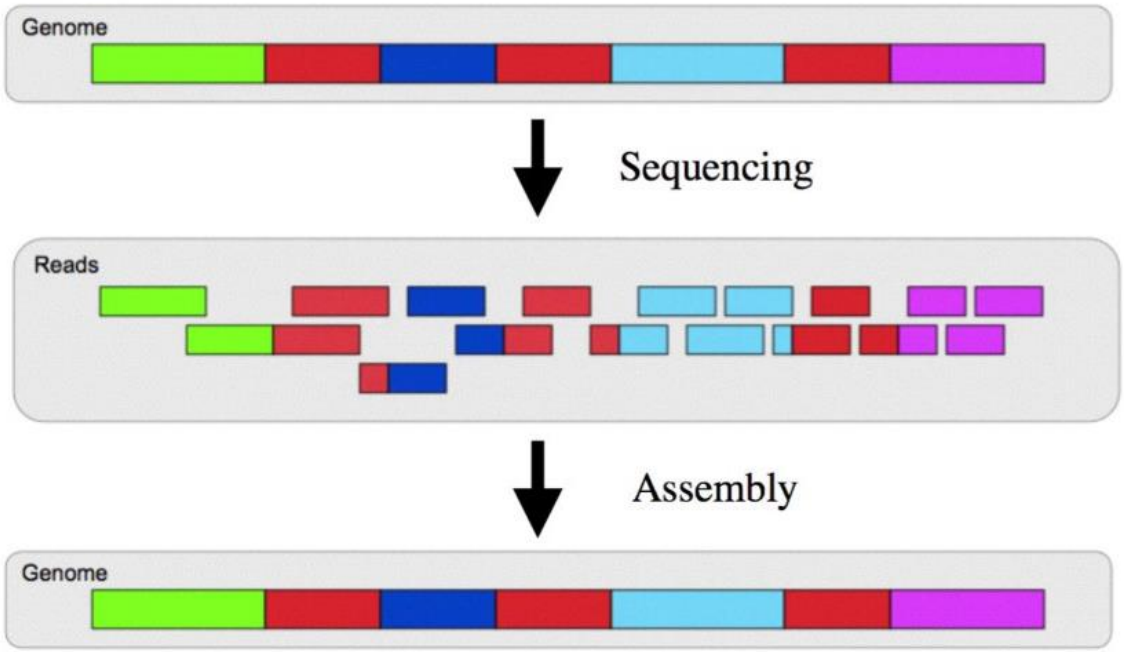
Genome assembly by short/long-read sequencing

2019 Dragon Star Bioinformatics Course (Day 2)

What is genome assembly?

- The genome assembly
  - Genome sequence produced after chromosomes have been fragmented
  - Those fragments have been sequenced
  - The resulting sequences have been put back together.



