CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

"Algorithms for Genomic Data"

Fall 2015

Smith College

Instructor: Prof. Sara Sheehan

Outline: 9/16

- Overlap graph assembly
- de Bruijn graph assembly

Assembly vocabulary

- **Long read**: a fragment that has been "read" from a genomic sequence (DNA for us), usually > 1000 bp
- **Short read**: same as a long read but usually < 1000 bp
- **Paired-end read**: both ends of a fragment are "read", but the portion between them is unknown
- **bp**: base pair
- **kb, Mb, Gb**: kilo bases 10^3, mega bases 10^6, giga bases 10^9

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
 GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
  GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
   GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
     GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
      GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
        GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
          GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
          GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
           GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
              GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
                GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
                GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
                  GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

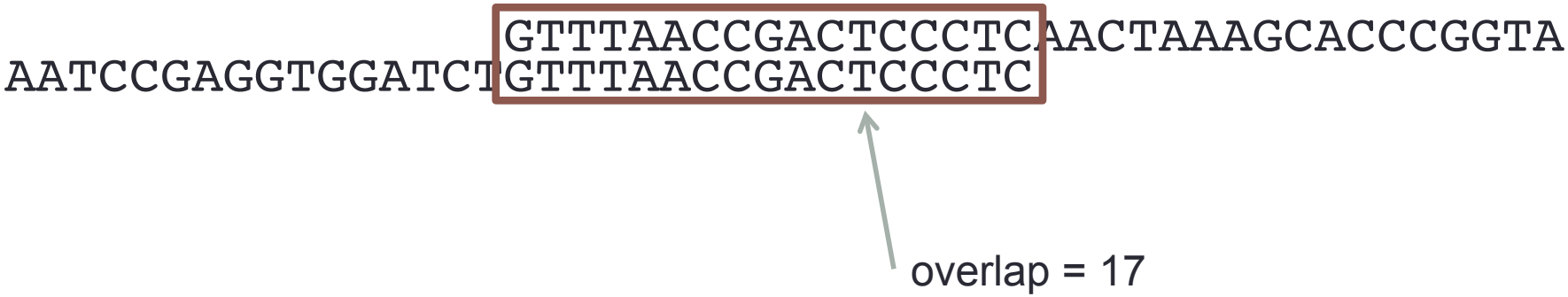
Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

```
                    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

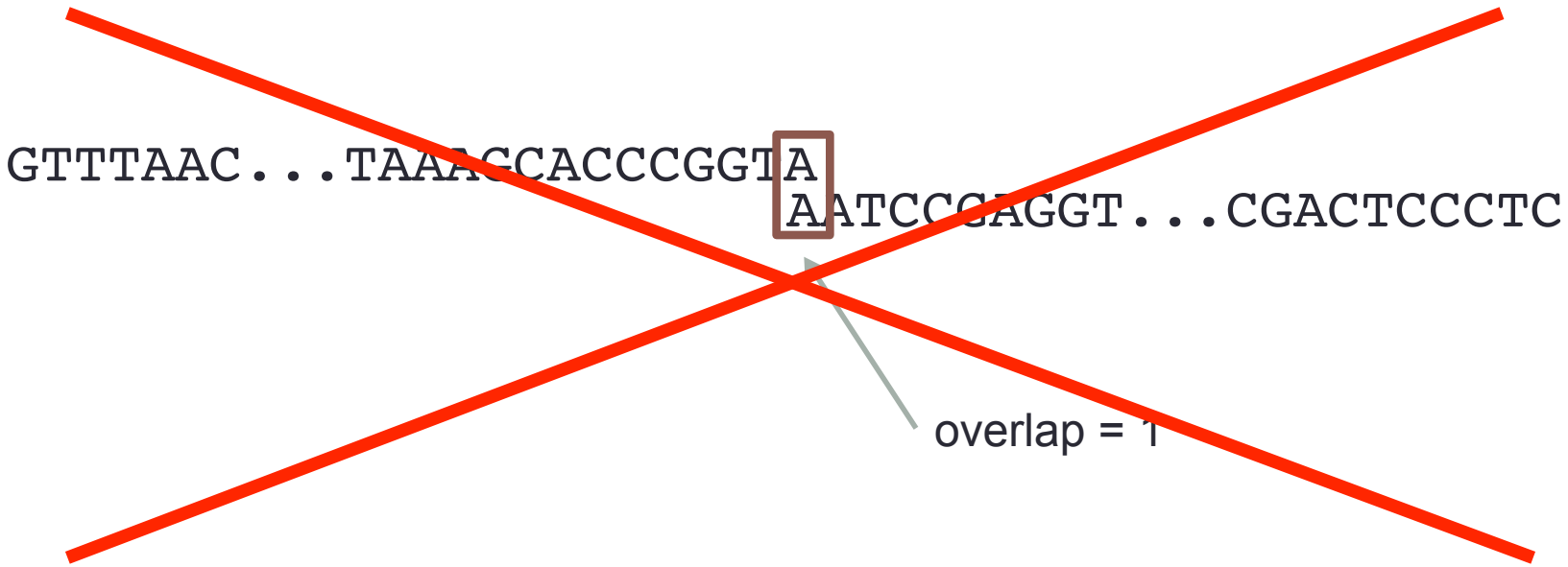
Overlap graph assembly

