Week 2: Genome assembly & annotation

- Genome assembly
  - Overlap layout consensus
  - de Bruijn graphs
- Genome annotation
  - Hidden Markov Models

Day 06: Genome assembly

- Genome assembly
  - Overlap layout consensus
  - de Bruijn graphs

Why sequence the genome?

- Determine the "complete" sequence of a haploid genome.
  - Previously "snippets" of the genome were available.
- Identify the sequence and location of every protein coding gene.
- Use as a "map" to then track the location and frequency of genetic variation.
- Unravel the genetic architecture of inherited and somatic traits/diseases.
- To understand genome and species evolution.

Sequencing of the human genome

"All the News That's Fit to Print"

The New York Times

Late Edition
New York: **Today**, afternoon thunderstorms, high 88. **Tonight**, showers end, low 67. **Tomorrow**, partly cloudy with showers late, high 81. **Yesterday**, high 88, low 74. Weather map, Page D8.

VOL. CXLIX . . No. 51,432 | NEW YORK, TUESDAY, JUNE 27, 2000 | 75 CENTS

$1 beyond the greater New York metropolitan area.

Genetic Code of Human Life Is Cracked by Scientists

JUSTICES REAFFIRM MIRANDA RULE, 7-2; A PART OF 'CULTURE'

By LINDA GREENHOUSE

WASHINGTON, June 26 — The Supreme Court reaffirmed the Miranda decision today by a 7-to-2 vote that erased a shadow over one of the most famous rulings of modern times and acknowledged that the Miranda warnings "have become part of our national culture."

The court said in an opinion by Chief Justice William H. Rehnquist that because the 1966 Miranda decision "announced a constitutional rule," a statute by which Congress had sought to overrule the decision was itself unconstitutional.

Miranda had appeared to be in jeopardy, both because of that long-ignored but recently rediscovered law, by which Congress had tried to overrule Miranda 32 years ago, and because of the court's perceived hostility to the original decision.

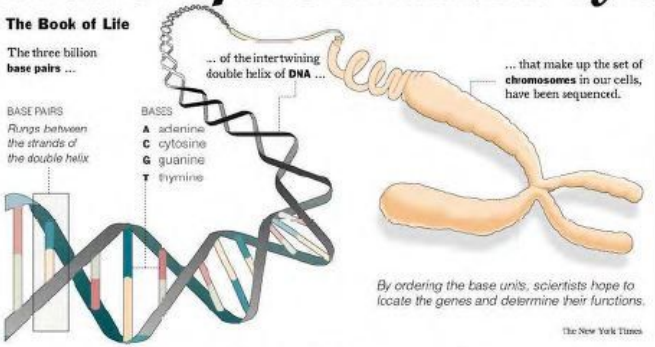
The chief justice said, though, that the 1968 law, which replaced the Miranda warnings with a case-by-case test of whether a confession was voluntary, could be upheld only if the Supreme Court decided to overturn Miranda. But with Miranda having

Justices Antonin Scalia and Clarence Thomas cast the dissenting votes.

The decision overturned a ruling last year by the federal appeals court in Richmond, Va., which held that Congress was entitled to the last word because Miranda's presumption that a confession was not voluntary unless preceded by the warnings was not required by the Constitution.

The decision today — only 14 pages long, in Chief Justice Rehnquist's typically spare style — brought an abrupt end to one of the odder episodes in the court's recent history, an intense and strangely delayed relighting of a previous generation's battle over the rights of criminal suspects. Miranda v. Arizona was a hallmark of the Warren Court, and Chief Justice Rehnquist, despite his record as an early and tenacious critic of the decision, evidently did not want its repudiation to be an imprint of his own tenure.

There was considerable drama in the courtroom today as the chief justice announced that he would de-



Science Times
A special issue

- Putting the genome to work.
- Some information has already paid research dividends.
- Two research methods, two results.
- From Mendel to helix to genome.
- More articles, charts and photos of the genome effort.

Section F

Francis S. Collins, head of the Human Genome Project, left, with J. Craig Venter, head of Celera Genomics, after the announcement yesterday that they had finished the first survey of the human genome.

Paul Hosefros/The New York Times

A SHARED SUCCESS

2 Rivals' Announcement Marks New Medical Era, Risks and All

By NICHOLAS WADE

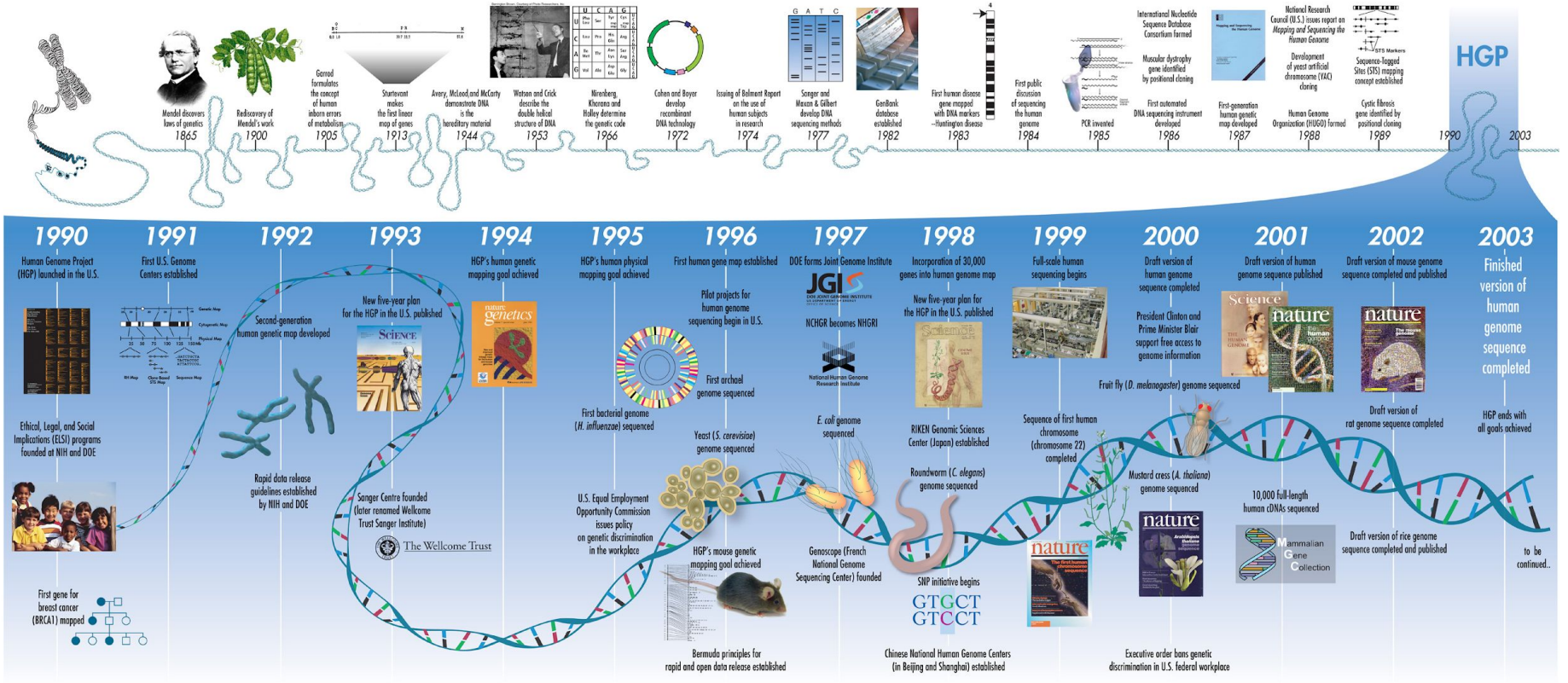
WASHINGTON, June 26 — In an achievement that represents a pinnacle of human self-knowledge, two rival groups of scientists said today that they had deciphered the hereditary script, the set of instructions that defines the human organism.

"Today we are learning the language in which God created life," President Clinton said at a White House ceremony attended by members of the two teams, Dr. James D. Watson, co-discoverer of the structure of DNA, and, via satellite, Prime Minister Tony Blair of Britain. [Excerpts, Page D8.]