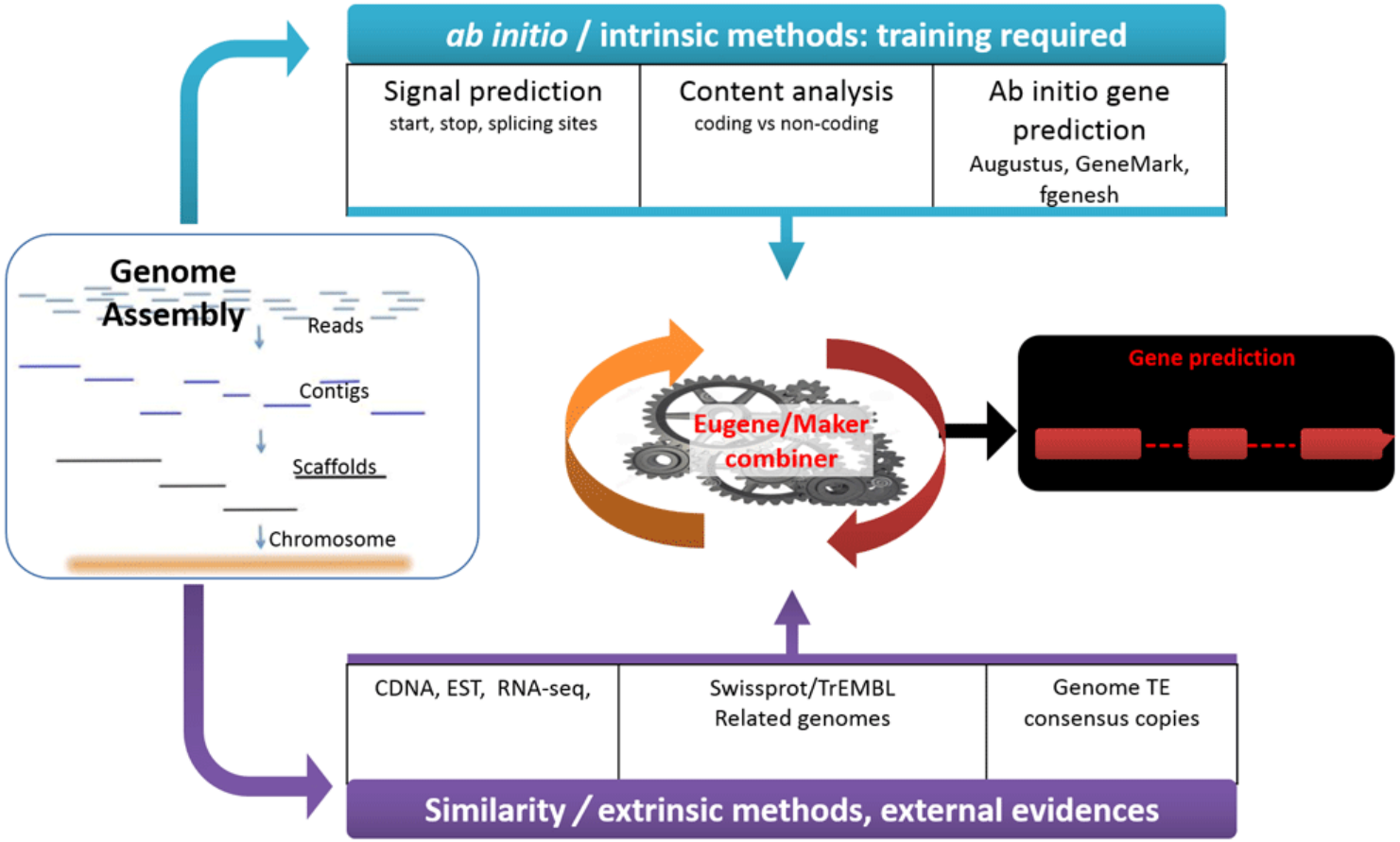
repeats

https://bioinformatics.ca/

Why genome assembly?

- Accurate assembly of genomes
  - Key to understanding genetic variation.
  - The more accurate the reference genome, the easier it is to map reads and interpret the functional impacts of genetic mutations
- most of the human genetic data analysis relies on reference genome
  - Although some reference-free variant caller are available
  - Therefore, accurate genome assembly and more complete (that incorporate population variation and alternative haplotypes) is important for human genetic studies.

Why genome assembly?



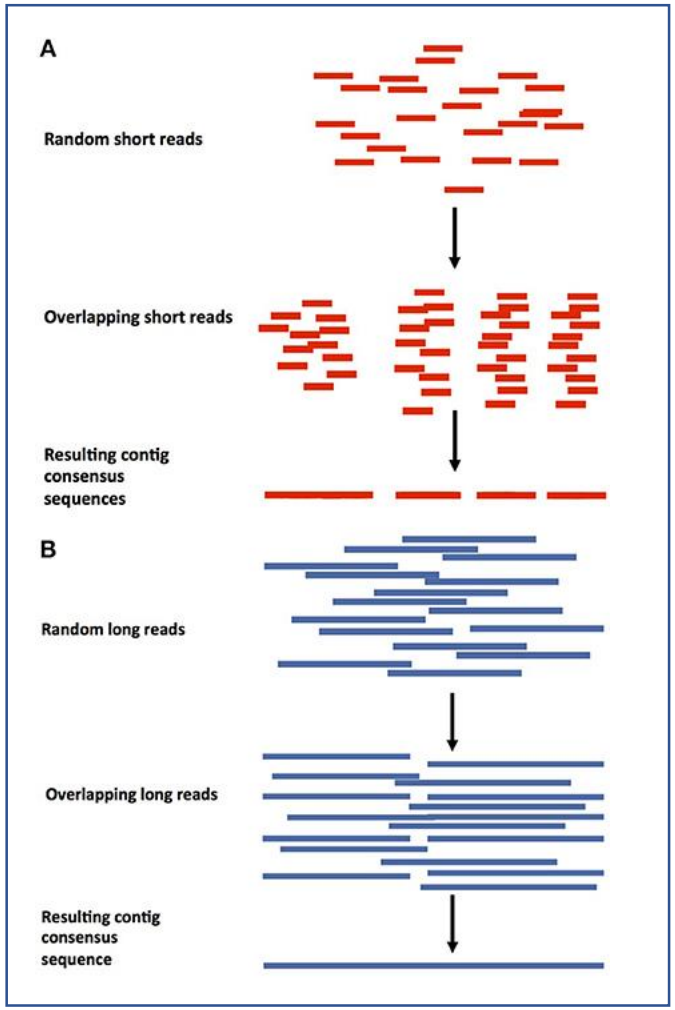
Dominguez Del Angel, F1000 Research, 2018

