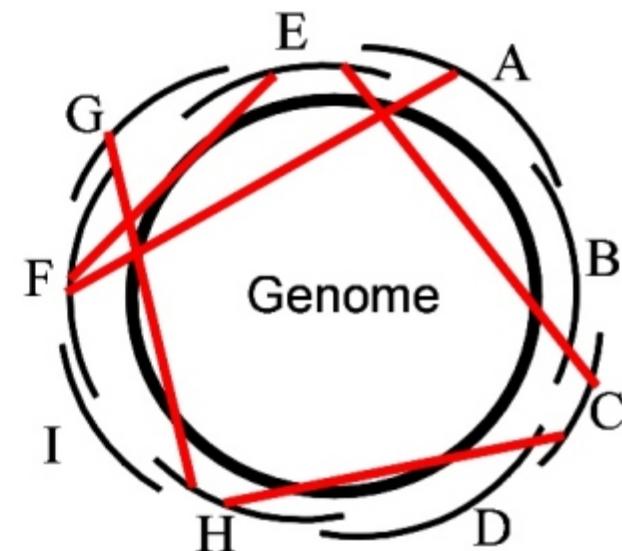
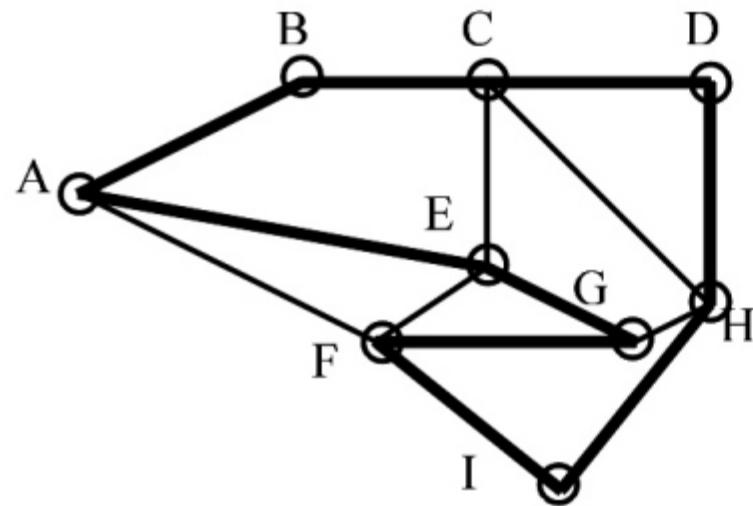