The teams' leaders, Dr. J. Craig Venter, president of Celera Genomics, and Dr. Francis S. Collins, director of the National Human Genome Research Institute, praised each other's contributions and signaled a spirit of cooperation from now on, even though the two efforts will remain firmly independent.

The human genome, the ancient script that has now been deciphered, consists of two sets of 23 giant DNA

Genome sequencing has a long history



https://www.flickr.com/photos/genomegov/26964377742/

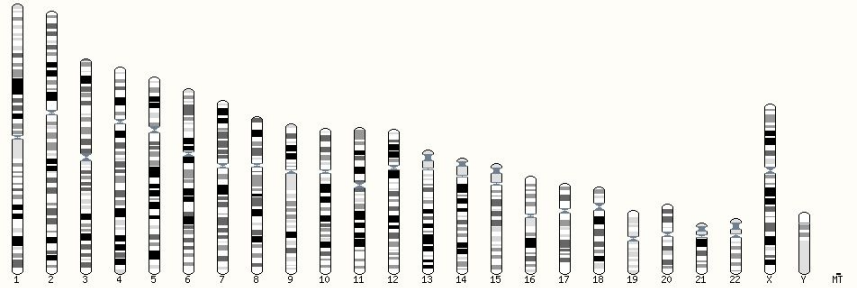
Genome sequencing technologies

Technology	Read length (bp)	Error rate	Native paired-end read support
ABI/Solid	75	Low (~2%)	Yes
Illumina/Solexa	100–150	Low (<2%)	Yes
IonTorrent	~200	Medium (~4%)*	No
Roche/454	400–600	Medium (~4%)*	No
Sanger	Up to ~2,000 bp	Low (~2%)	Yes
Pacific Biosciences	Up to ~15,000‡	High (~18%)	Yes (in strobe read mode)
Oxford nanopore	Up to ~20,000	High (~12%)	

Nagarajan & Pop (2013)

Genome sequence build

The Human Genome – Summary



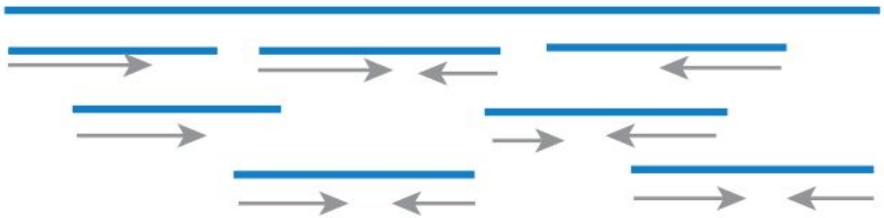
Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28, Dec 2013
Base Pairs	4,537,931,177
Golden Path Length	3,096,649,726
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2019
Database version	98.38
Gencode version	GENCODE 32

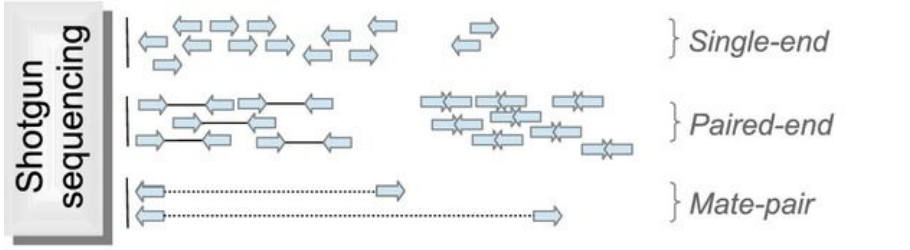
Gene counts (Primary assembly)

Coding genes	20,444 (incl 667 readthrough)
Non coding genes	23,949
Small non coding genes	4,871
Long non coding genes	16,857 (incl 304 readthrough)
Misc non coding genes	2,221
Pseudogenes	15,214 (incl 8 readthrough)
Gene transcripts	227,530

http://useast.ensembl.org/Homo_sapiens/Location/Genome

Genome assembly & annotation – Overview





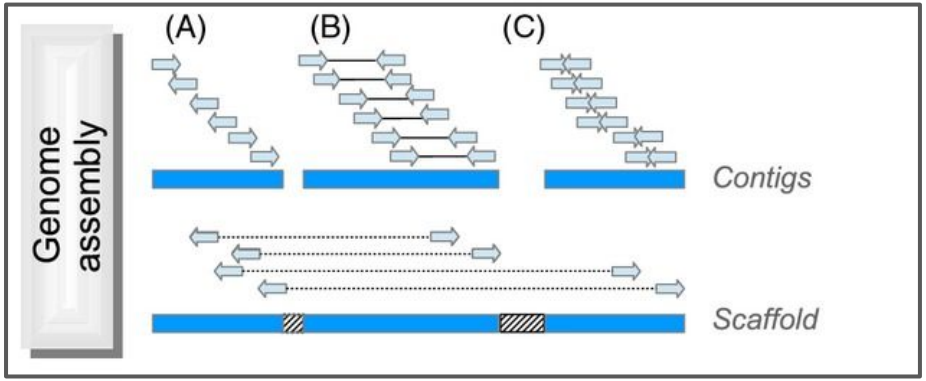
read: a short/long word that comes out of sequencer

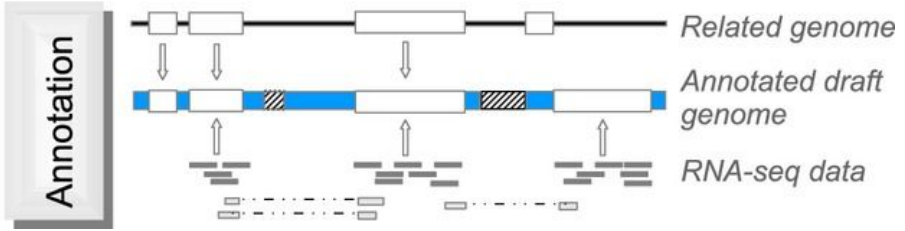
mate pair: a pair of reads from two ends of the same insert fragment

contig: a contiguous sequence formed by several overlapping reads with no gaps

scaffold: an ordered and oriented set of contigs, usually by mate pairs

consensus sequence: derived from the sequence multiple alignment of reads in a contig





Ekblom & Wolf (2014)