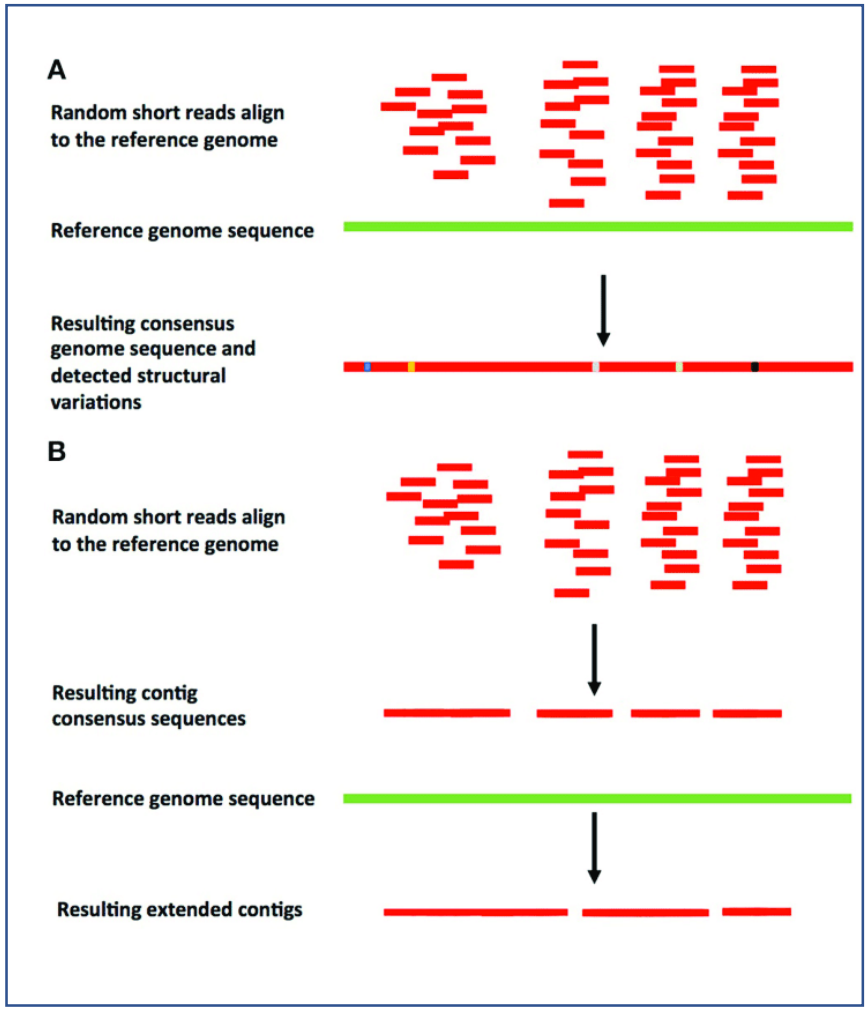
What types of genome assembly?

- *de novo* assembly:
  - The entire assembled sequence is resolved from raw sequence data without comparison to a reference genome sequence.

- Comparative assembly:
  - You have a "reference" genome to guide the assembly process.
    - For example, sequence and assemble a new human genome, using GRCh38 as a guide.
  - Or sequence a chimp genome using the existing human assembly as a guide.

