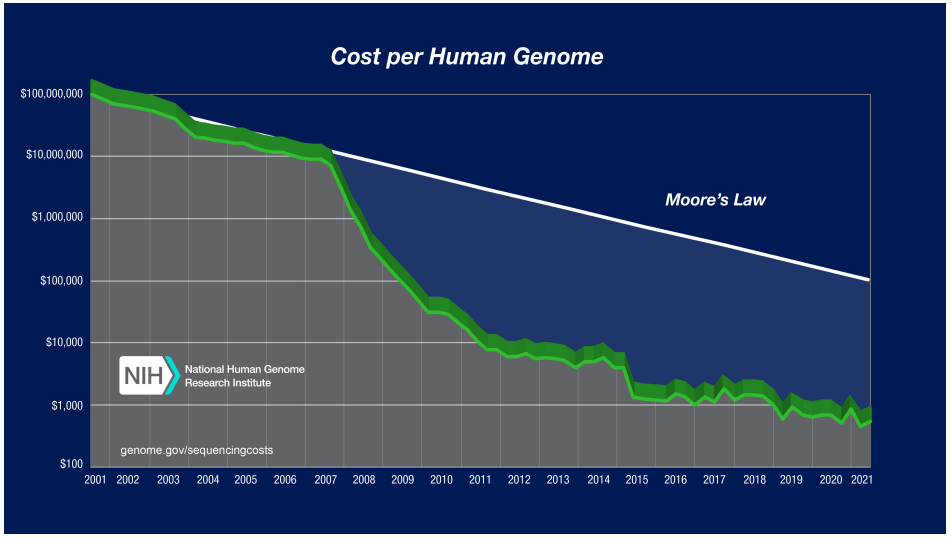
Genome Assembly



High Throughput Sequencing

Example applications:

- Sequencing a genome (DNA)
- Sequencing a transcriptome and gene expression studies (RNA)
- ChIP (chromatin immunoprecipitation)

Example platforms:

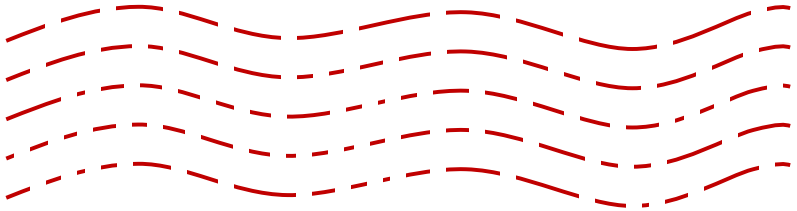
- 454
- Illumina
- SOLiD

Sequencing Output

- Hundreds of millions of sequencing reads, each ~200 nts in length
- We need to re-assemble the genome from these sequencing reads