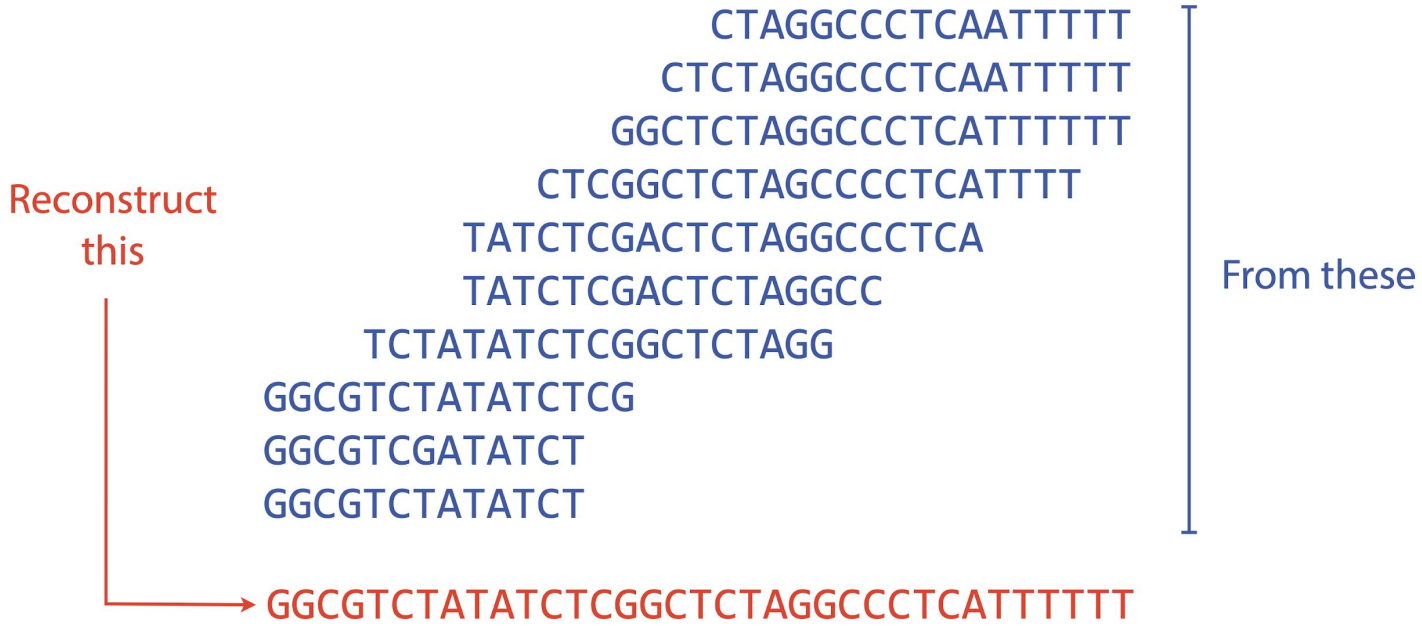
Genome assembly - the problem

Input: GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

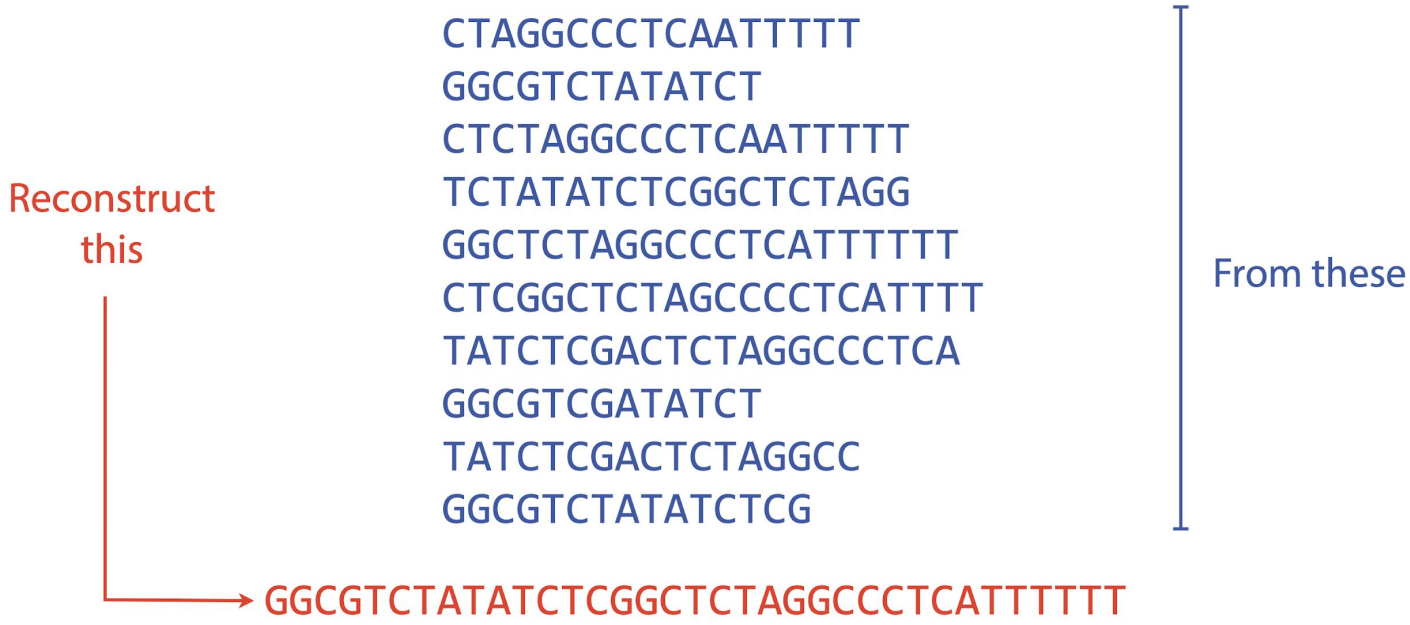
Copy:
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Fragment:
GGCGTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
GGC GTCTATAT CTCGGCTCTAGGCCCTCA TTTTTT
GGCGTC TATATCT CGGCTCTAGGCCCT CATTTTTT
GGCGTCTAT ATCTCGGCTCTAG GCCCTCA TTTTTT