consensus

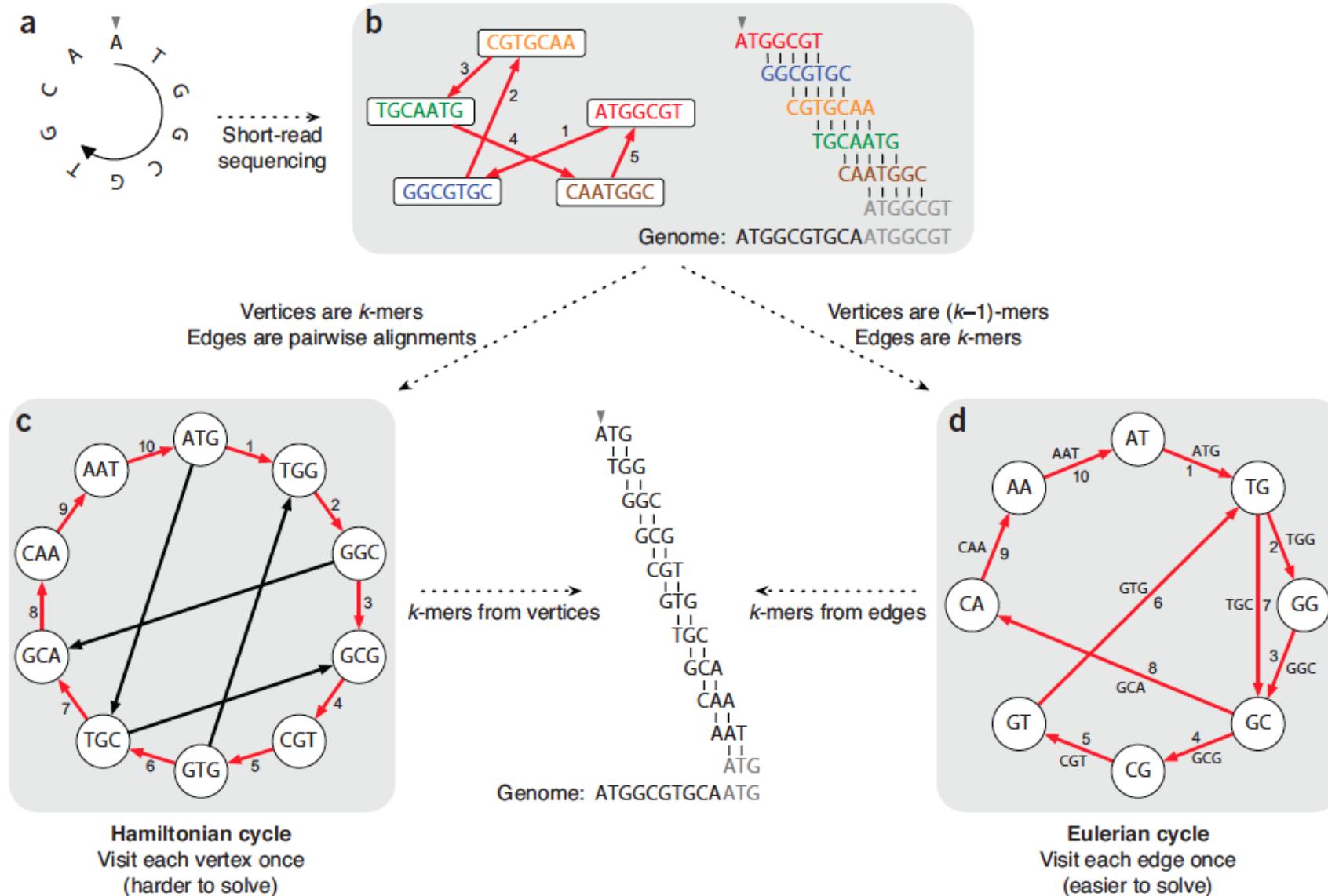


DE NOVO ASSEMBLY

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



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

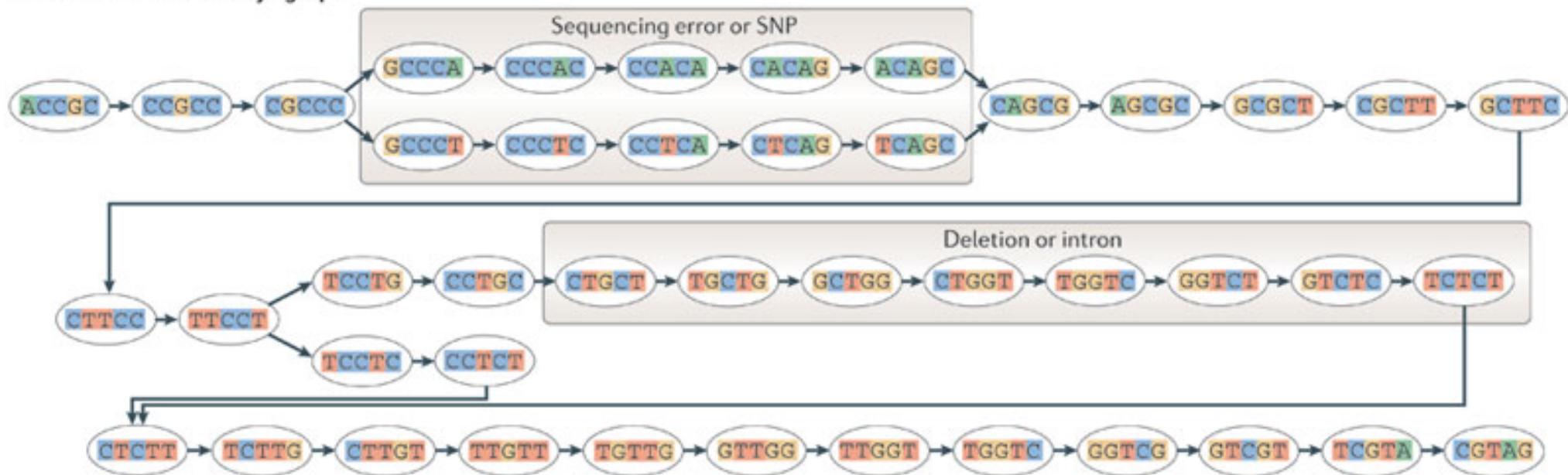
De novo assembly (De Bruijn graph construction)

a Generate all substrings of length k from the reads

ACAGC	TCCTG	GTCTC		AGCGC	CTCTT	GGTCG	
CACAG	TTCCT	GGTCT		CAGCG	CCTCT	TGGTC	
CCACA	CTTCC	TGGTC	TGTTG	TCAGC	TCCTC	TTGGT	
CCCAC	GCTTC	CTGGT	TTGTT	CTCAG	TTCCT	GTTGG	
GCCCC	CGCTT	GCTGG	CTTGT	CCTCA	CTTCC	TGTTG	
CGCCC	GCGCT	TGCTG	TCTTG	CCCTC	GCTTC	TTGTT	CGTAG
CCGCC	AGCGC	CTGCT	CTCTT	GCCCT	CGCTT	CTTGT	TCGTA
ACCGC	CAGCG	CCTGC	TCTCT	CGCCC	GCGCT	TCTTG	GTCGT
ACCGCCCCACAGCGCTTCCCTGCTGGTCTCTTGTGTTG				CGCCCTCAGCGCTTCCCTCTTGTGGTCGTAG			Reads