De novo vs reference-guided assembly





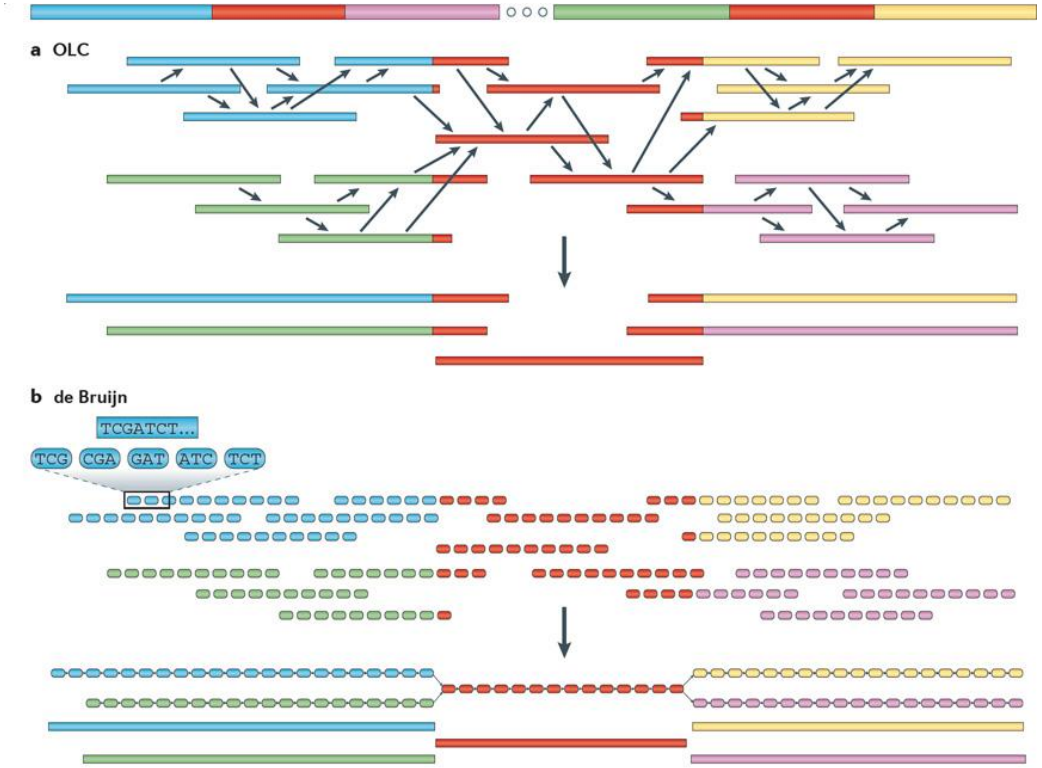
Kyriakidou, Frontiers in Plant Science, 2018

De Novo Assembly paradigms

- Overlap-layout-consensus (OLC) methods
  - Greedy (TIGR Assembler, phrap, CAP3...)
  - Graph-based (Celera Assembler, Arachne)
  - Examples are Arachne, Celera Assembler, CAP3, PCAP, Phrap, Phusion and Newbler
- De bruijn (k-mer) graph
  - Especially useful for assembly from short reads
  - Stacking overlapping sequences of genomic fragments of a defined size (the k-mer), generated by breaking each read into k-mer size.
  - Examples are Euler-USR, Velvet, ABySS, AllPath-LG and SOAPdenovo.
- Extensions and other approaches
  - String graph
  - Hybrid approach

OLC versus de Bruijn graph

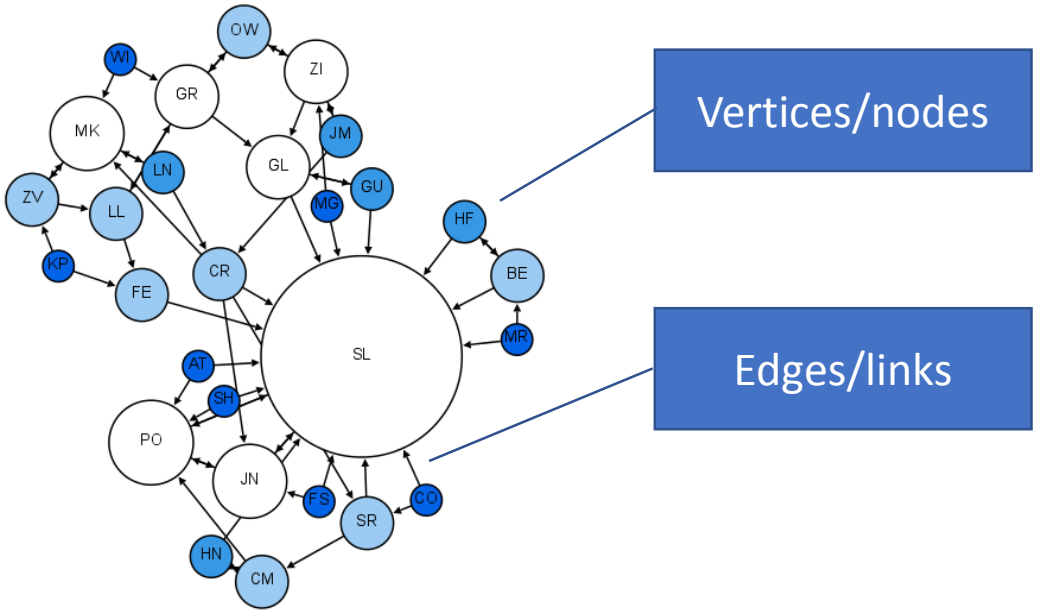
- OLC: pairwise alignment between reads are detected and merged
- de Buijn: reads are decomposed into k-mers and merged