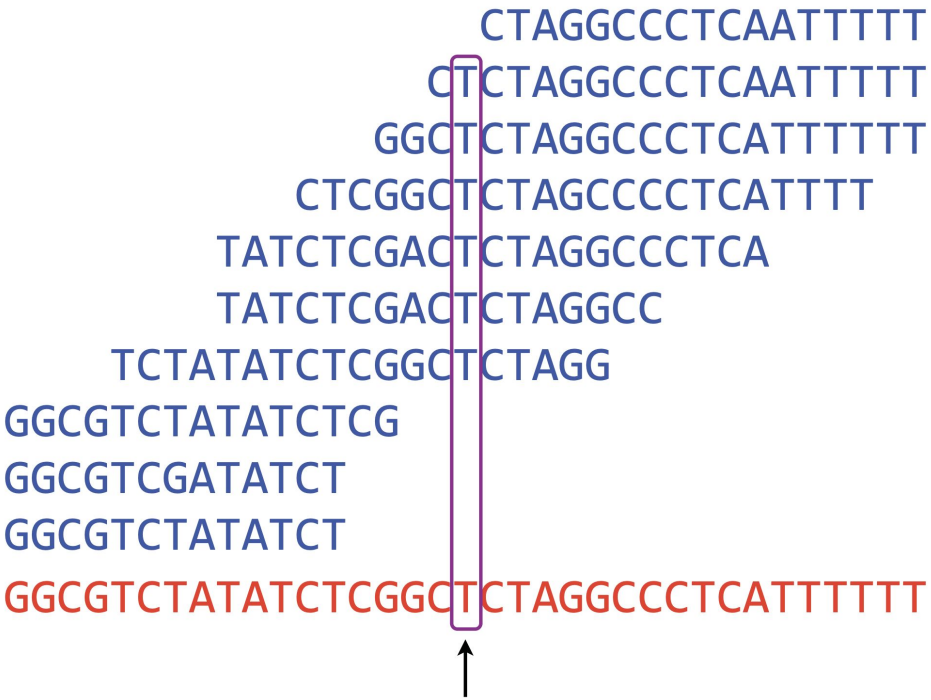
Genome assembly - the problem



Genome assembly - the problem



Genome assembly - the problem



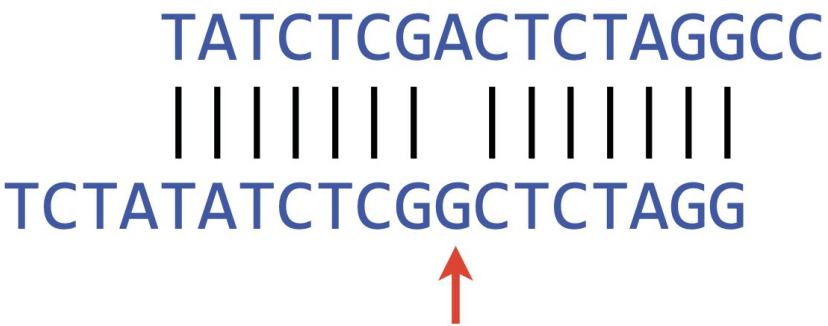
177 nucleotides in total

Genome is 35 nucleotides

Average coverage = 177 / 35 ~ 7x

Genome assembly - the problem

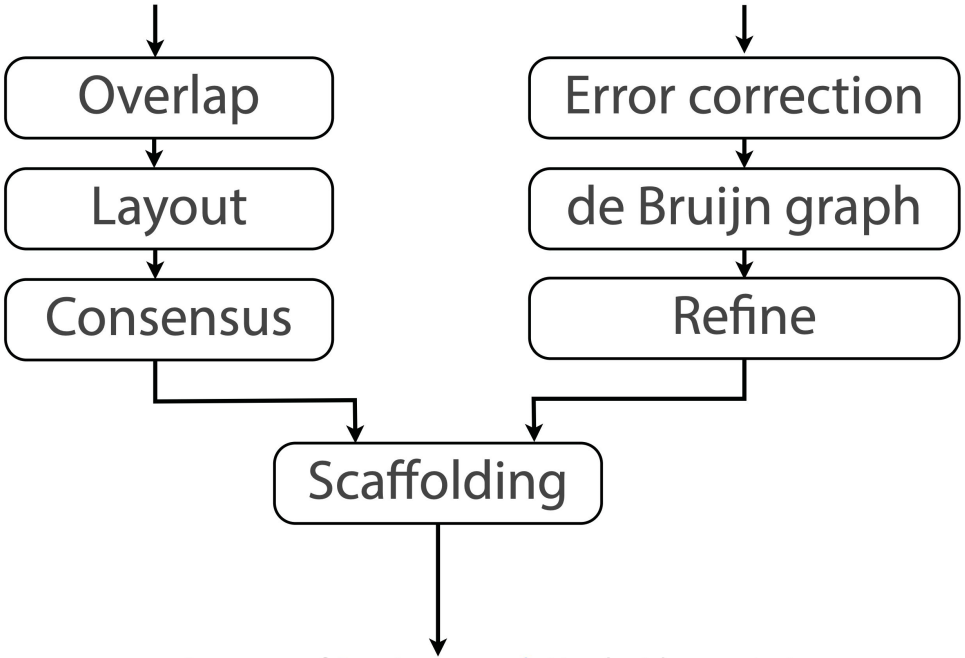
Say two reads truly originate from overlapping stretches of the genome.
Why might there be differences?



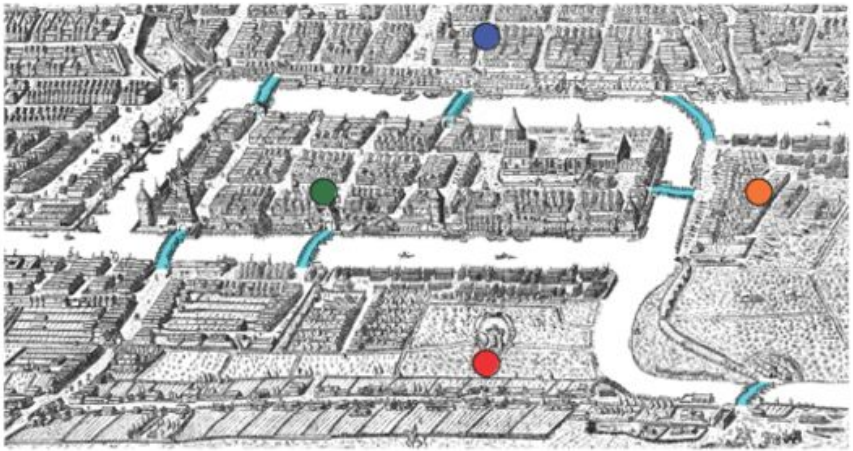
1. Sequencing error
2. Ploidy

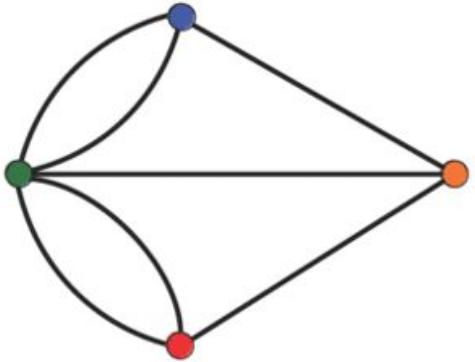
Genome assembly - two major approaches

1. Overlap-Layout-Consensus (OLC) assembly
2. de Bruijn graph (DBG) assembly



Introduction to graph theory





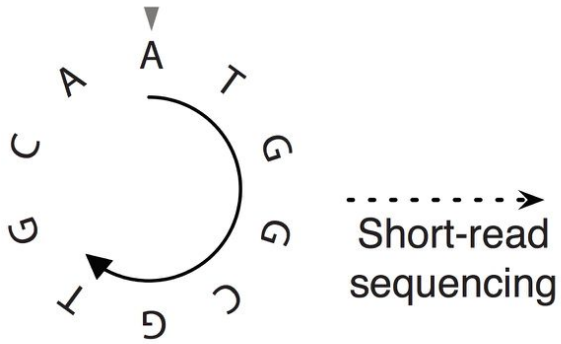
[1736] Seven Bridges of Königsberg

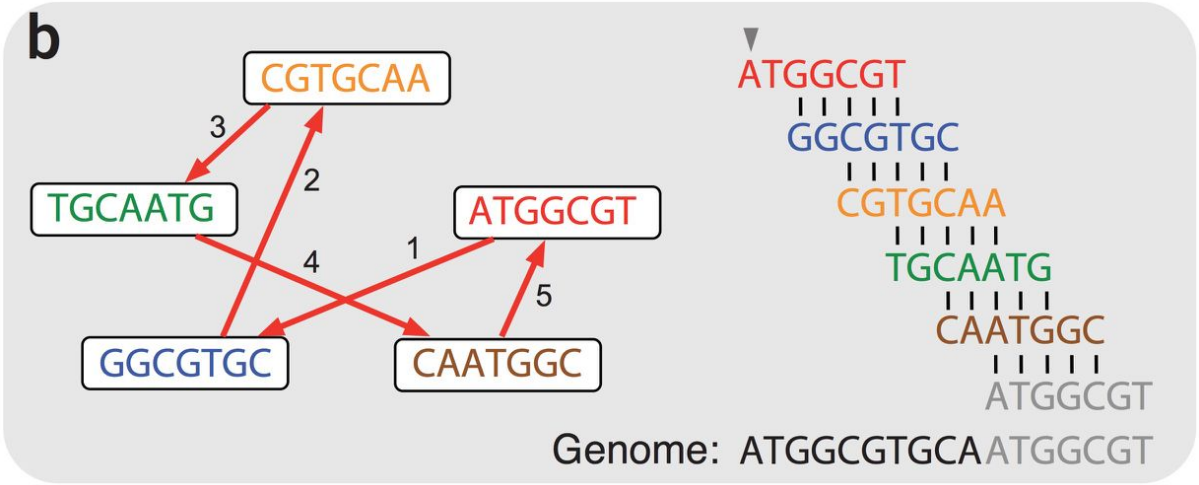
Can one walk through the city so that they cross each of those bridges once and only once?

Graph: (Nodes, Edges, Weights)

(Eulerian) Path exists if the graph contains zero or two vertices that have an odd degree.

Overlap graphs for sequence assembly





Nodes: Reads

Edges: Alignments between reads (≥ 5 bases)

Genome: Walk along a Hamiltonian cycle (visit each vertex exactly once) to combine alignments between successive reads and reconstruct the full circular genome.

The shortest common superstring problem

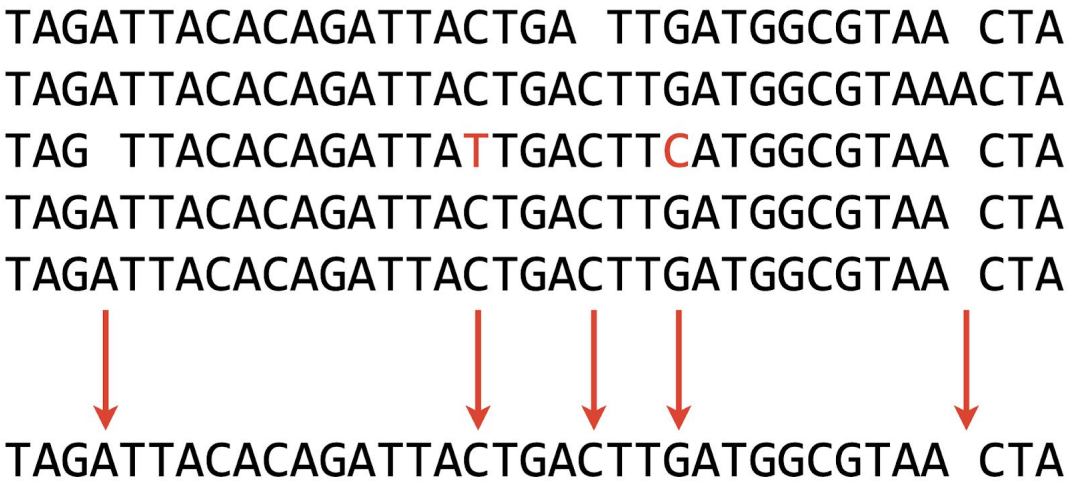
Find a shortest circular 'superstring' that contains all possible 'substrings' of length k (k-mers) over a given alphabet.