DE NOVO ASSEMBLY

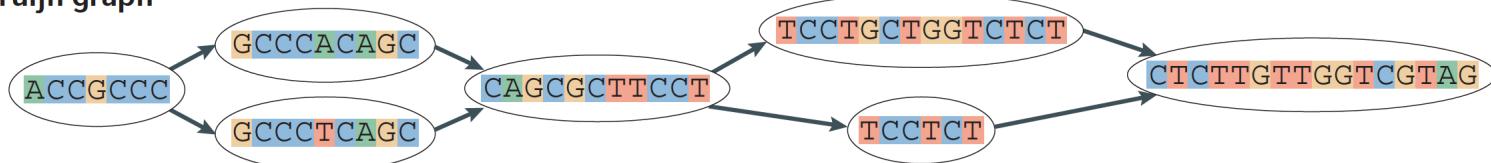
b Generate the De Bruijn graph



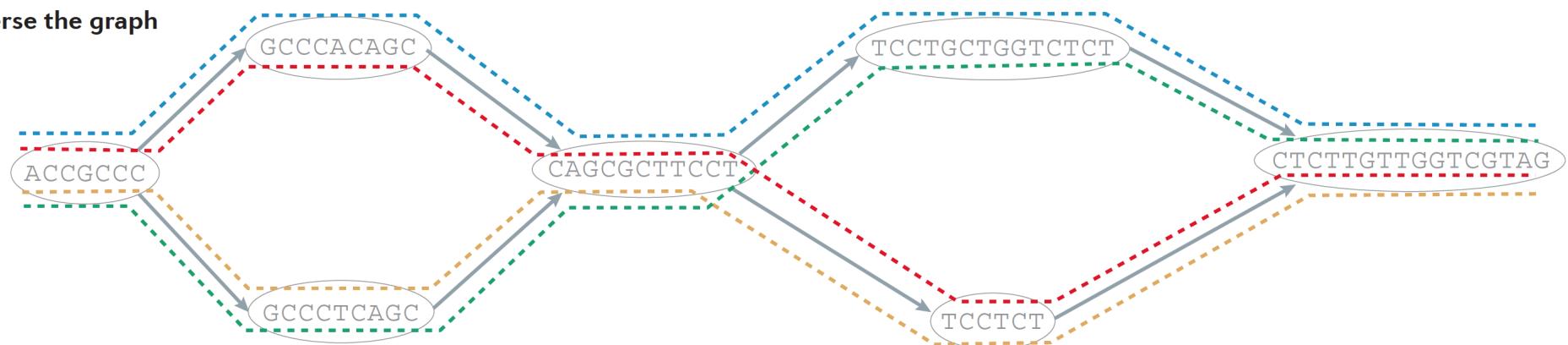
Branching paths can represent different splicing variants encoded by the gene