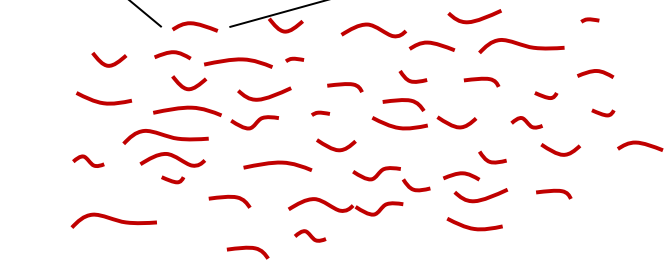
```
- ACGTAGTCGTAGTCGCTTACGATTCGTATGCGTACGTGTAGTCTACGTGTA
- CCGCGCGTACTCTCGATGTACGTGTCAGTTACGATGTACTGTAGTCAGTGT
- TTTGATCGTAGTGTCACTGAGCAAACACCATTACTTACTATCTTGGACATC
- TGGGGCGATCGAGGATTCTAGTTATCGAGTGTCGGCGATTATGCGATCGAA
- GGCATCATACGTAGCGTATCGATTAGCAGTGCCGCGCTATACGTCGTCGAT
- AGCGCGGGTCTGGACGACGGCTAGCGTCATCGTCGATCGATCGATCGTAAA
- TCGGTCTCGATCGAGTGTCTGCGCGTCTCTGAGAGGCTAGGTAGAGAGCTG
- CTCCTCCGAGCGTAGTCGCGATACGTATCGGATCTGACGTATGCAGTCGAT
- CCGTATGCGTAGTCGCGATCGATCGTGATTTGAGTCGCTAGATTACGTAGT
- TACGGCGACGTTACGCGTAGTCATGCATGCATGAGCCTAGTTGCGTATGCG
- GAAGCGTACATGCTAGTCTACGTCTCGCAGTTCGTATGCGATCGTATGACG
- ACGCACGATCGATCTATGCATCGATGCTAGTCGTATGCTAGTCTATGCGAT
- ACGTAGTCGTAGTCGCTTACGATTCGTATGCGTACGTGTAGTCTACGTGTA
- CCGCGCGTACTCTCGATGTACGTGTCAGTTACGATGTACTGTAGTCAGTGT
- TTTGATCGTAGTGTCACTGAGCAAACACCATTACTTACTATCTTGGACATC
- TGGGGCGATCGAGGATTCTAGTTATCGAGTGTCGGCGATTATGCGATCGAA
- GGCATCATACGTAGCGTATCGATTAGCAGTGCCGCGCTATACGTCGTCGAT
- AGCGCGGGTCTGGACGACGGCTAGCGTCATCGTCGATCGATCGATCGTAAA
- TCGGTCTCGATCGAGTGTCTGCGCGTCTCTGAGAGGCTAGGTAGAGAGCTG
- CTCCTCCGAGCGTAGTCGCGATACGTATCGGATCTGACGTATGCAGTCGAT
- CCGTATGCGTAGTCGCGATCGATCGTGATTTGAGTCGCTAGATTACGTAGT
- TACGGCGACGTTACGCGTAGTCATGCATGCATGAGCCTAGTTGCGTATGCG
- GAAGCGTACATGCTAGTCTACGTCTCGCAGTTCGTATGCGATCGTATGACG
- ACGCACGATCGATCTATGCATCGATGCTAGTCGTATGCTAGTCTATGCGAT
- GTGCGTAGTCGTATATGCCTAGCATCGTTAGTCTAGCGTAGTCGTACGTAC
- ATCATCGGCGATAGTCTACGTAGTTATATCTACGCGGCGCCCATTCTGCAA
- CGAGTAGGAGTCGTAGTCGTAGTCGATGCATCGAGTCTGCGATAGTCGTAG
- GACAGTGCGGACGTATGTATAGCAGCAGTTACGTACGTAGTTGACGTAGTA
- CAACTTTCGGCATCTTCCGGTTCTCTCTCTCTAGATAGAGACTTACGATCG
- TCCGCGCATCGTAGTCGCGCGCGACTATGCGCGACACACGTGTAGCATACC
- AGACTGAGTATATCGGCGCGTATGCGACTGTAGCTATATACGGCATGCGTC
- GGATCGAGTCATCGTCGATCGAGTCGTAGCAGCAGTCGTAGCTAATAGTCG
- GCGATCGTAGTCAGTTCGAGTCATGCGATCTGATGCGAGTACGTAGTCGTA
- TTATTATCGCGCAGTGTCGTACGTAGTCGTAGTCATCGATGCGTATATACT
- AGTACGTAGTATCTGAGCGTCTCTTCTACGGACGATCGATCGTACACGTGA
- TTTATTACGACGATACGTGCGCATTGCGTATGCAGTGTCAGTCAGTCGTAG
- GTGCGTAGTCGTATATGCCTAGCATCGTTAGTCTAGCGTAGTCGTACGTAC
- ATCATCGGCGATAGTCTACGTAGTTATATCTACGCGGCGCCCATTCTGCAA
- CGAGTAGGAGTCGTAGTCGTAGTCGATGCATCGAGTCTGCGATAGTCGTAG
- GACAGTGCGGACGTATGTATAGCAGCAGTTACGTACGTAGTTGACGTAGTA
- CAACTTTCGGCATCTTCCGGTTCTCTCTCTCTAGATAGAGACTTACGATCG
- TCCGCGCATCGTAGTCGCGCGCGACTATGCGCGACACACGTGTAGCATACC
- AGACTGAGTATATCGGCGCGTATGCGACTGTAGCTATATACGGCATGCGTC
- GGATCGAGTCATCGTCGATCGAGTCGTAGCAGCAGTCGTAGCTAATAGTCG
- GCGATCGTAGTCAGTTCGAGTCATGCGATCTGATGCGAGTACGTAGTCGTA
- TTATTATCGCGCAGTGTCGTACGTAGTCGTAGTCATCGATGCGTATATACT
- AGTACGTAGTATCTGAGCGTCTCTTCTACGGACGATCGATCGTACACGTGA
- TTTATTACGACGATACGTGCGCATTGCGTATGCAGTGTCAGTCAGTCGTAG
```