Consider:
- alphabet: 0 & 1
- all 3-mers: 000, 001, 010, 011, 100, 101, 110, 111.
(There exist n^k k-mers in an alphabet containing n symbols.)

A concatenation contains all 3-mers: 000001010011100101110111

The circular superstring **0001110100** contains all 3-mers & each 3-mer exactly once.

Building a consensus



Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote

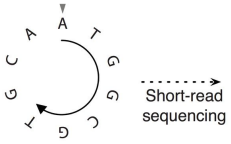
Issues

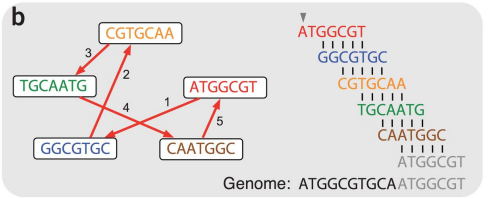
Four hidden assumptions that **do not hold** for real sequencing:

1. We can generate all k-mers present in the genome.
2. All k-mers are error free.
3. Each k-mer appears at most once in the genome.
4. The genome consists of a single circular chromosome.

E.g., a technology that generates 100-nucleotide long reads:

- may miss some 100-mers present in the genome (even if the read coverage is high)
- the 100-mers that it does generate typically have errors.



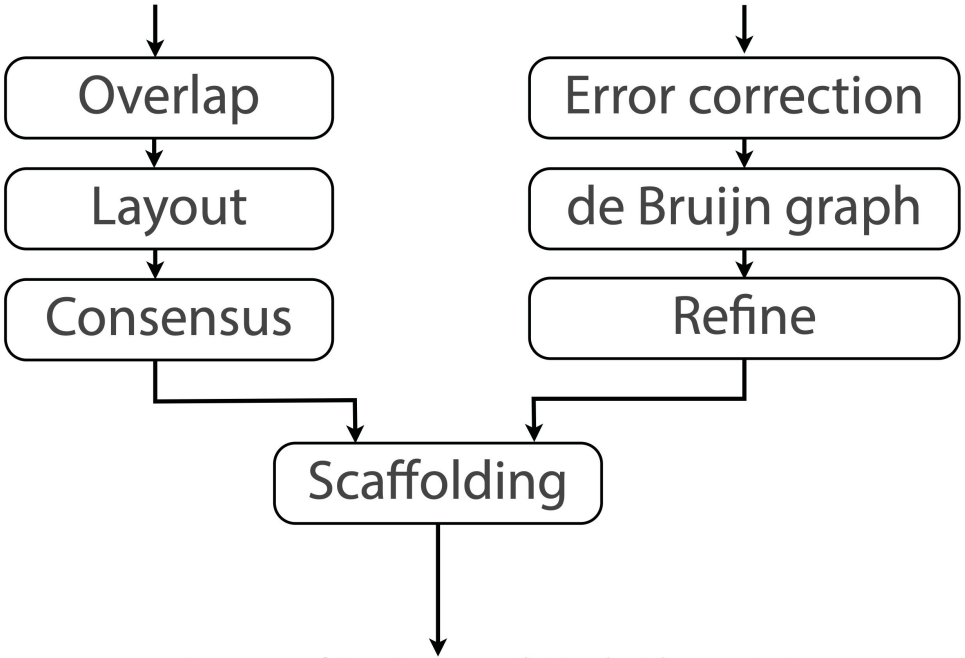


The main drawback of **OLC** is that building the overlap graph can be very slow.

E.g., 2nd-generation sequencing datasets contain ~100s of millions or billions of reads, hundreds of billions of nucleotides total.

Genome assembly - two major approaches

1. Overlap-Layout-Consensus (OLC) assembly
2. de Bruijn graph (DBG) assembly



de Bruijn graphs: the 'superstring problem'

Find a shortest circular 'superstring' that contains all possible 'substrings' of length k (k-mers) over a given alphabet.

Consider:
- alphabet: 0 & 1
- all 3-mers: 000, 001, 010, 011, 100, 101, 110, 111.
(There exist n^k k-mers in an alphabet containing n symbols.)

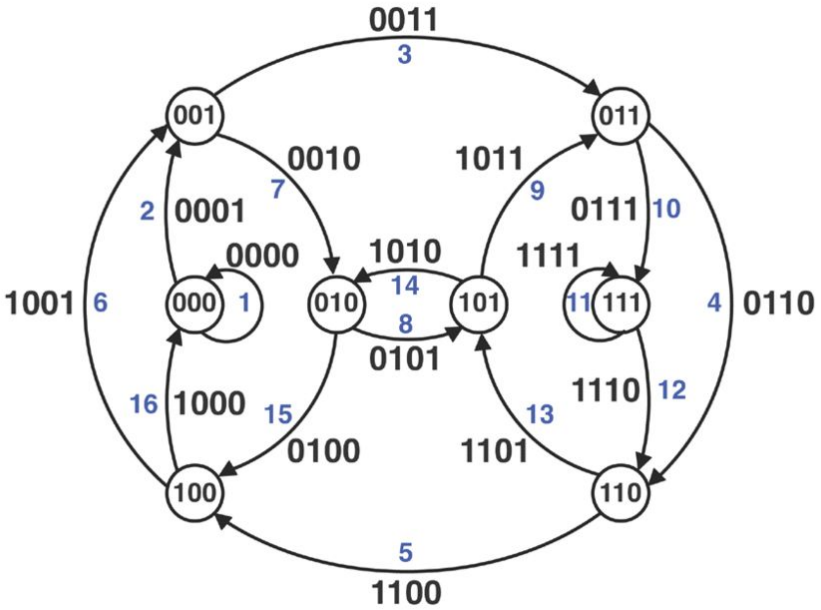
A concatenation contains all 3-mers:
000001010011100101110111

The circular superstring **0001110100** contains all 3-mers & each 3-mer exactly once.

How can we construct a superstring for all k-mers in the case of an arbitrary value of k and an arbitrary alphabet?

Construct a **directed graph**: all prefixes & suffixes – (k-1)-mers – as nodes; all k-mers as edges.

de Bruijn graphs: the 'superstring problem'



k = 4 | Two-character alphabet: digits 0 & 1

Does this graph have an Eulerian cycle? [Balanced?]

Following the blue numbered edges in order from 1 to 16 traces the cyclic superstring 0000110010111101.

How can we construct a superstring for all k-mers in the case of an arbitrary value of k and an arbitrary alphabet?

Construct a **directed graph**: all prefixes & suffixes – (k-1)-mers – as nodes; all k-mers as edges.

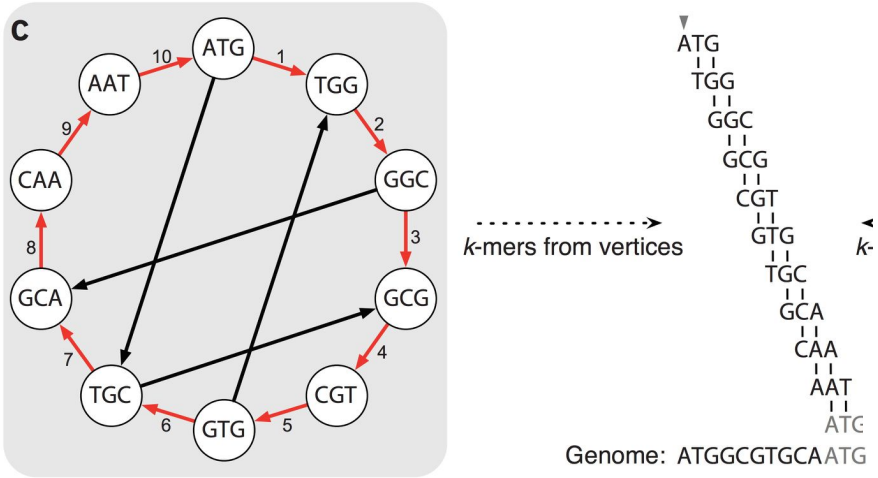
Overlap graphs for sequence assembly

Nodes: k-mers
Edges: Overlap between k-mers
Genome: Walk the Hamiltonian cycle

Large genomes result in too many reads:

- 10^6 reads $\longrightarrow$ 10^{12} pairwise alignments.
- 10^9 reads $\longrightarrow$ 10^{18} alignments.