Chaisson et al, Nature Reviews Genetics, 2015

Graph theory

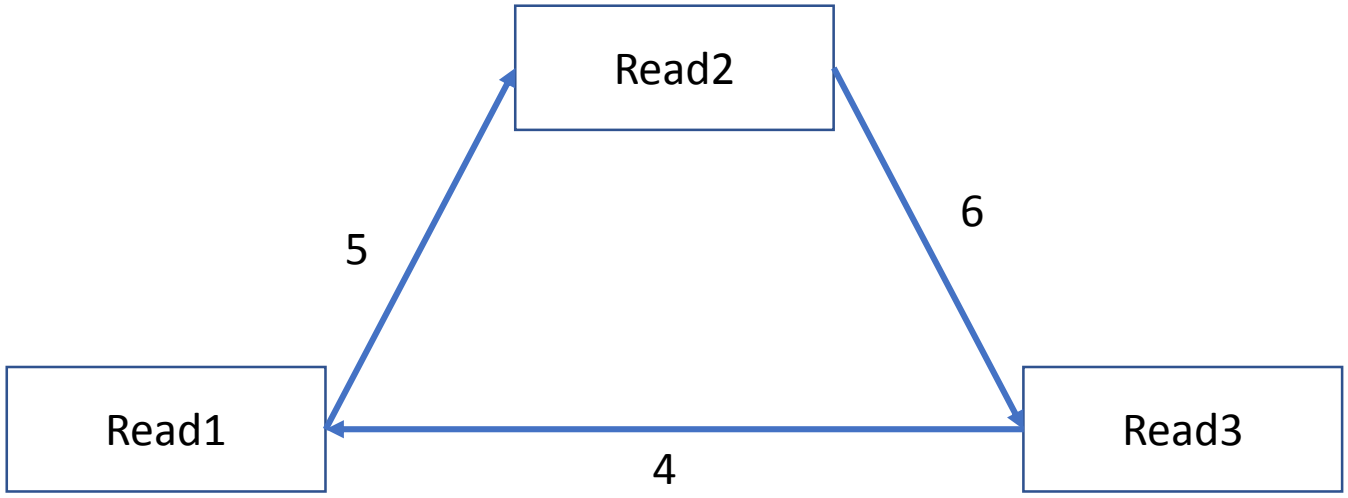
- Graph:
  - Mathematical structures used to model pairwise relations between objects.
  - A graph is made up of vertices (also called nodes or points) which are connected by edges (also called links or lines).



https://en.wikipedia.org/wiki/Graph_theory

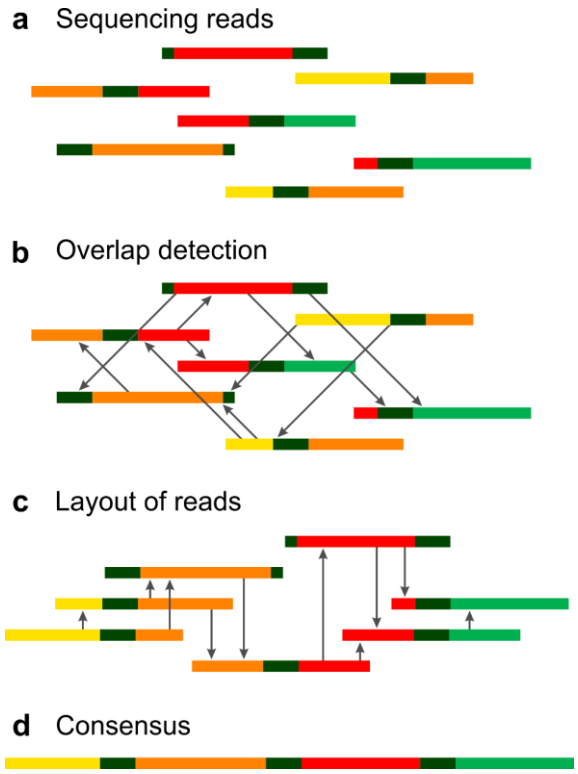
Directed graph

- G(V, E): a set of vertices (V) and a set of directed edges (E)
- Vertex and edge may have weights associated with them
- Here, V= {Read1, Read2, Read3}, E = {(Read1, Read2), (Read2, Read3), (Read3, Read1)} with weights of 5, 6 and 4.



Overlap-based approach

- Overlap graph:
  - Nodes are initially formed by the sequences of the individual reads
  - Edges are represented by the sequence overlaps between these reads.