DNA Sequencing



DNA Sequencing

CGTAGTAGTCACAGTCTACGTATATGGGCTCAGCATATAGCGTATAGCGGACTTAGCCATCGTA



DNA Sequencing

>CGTAGTAGTCACAGTCTACGTATATGGGCTCAGCATATAGCGTATAGCGGACTTAGCCATCG
>GCGTATAGTCTATATACGACTTATCGGCTCGGTCGCAGAGCAGATATATGCAGTTATATGCTAG
>CCTACGTTATATCGATACTACTAGCTCGTCATGAGCGAGTAGATAGTATGACGAGCGACGATC
...
>CGATATTAGCCTAGCATCATTACGGCGAGACTCTCGGCTCGCTATATAGCGCTATAGCGAT
>CGGCTATAGCGCATATGCTCAGTAGCTATTAGCAGTATTACGATTATAGTCTCGGCGCATTAC
>TTTCGCGGGATAAGTCTTCGTCTTATGCGACGATTATACGGCCGTATATTTGCATTTAGCATTT
>GGCGTATGGCGGATATCGGCGCGTCATAGCGCCGATTAGGCTACGCCGATGCATCG
>CGCGATCGCGCGCGATCGCGTCAGTCGCGCAGTAGCGCGGCATAGTCGTATCGGCGCCG
>TGACAGAAGCTATAAGAGTCAGTAGATCTGAGTATTAGCATTATCGGCGCGATGCGCGATAACG
>GCGTATAGTCTATATACGACTTATCGGCTCGGTCGCAGAGCAGATATATGCAGTTATATGCTA
>CGCGATCGCGCGCGATCGCGTCAGTCGCGCAGTAGCGCGGCATAGTCGTATCGGCGCCGATCGC
>ATAGCAGCACGTGATAGGATATGCTGCTCGTTCGACTATCATATCTGCGTCTGGTCTAGCA
>TGACAGAAGCTATAAGAGTCAGTAGATCTGAGTATTAGCATTATCGGCGCGATGCGCGA
>CCTACGTTATATCGATACTACTAGCTCGTCATGAGCGAGTAGATAGTATGACGAGCGACGATCC
>CGTAGTAGTCACAGTCTACGTATATGGGCTCAGCATATAGCGTATAGCGGACTTAGCCATCG
>TTTCGCGGGATAAGTCTTCGTCTTATGCGACGATTATACGGCCGTATATTTGCATTTAGCATTT
>GGCGTATGGCGGATATCGGCGCGTCATAGCGCCGATTAGGCTACGCCGATGCATCGTCGAGTA