There is no known efficient algorithm for finding a Hamiltonian cycle in a large graph with millions (let alone billions) of nodes.



Compeau et al. (2011)

de Bruijn graphs for sequence assembly

Break reads into shorter k-mers!

Resulting k-mers often represent nearly all k-mers from the genome for sufficiently small k.

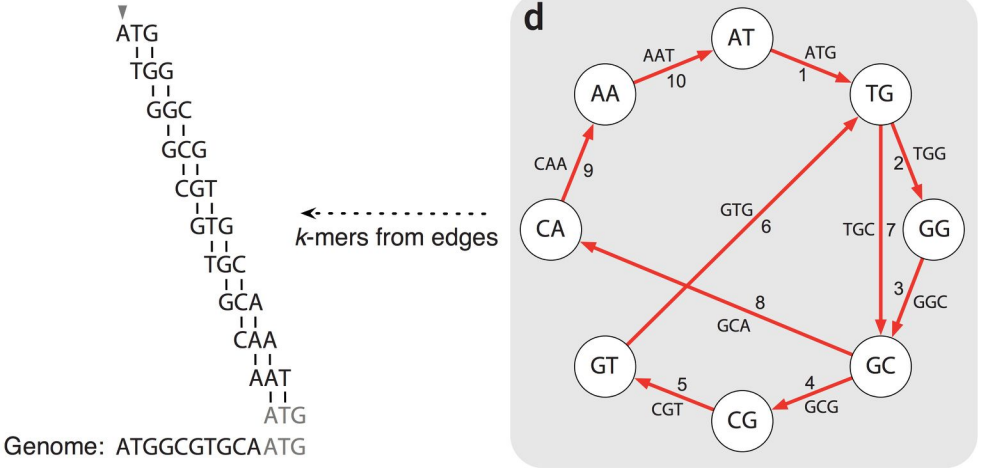
→

Finding a path that visits all *edges* of a graph exactly once is much easier!

Nodes: (k-1)-mers [in- & out-degrees?]
Edges: k-mers
Genome: Walking the Eulerian cycle

Send out an ant, find a cycle; Note edges traversed, send out another ant, … until all of the graph's edges have been explored. Combine all cycles to form an Eulerian cycle!

Computationally tractable; time roughly proportional to the number of edges.

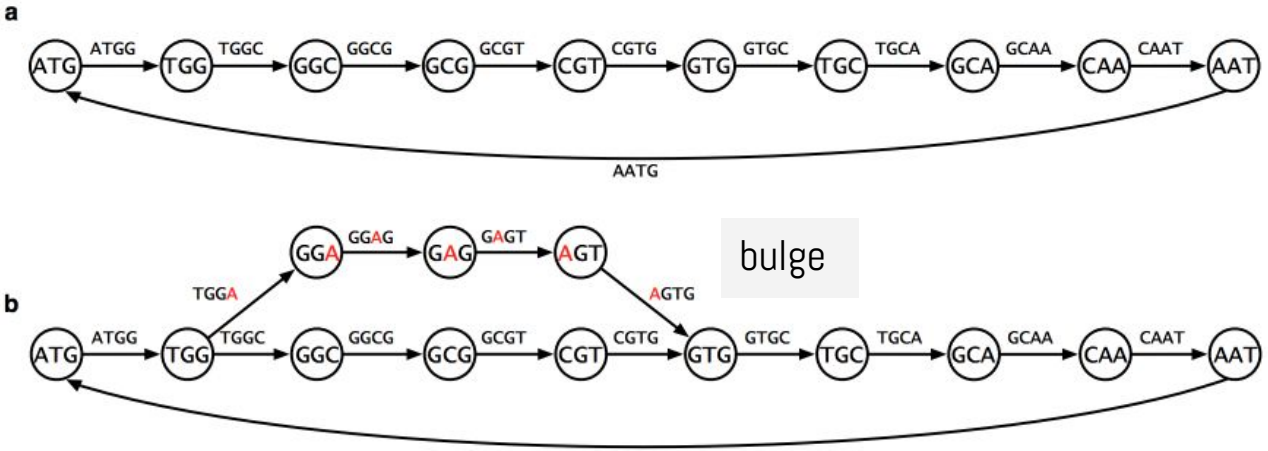


Compeau et al. (2011)

de Bruijn graphs for sequence assembly – not so easy w/ real data

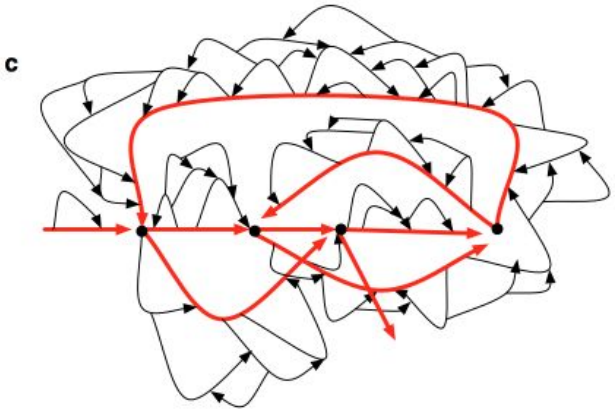
- Not all k-mers in the genome
  - Read breaking
- Errors in reads
  - Error correcting reads
  - Removing bulges in de Bruijn graphs
- DNA repeats
  - Incorporating k-mer multiplicity
- Multiple and linear chromosomes
  - Cycles to paths
- Unsequenced regions
  - Scaffolds

de Bruijn graphs from reads with sequencing errors



TGG***A***GTG (incorrect)

TGG***C***GTG (correct)



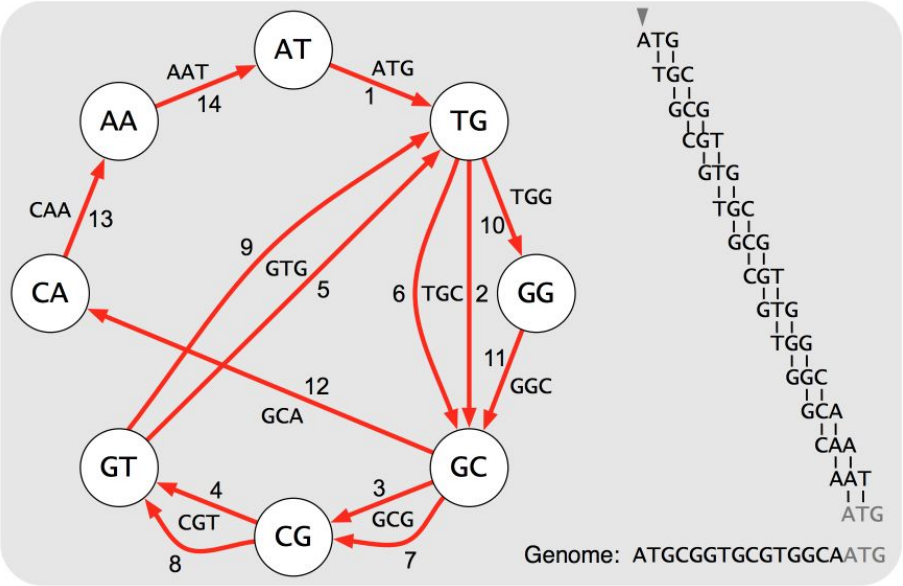
The process of bulge removal should leave only the red edges remaining, yielding an Eulerian path in the resulting graph.