Vollmers et al, PLoS ONE, 2017

Three steps: overlap-layout-consensus

- Overlap:
  - The identification of overlaps between all sequencing reads is the first and most time consuming step of the assembly process.
  - Pairwise overlaps of variable length are identified between all reads in the dataset and represented as edges between these reads in the resulting overlap graph
- Layout:
  - This information is used in the next step to layout the reads into the most probable contiguous sequence stretches
- Consensus:
  - In the final step the consensus sequence is determined for each contig by choosing the nucleotide, which is represented by the majority of the overlapping reads for every sequence position

Constructing the overlap graph

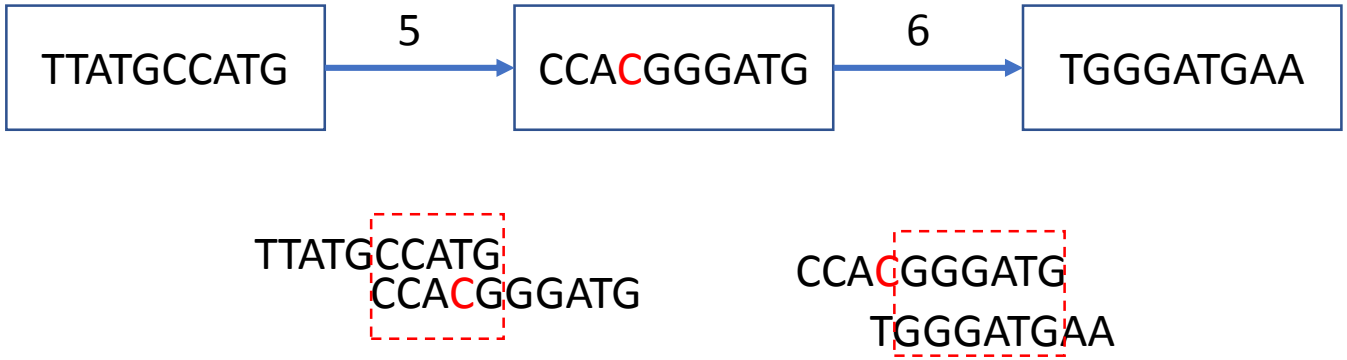
- Create the overlap graph for the following string, where overlap means suffix/prefix match of >=5bp
  - TT**ATG**CC**ATG**GG**ATG**AA
- Note that TT, CC, GG and AA are separated by 3 identical ATG
- Assume that we generated a few short reads from this string: TTATGCCATG, CCATGGGATG, TGGGATGAA
- Vertices/nodes are formed by the sequences of the individual reads; edges are represented by the sequence overlaps between these reads

Constructing the overlap graph

- Genome: TT**ATG**CC**ATG**GG**ATG**AA
- Vertices: {TTATGCCATG, CCATGGGATG, TGGGATGAA}
- Edges: {(TTATGCCATG, CCATGGGATG), (CCATGGGATG, TGGGATGAA)}

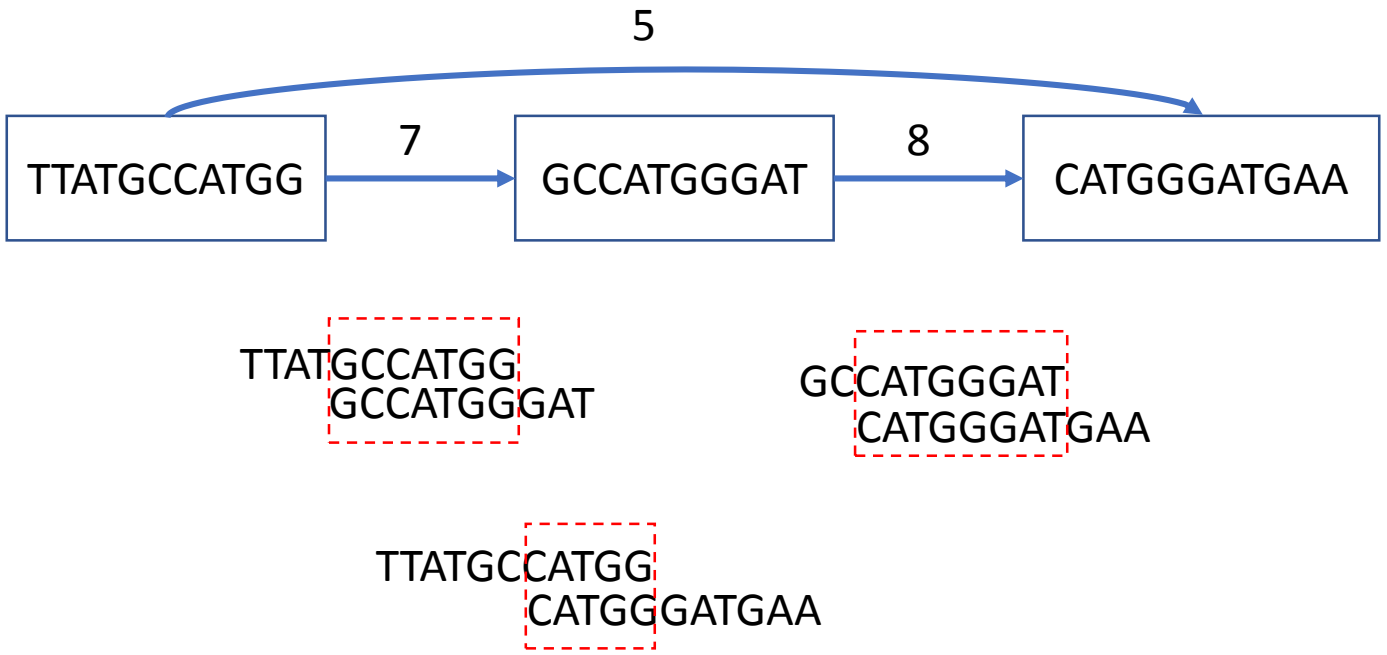
Sometimes we need to tolerate sequencing errors