```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```

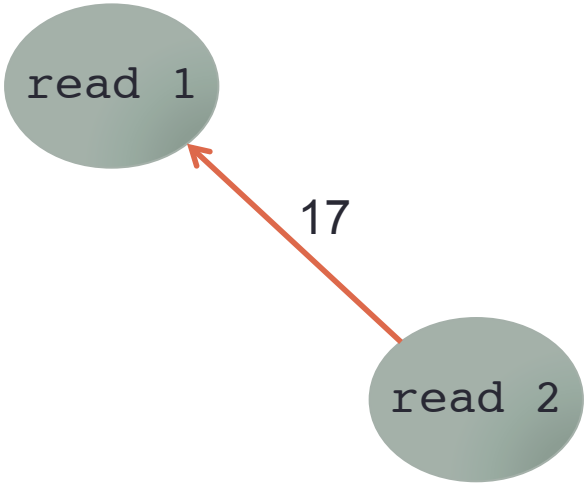


Overlap graph assembly

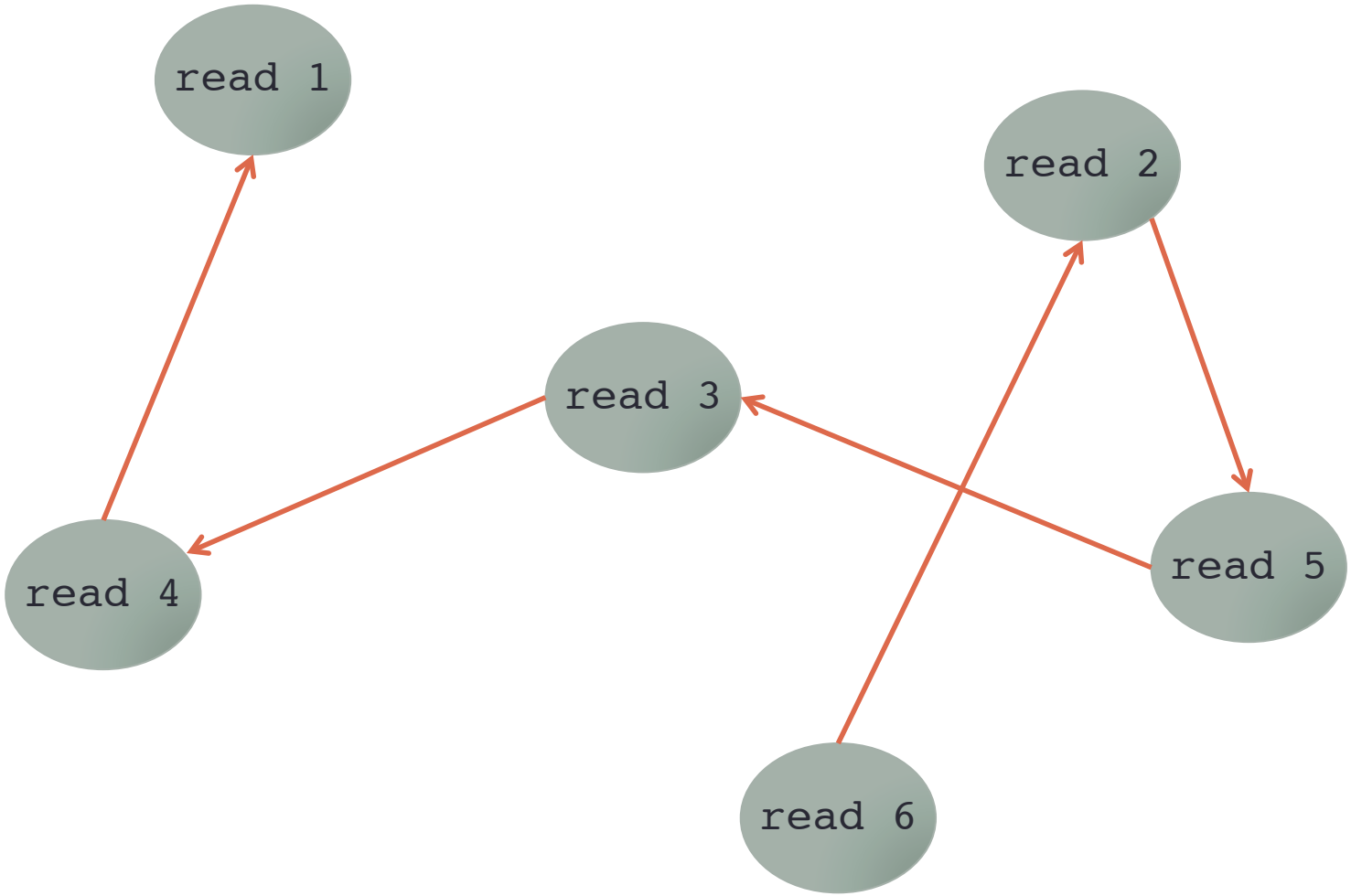
```
read 1:    GTTTAACCGACTCCCTCAACTAAAGCACCCGGTA
read 2:    AATCCGAGGTGGATCTGTTTAACCGACTCCCTC
```