Compeau et al. (2011)

de Bruijn graphs for sequence assembly – not so easy w/ real data

- Not all k-mers in the genome
  - Read breaking
- Errors in reads
  - Error correcting reads
  - Removing bulges in de Bruijn graphs
- DNA repeats [E.g., ATGCATGC → four 3-mers: ATG, TGC, GCA & CAT]
  - Incorporating k-mer multiplicity
- Multiple and linear chromosomes
  - Cycles to paths
- Unsequenced regions
  - Scaffolds

de Bruijn graphs for dealing with repeats



Genome: **ATGCGTGCGTGGCA**

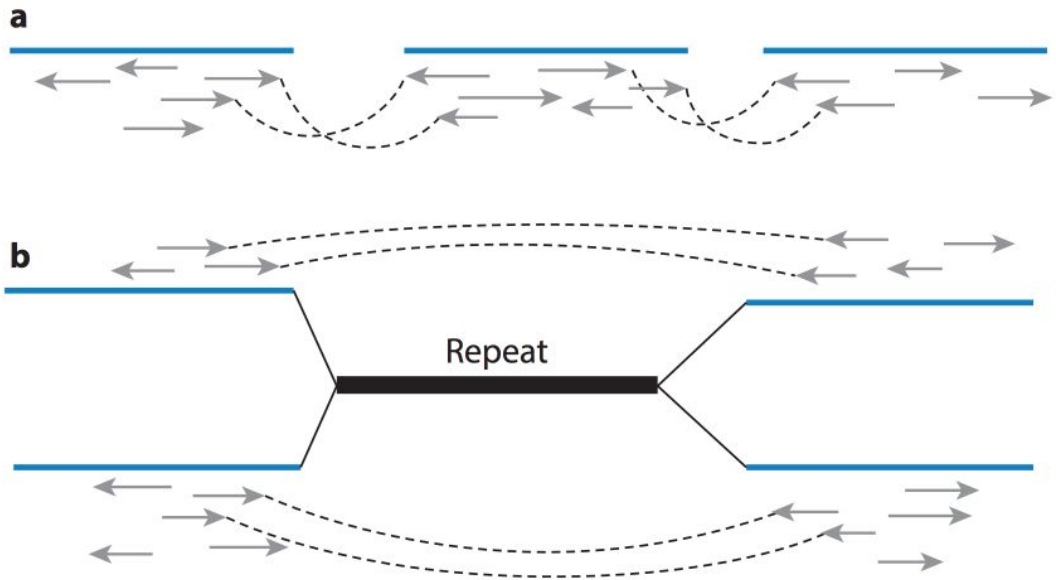
Incorporate k-mer multiplicity:

- Four 3-mers **TGC**, **GCG**, **CGT**, and **GTG**: multiplicity 2
- Six 3-mers **ATG**, **TGG**, **GGC**, **GCA**, **CAA**, and **AAT**: multiplicity 1

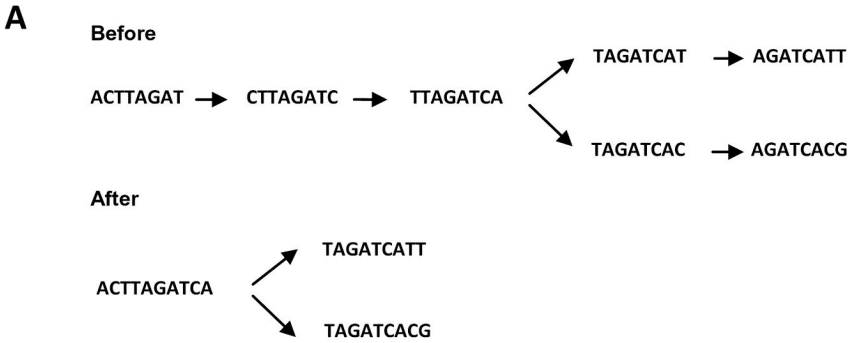
Compeau et al. (2011)

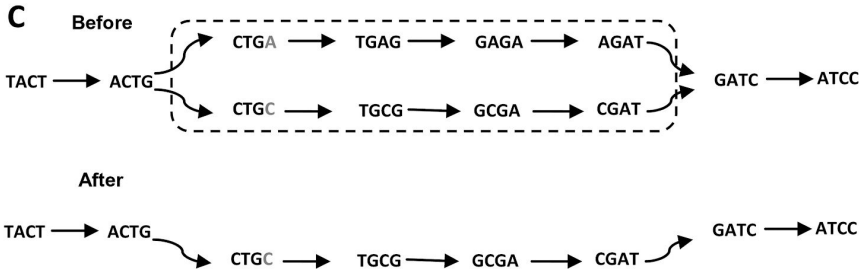
de Bruijn graphs for sequence assembly – not so easy w/ real data

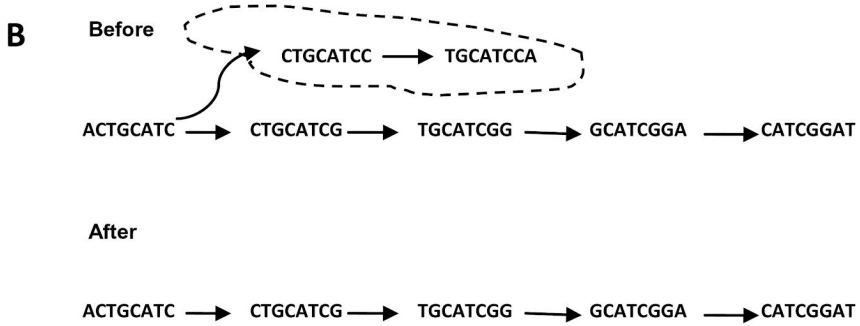
- Not all k-mers in the genome
  - Read breaking
- Errors in reads
  - Error correcting reads
  - Removing bulges in de Bruijn graphs
- DNA repeats
  - Incorporating k-mer multiplicity
- Multiple and linear chromosomes
  - Cycles to paths
- Unsequenced regions
  - Scaffolds

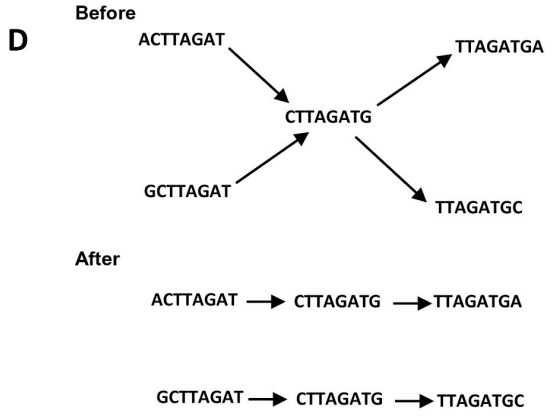


Simplifying de Bruijn graphs



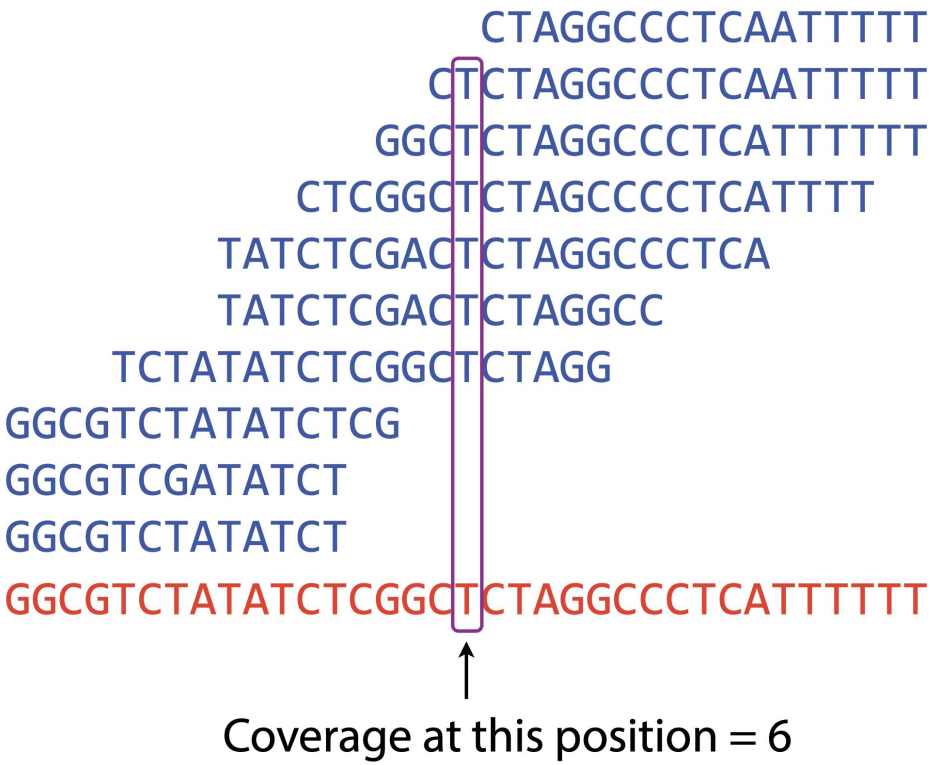






El-Metwally et al. (2013)

Genome assembly - quality & contiguity



177 nucleotides in total

Genome is 35 nucleotides

Average coverage = 177 / 35 ~ 7x

N50: A statistic used for assessing the contiguity of a genome assembly.