- Genome: TT**ATG**CC**ATG**GG**ATG**AA
- Vertices: {TTATGCCATG, CCACGGGATG, TGGGATGAA}
- Edges: {(TTATGCCATG, CCACGGGATG), (CCATGGGATG, TGGGATGAA)}

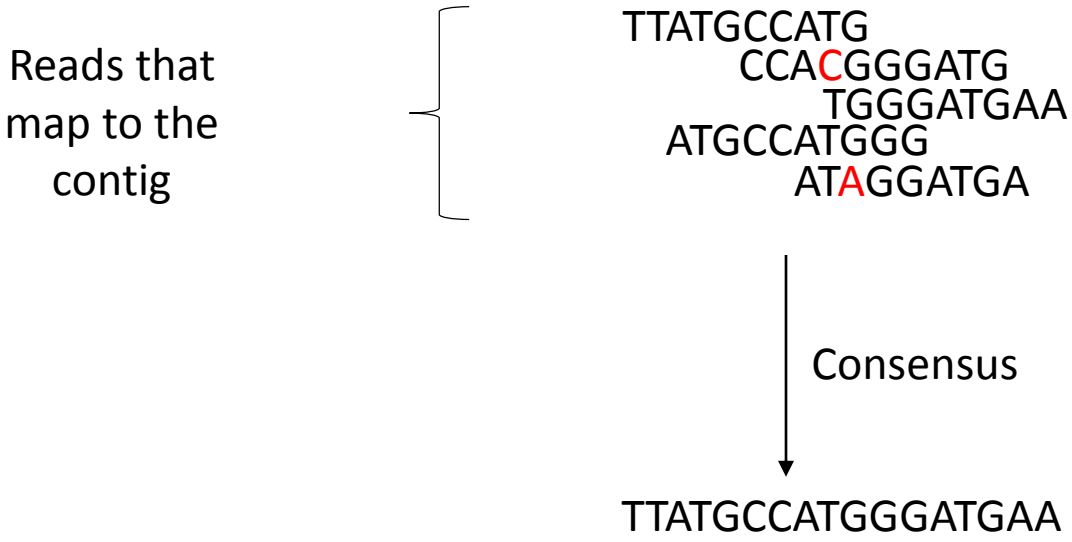


Sometimes graphs can be reduced by eliminating transitive edges

- Genome: TT**ATG**CC**ATG**GG**ATG**AA



Consensus step: inferring the most likely nucleotide at each position

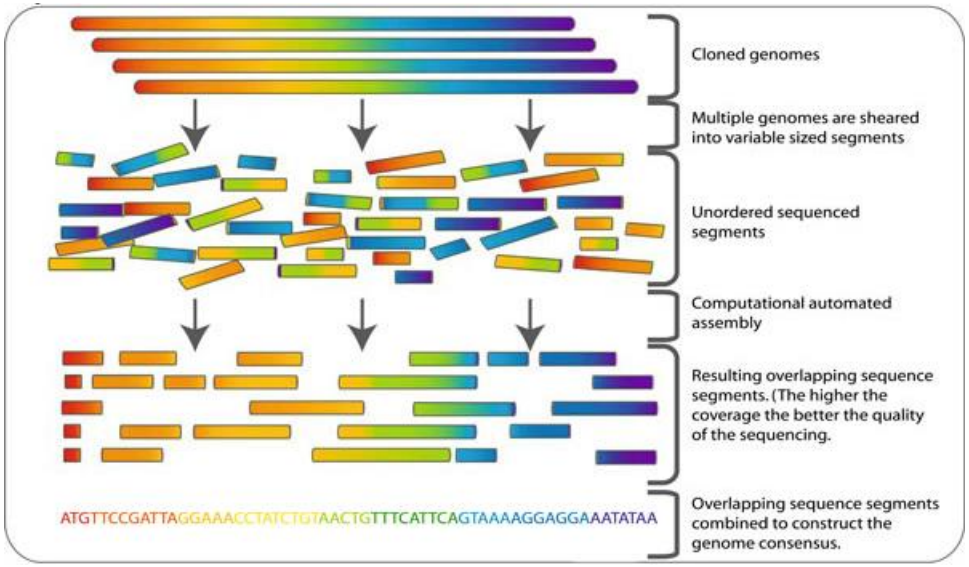


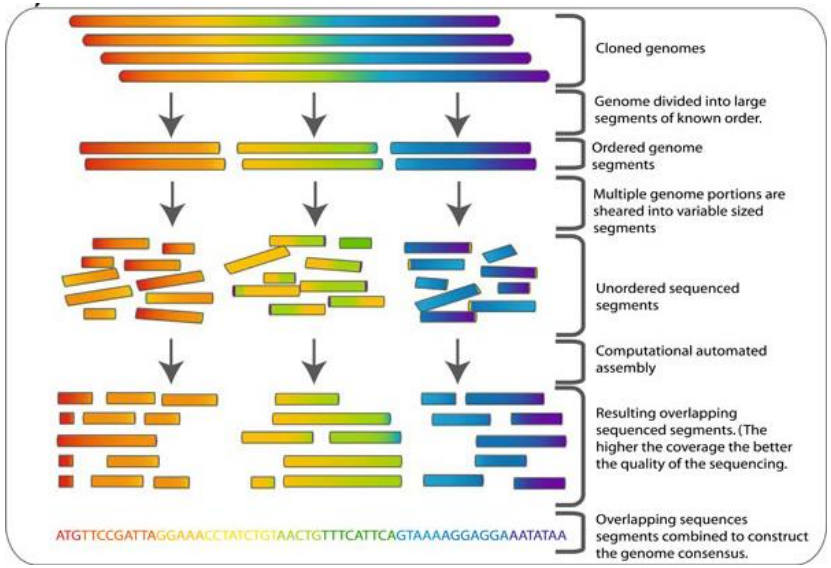
Note: sequencing error or mapping error (map to repetitive regions) or ploidy can result in mismatches

Note: For long-read sequencing, it is common to do error correction first to reduce error rates of reads, and then use a fraction of the longest reads for overlap-based assembly

Comparison of two approaches for large genomes such as human genomes

- Whole genome shotgun sequencing vs hierarchical approach





Commins et al, Biol Proced Online. 2009

Now let's go back 30 years

- How should we sequence and assemble a human genome?
- Begun formally in 1990, the U.S. Human Genome Project was a 13-year effort coordinated by the U.S. Department of Energy (DOE) and the National Institutes of Health (NIH).
- Primary goals were to
  - Discover the complete set of human genes and make them accessible for further biological study
  - Determine the complete sequence of DNA bases in the human genome.

- Two thoughts: whole-genome shotgun versus physical mapping followed by BAC-by-BAC sequencing

Weber et al, Genome Research, 1997

Primary Human Genome Project Sequencing Sites

- U.S.DOE Joint Genome Institute, Walnut Creek, California, USA*
- Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, Houston, Texas, USA*