Overlap graph

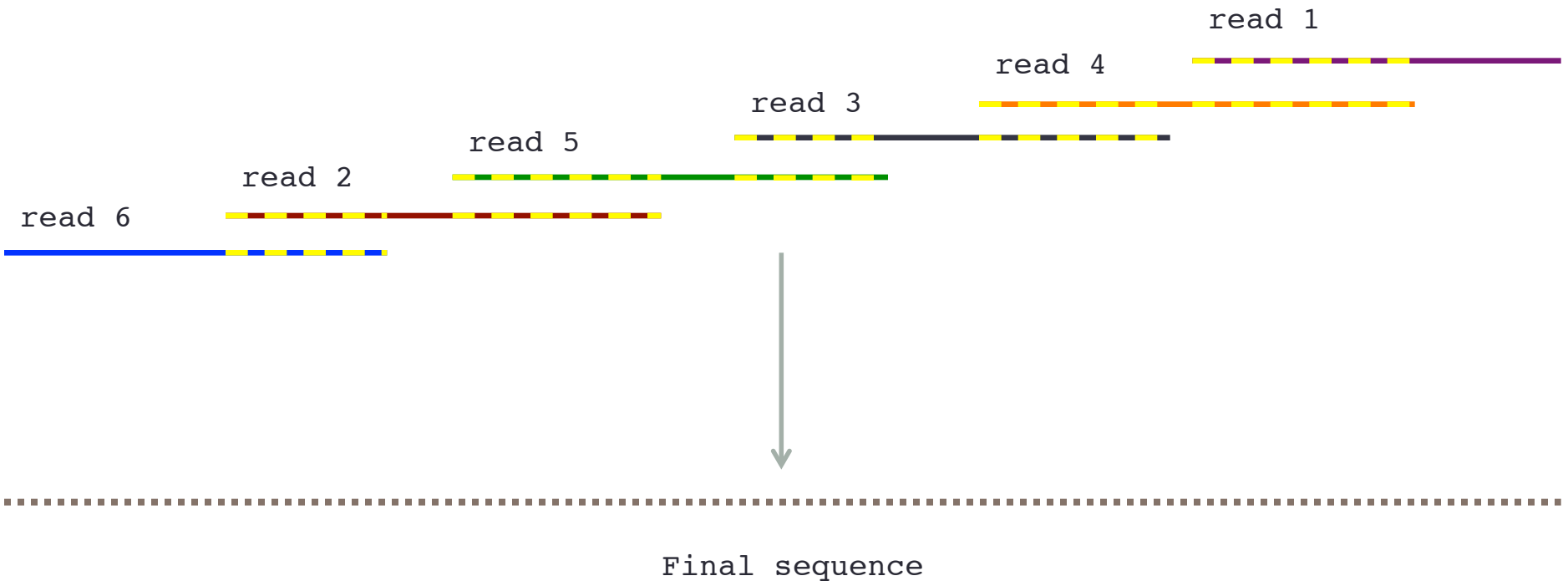


Overlap graph

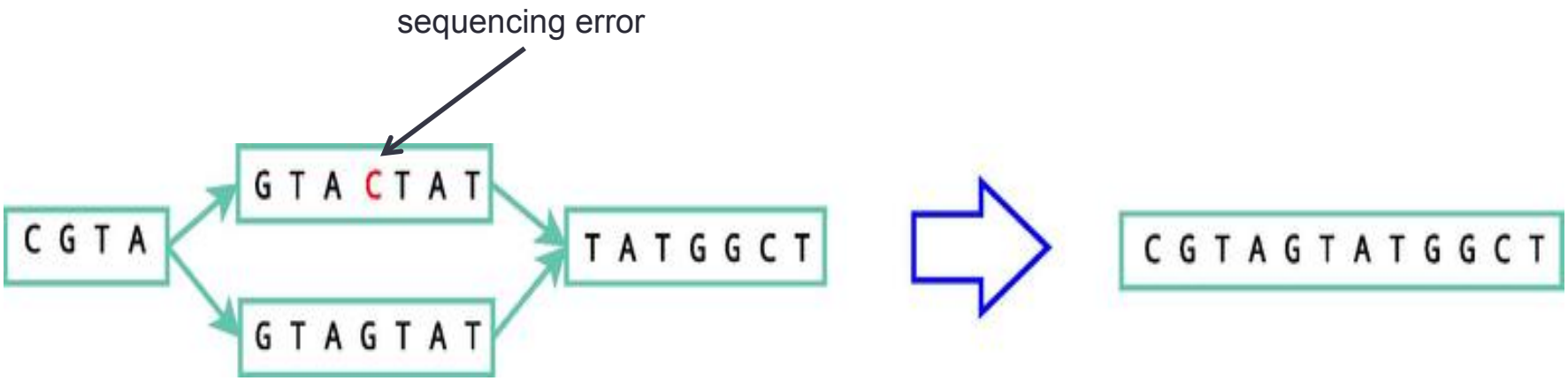


What is the runtime for creating the overlap graph?