- The contigs in an assembly are sorted by size and added, starting with the largest.
- N50 is the size of the contig that makes the total greater than or equal to 50% of the genome size.

Algorithms for genome assembly

- Error detection and correction based on sequence composition of the reads.
- Graph construction to represent reads/k-mers and their shared sequence.
- Graph adjustments:
  - Reduction of simple non-intersecting paths to single nodes.
  - Removal of error-induced paths (recognized as spurs or bubbles).
  - Collapse of polymorphism-induced complexity (bubbles).
  - Simplification of tangles (using information outside the graph: individual, paired-end, or mate-pair reads to constraints on path distance & outcome.
- Conversion of reduced paths to contigs and scaffolds.
- Reduction of alignments to a consensus sequence.

Genome sequencing technologies

Technology	Read length (bp)	Error rate	Native paired-end read support
ABI/Solid	75	Low (~2%)	Yes
Illumina/Solexa	100–150	Low (<2%)	Yes
IonTorrent	~200	Medium (~4%)*	No
Roche/454	400–600	Medium (~4%)*	No
Sanger	Up to ~2,000 bp	Low (~2%)	Yes
Pacific Biosciences	Up to ~15,000‡	High (~18%)	Yes (in strobe read mode)
Oxford nanopore	Up to ~20,000	High (~12%)	

Nagarajan & Pop (2013)

Genome assemblers

Assemblers	Technology	Availability	Notes
ALLPATHS-LG	Illumina, Pacific Biosciences	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG	Requires a specific sequencing recipe (BOX 3)
SOAPdenovo	Illumina	http://soap.genomics.org.cn/soapdenovo.html	Also used for transcriptome and metagenome assembly
Velvet	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/velvet	May have substantial memory requirements for large genomes
ABySS	Illumina, SOLiD, 454, Sanger	http://www.bcgsc.ca/platform/bioinfo/software/abyss	Also used for transcriptome assembly

Nagarajan & Pop (2013)

Metagenome, Transcriptome, Single-cell assemblers

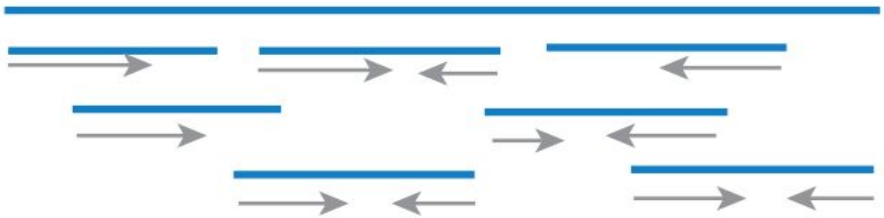
Assemblers	Technology	Availability	Notes
Metagenome assemblers			
Genovo	454	http://cs.stanford.edu/group/genovo	Uses a probabilistic model for assembly
MetaVelvet	Illumina, SOLiD, 454, Sanger	http://metavelvet.dna.bio.keio.ac.jp	Based on Velvet
Meta-IDBA	Illumina	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba	Based on IDBA
Transcriptome assemblers			
Trinity	Illumina, 454	http://trinityrnaseq.sourceforge.net	Tailored to reconstruct full-length transcripts; may require substantial computational time
Oases	Illumina, SOLiD, 454, Sanger	http://www.ebi.ac.uk/~zerbino/oases	Based on Velvet
Single-cell assemblers			
SPAdes	Illumina	http://bioinf.spbau.ru/en/spades	
IDBA-UD	Illumina	http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud	Based on IDBA

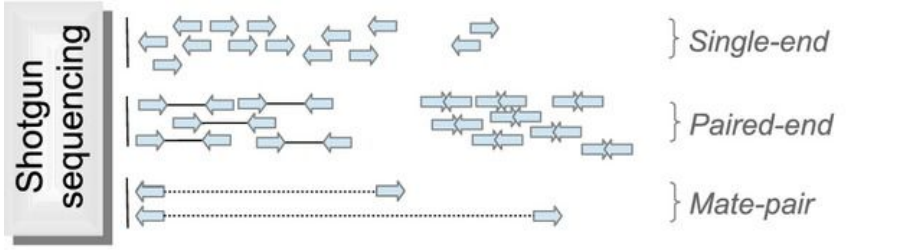
Nagarajan & Pop (2013)

Algorithms for genome assembly

- New sequencing technologies ⟶ a different best computational strategy.
- Factors that influence the choice of algorithms:
  - Quantity of data (read length and coverage)
  - Quality of data (including error rates)
  - Genome structure (e.g., GC content and the number and size of repeated regions).
- de Bruijn graphs are best suited for current short-read sequencing technologies
  - Produce very large numbers of reads
  - Can represent genomes with repeats
    [Overlap methods need to mask repeats > read length]
- Long-read technology growing at a rapid pace.

Genome assembly & annotation – Overview





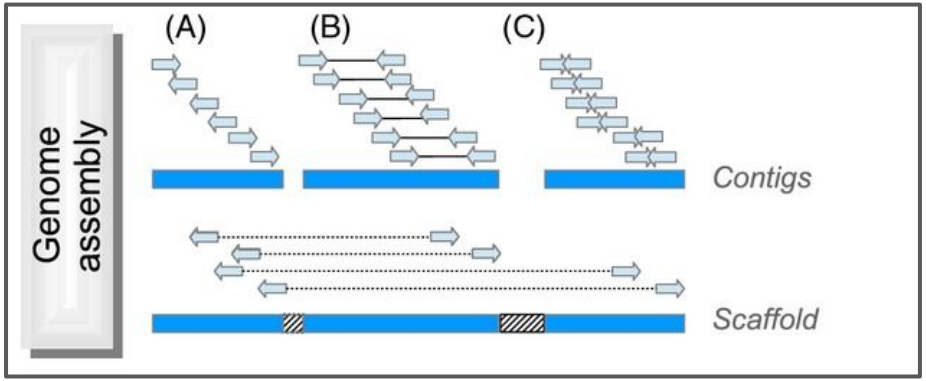
read: a short/long word that comes out of sequencer

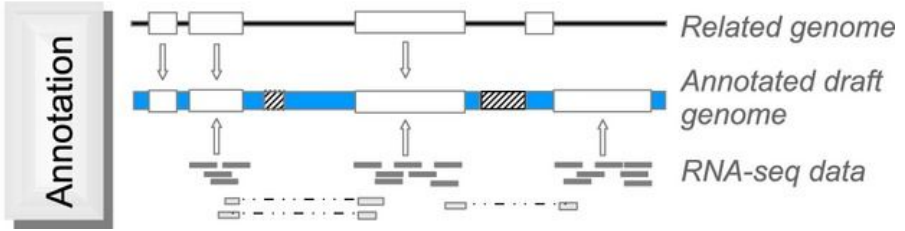
mate pair: a pair of reads from two ends of the same insert fragment

contig: a contiguous sequence formed by several overlapping reads with no gaps

scaffold: an ordered and oriented set of contigs, usually by mate pairs

consensus sequence: derived from the sequence multiple alignment of reads in a contig





Ekblom & Wolf (2014)