Assembly

>GGGCAGGC
>GGCTAGGG
>TAGGGCA

Assembly

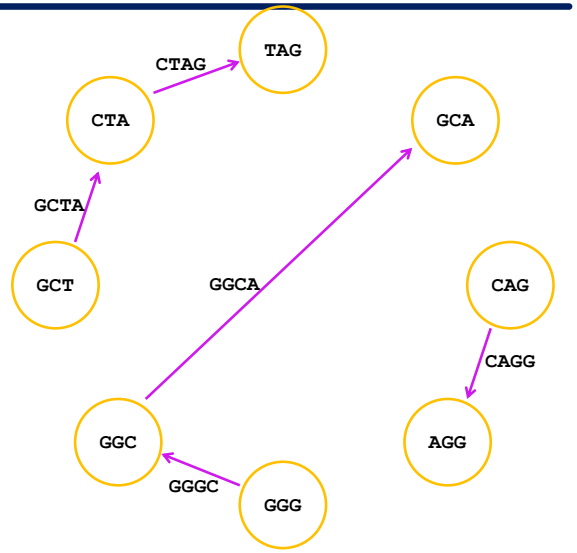
>GGGCAGGC
>GGCTAGGG
>TAGGGCA

CTAG
GGGC
CAGG
GGCA
GCTA
AGGC
TAGG
AGGG
GCAG
GGCT
GGGC
TAGG
GGGC
AGGG

Assembly

>GGGCAGGC
>GGCTAGGG
>TAGGGCA

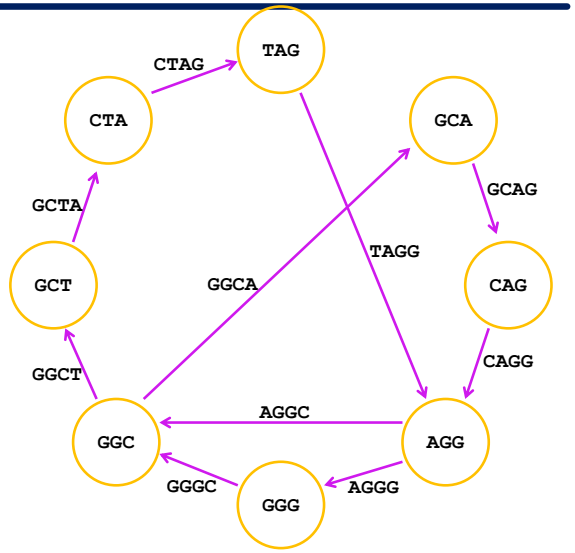
CTAG
GGGC
CAGG
GGCA
GCTA
AGGC
TAGG
AGGG
GCAG
GGCT
GGGC
TAGG
GGGC
AGGG



Assembly

>GGGCAGGC
>GGCTAGGG
>TAGGGCA

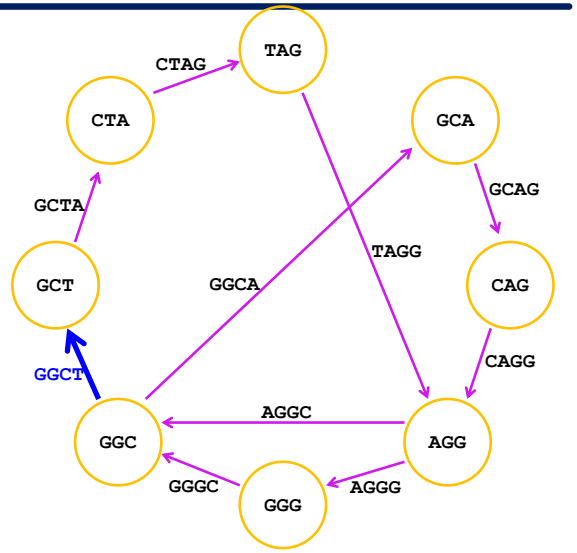
CTAG
GGGC
CAGG
GGCA
GCTA
AGGC
TAGG
AGGG
GCAG
GGCT
GGGC
TAGG
GGGC
AGGG



Assembly

GGCT

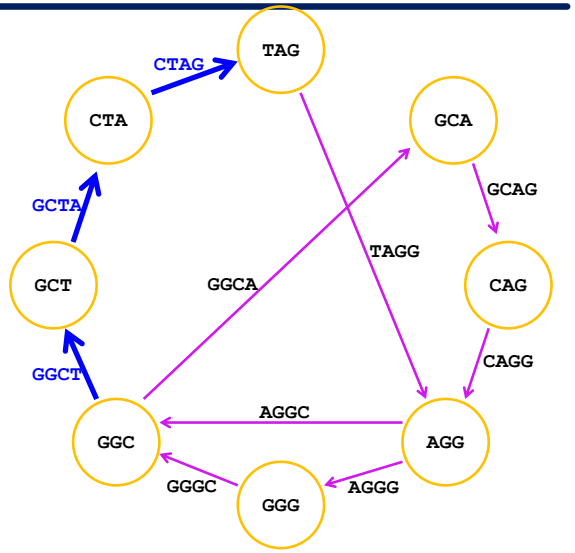
>GGGCAGGC
>GGCTAGGG
>TAGGGCA



Assembly

GGCTAG

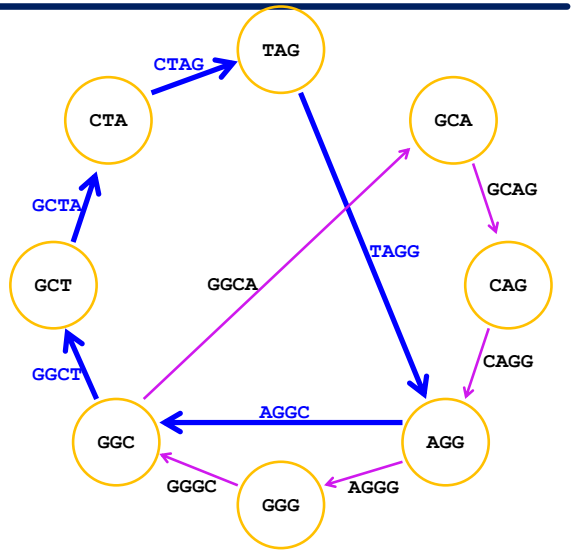
>GGGCAGGC
>GGCTAGGG
>TAGGGCA



Assembly

GGCTAGGC

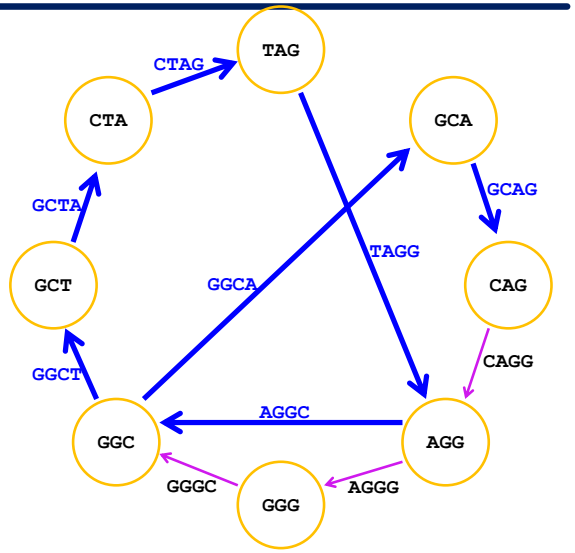
>GGGCAGGC
>GGCTAGGG
>TAGGGCA



Assembly

GGCTAGGCAG

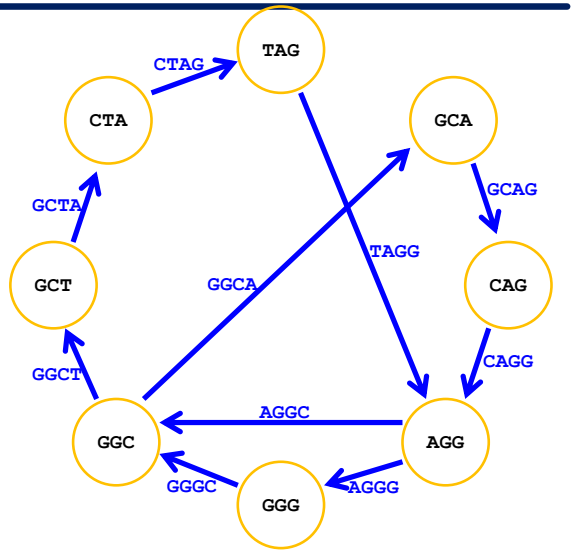
>GGGCAGGC
>GGCTAGGG
>TAGGGCA



Assembly

GGCTAGGCAGGGC

>GGGCAGGC
>GGCTAGGG
>TAGGGCA



Challenges

- Repeats
- Sequencing errors

Implementation

deBruijn graph can be implemented with a hash table

- Each entry in hash table corresponds to an edge in the graph (each key is a *k*-mer and each value is the number of occurrences of the *k*-mer).
- Nodes are stored implicitly.

Implementation

GGCT

Assembly corresponds to Euler path through graph

- Genome sequence starts with any *k*-mer (edge in the graph)
- Repeatedly extend genome sequence forward, one nucleotide at a time, until no further extensions are possible