Perfect graph traversal



Bubbles

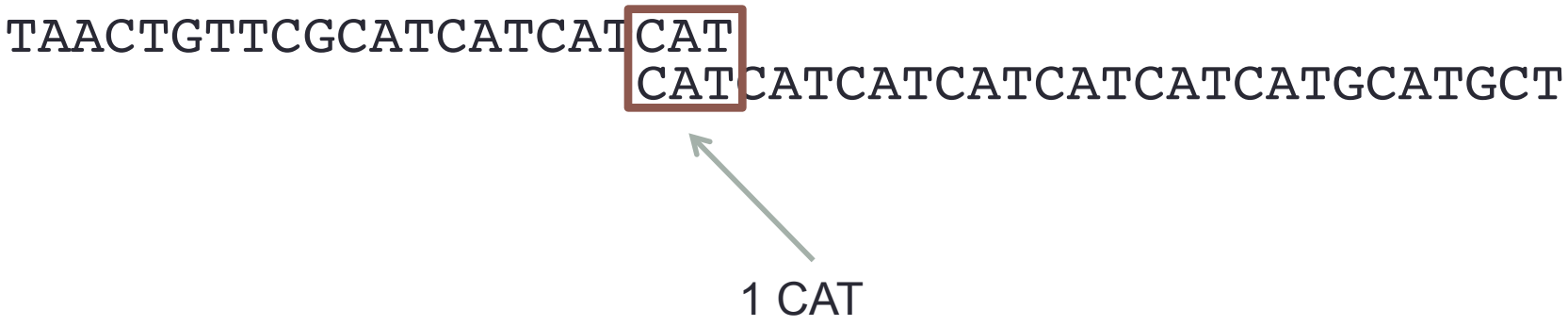


Velvet Assembler, Wikipedia

Repeats, example 2

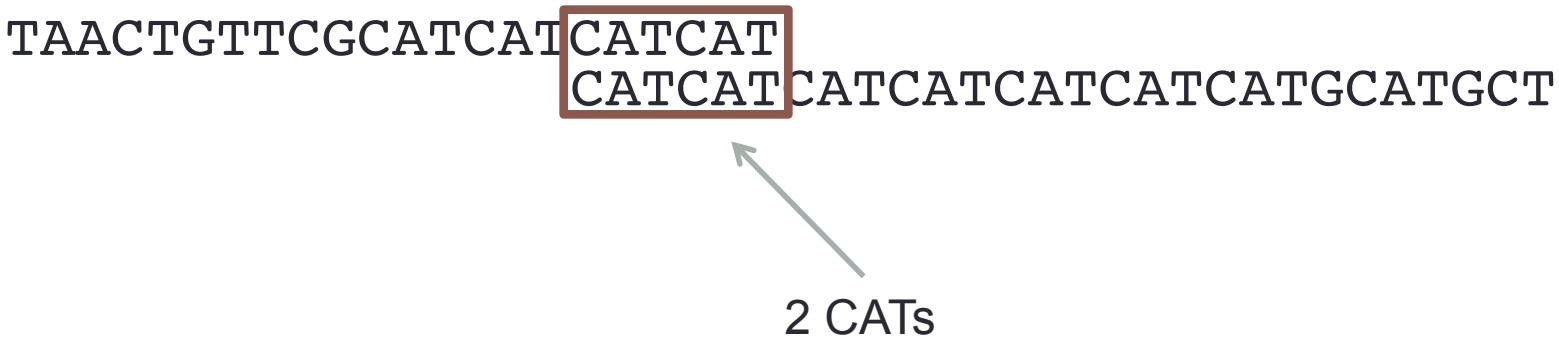
Overlap graph assembly

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```



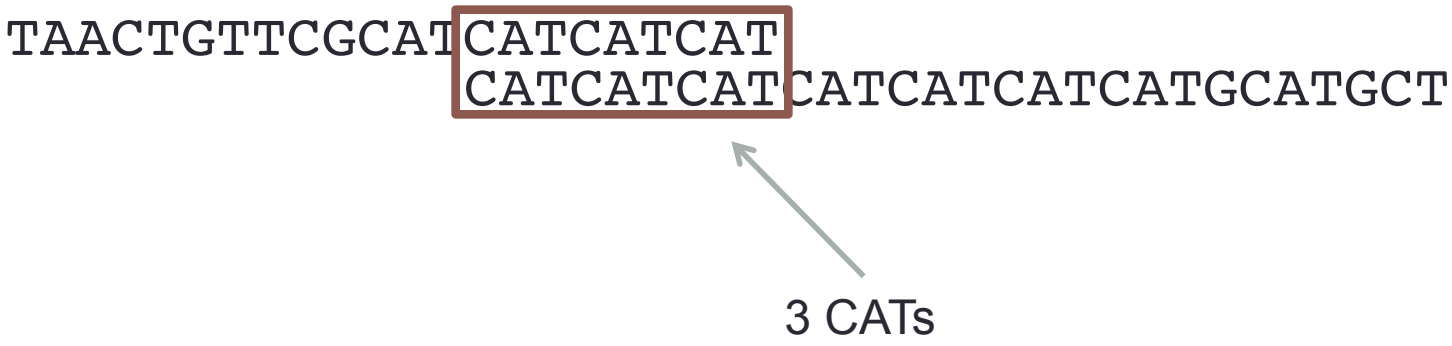