DE NOVO ASSEMBLY

c Collapse the De Bruijn graph



d Traverse the graph

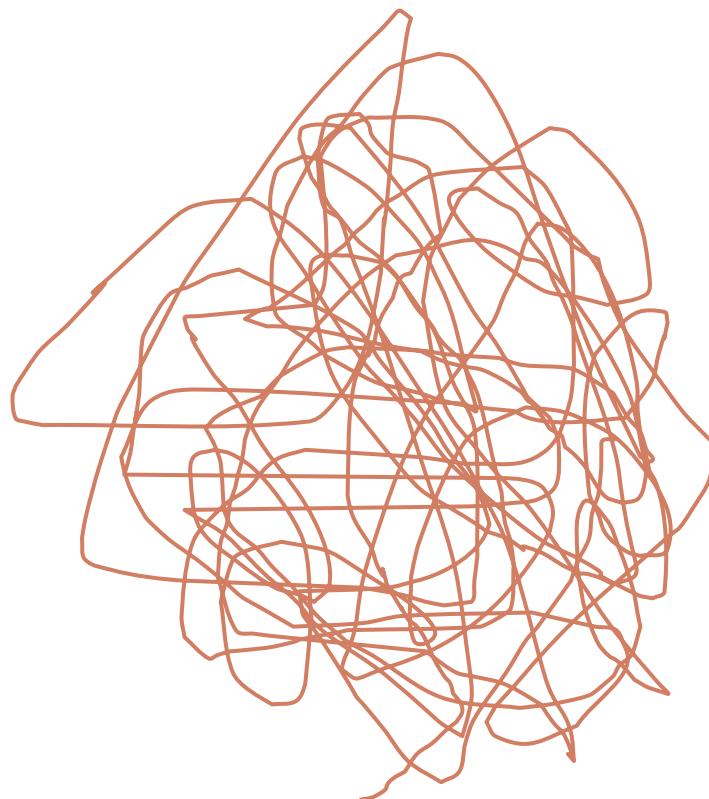


e Assembled isoforms

— ACCGCCACAGCGCTTCCTGCTGGTCTCTTGTGGTCGTAG
- - - ACCGCCACAGCGCTTCCT - - - CTTGTTGGTCGTAG
- - - ACCGCCCTCAGCGCTTCCT - - - CTTGTTGGTCGTAG
- - - ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTGGTCGTAG

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

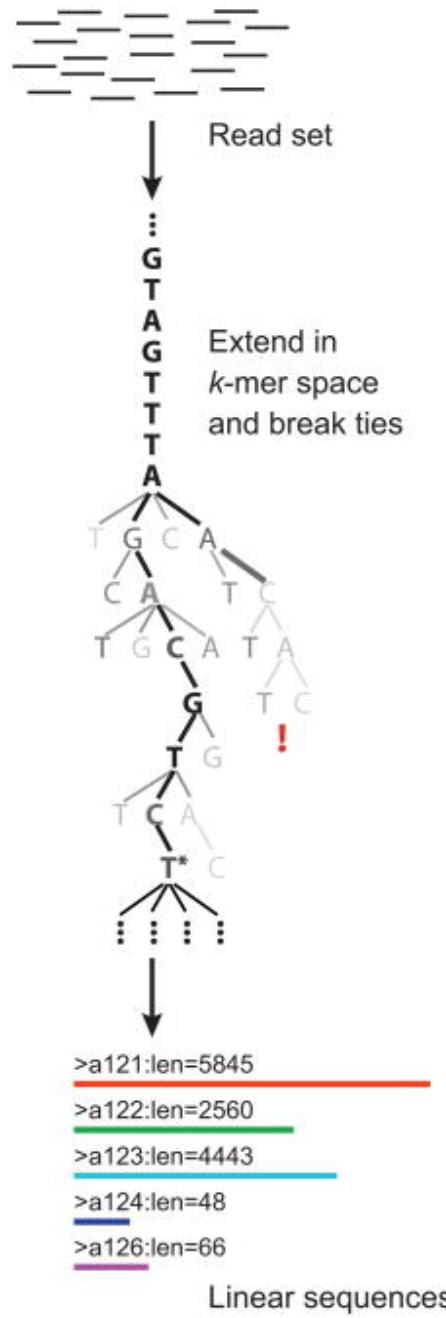


de-Brujin
graphs

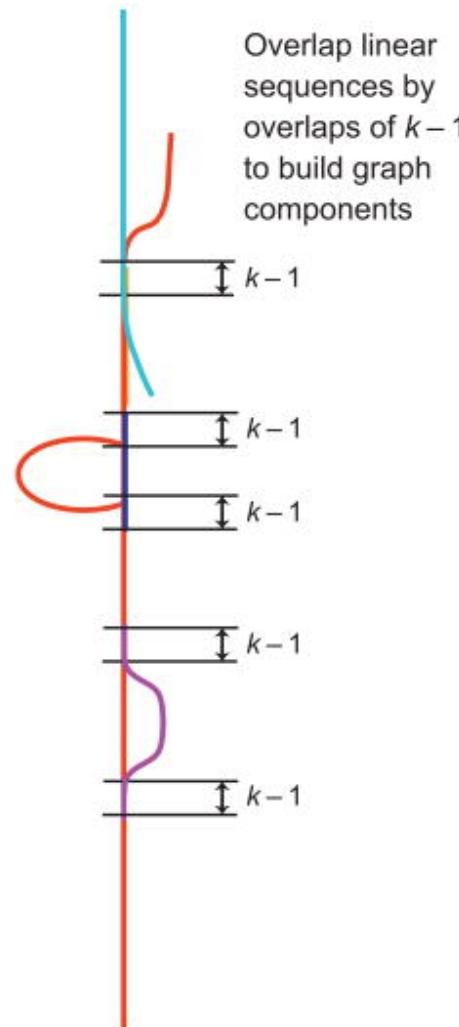
Transcripts
+
Isoforms



a



b



c