  - The genome sequence is extended and a nucleotide added to its end if there exists a nucleotide (A, C, G, or T) that can be added to the end of the *k*-1 final nucleotides of the genome sequence to form a *k*-mer that is an edge in the graph.
  - If there are multiple individual nucleotides that can be added to the final *k*-1 nucleotides in the genome sequence to form *k*-mer edges in the graph, then the nucleotide resulting in the *k*-mer edge with the largest number of occurrences is chosen.
  - Each time the genome sequence is extended by a nucleotide, the corresponding *k*-mer edge is removed from the graph.

Implementation

GGCT

Assembly corresponds to Euler path through graph

- Genome sequence starts with any *k*-mer (edge in the graph)
- Repeatedly extend genome sequence backward, one nucleotide at a time, until no further extensions are possible
  - The genome sequence is extended and a nucleotide added to its front if there exists a nucleotide (A, C, G, or T) that can be added to the front of the *k*-1 first nucleotides of the genome sequence to form a *k*-mer that is an edge in the graph.
  - If there are multiple individual nucleotides that can be prepended to the first *k*-1 nucleotides in the genome sequence to form *k*-mer edges in the graph, then the nucleotide resulting in the *k*-mer edge with the largest number of occurrences is chosen.
  - Each time the genome sequence is extended by a nucleotide, the corresponding *k*-mer edge is removed from the graph.