Overlap graph assembly

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```



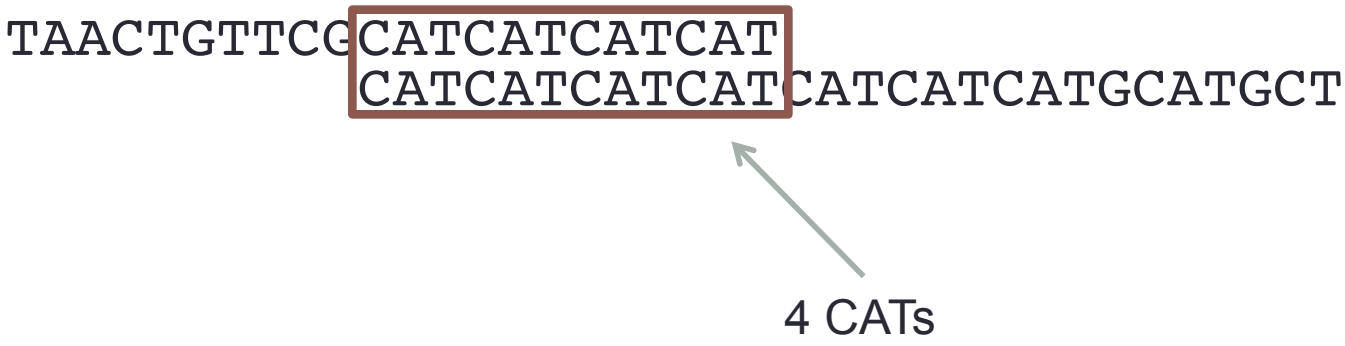
Overlap graph assembly

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```



Overlap graph assembly

```
read 1:    TAACTGTTCGCATCATCATCAT
read 2:    CATCATCATCATCATCATCATGCATGCT
```



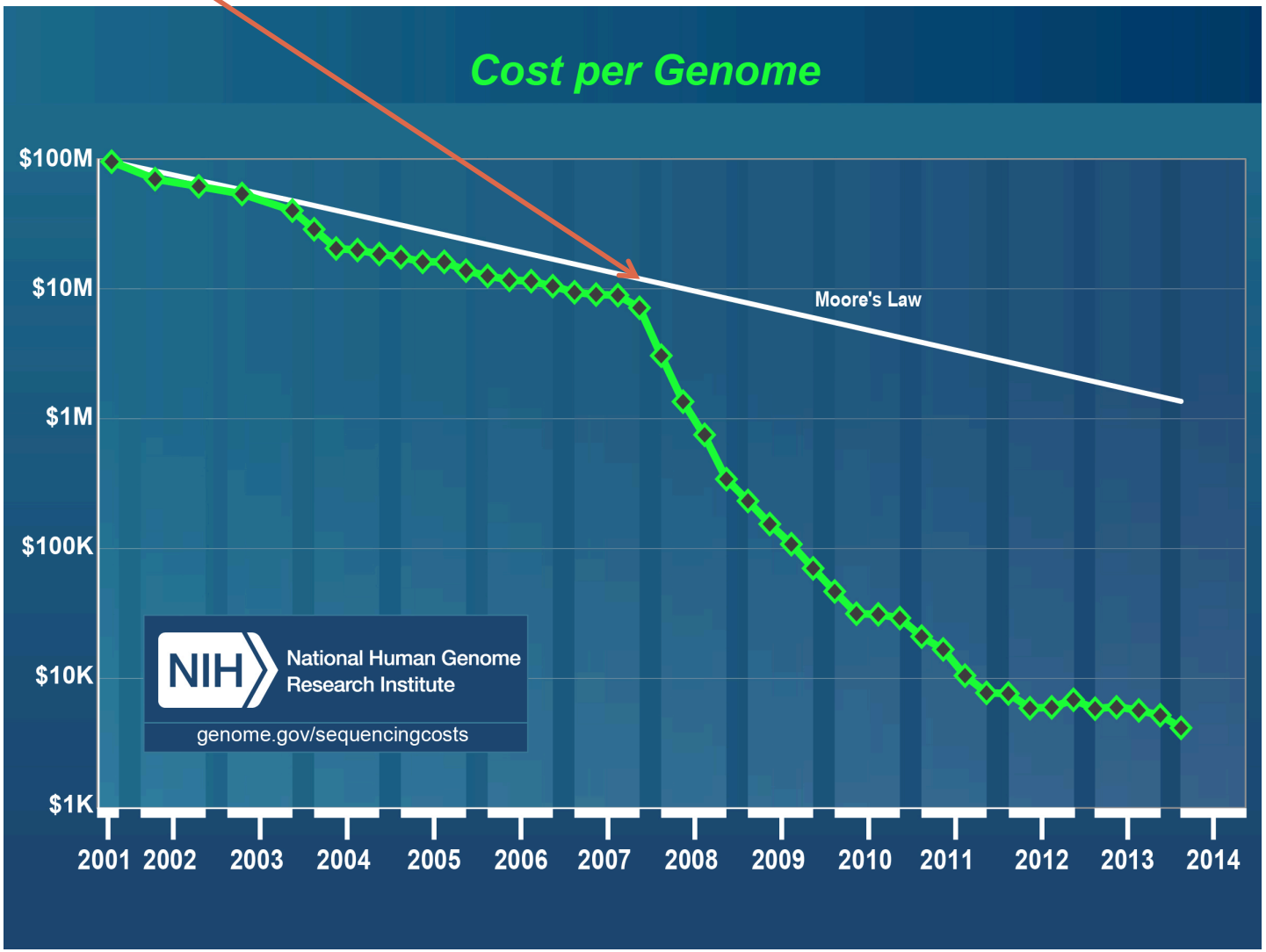
Repeats

- Hardest part of assembly
- Sometimes cannot be resolved
- **Contigs**: assembled sequence, separated by unassembled regions

Assembly evaluation

- **N50**: for a set of contigs, N50 is the length such that at least half the bases of the assembly are in a contig with length N50 or longer, and at least half the bases are in a contig with length N50 or shorter.

Long -> Short reads



Assembly vocabulary (cont)

- **Coverage**: total number of bases in all reads, divided by the length of the genome (short reads: need higher coverage)

- **k-mer**: a genomic “word” of length *k*

de Bruijn graphs