- The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, United Kingdom*
- Washington University School of Medicine Genome Sequencing Center, St. Louis, Missouri, USA*
- Whitehead Institute/MIT Center for Genome Research, Cambridge, Massachusetts, USA*

Additional Contributors to HGP Research

- Australian Genome Research Facility
- Beijing Genomics Institute/Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China*
- Chromosome 21 Consortium
- Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, Cold Spring Harbor, New York, USA*
- Chromosome 13 - Sanger Centre, UK
- Computational Biology at Oak Ridge National Laboratory, Oak Ridge, TN, USA
- European Bioinformatics Institute
- Foundation Jean Dausset-CEPH
- Généthon
- Genome Database
- Genoscope and CNRS UMG-8030, Evry, France*
- German Human Genome Project
- Gesellschaft fur Biotechnologische Forschung mbH, Braunschweig, Germany*
- GTC Sequencing Center, Genome Therapeutics Corporation, Waltham, Massachusetts, USA*
- Department of Genome Analysis, Institute of Molecular Biotechnology, Jena, Germany*
- The Institute for Systems Biology, Seattle, Washington, USA*
- Japan Science and Technology Corporation Sequencing Projects
- Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan*
- Max Planck Institute for Molecular Genetics, Berlin, Germany*

https://web.ornl.gov/sci/techresources/Human_Genome/research/centers.shtml

How should we assemble a human genome?

- Two thoughts: whole-genome shotgun versus physical mapping followed by BAC-by-BAC sequencing

PERSPECTIVE

Human Whole-Genome Shotgun Sequencing

James L. Weber[1,3] and Eugene W. Myers[2]

[1]Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449;
[2]Department of Computer Science, University of Arizona, Tucson, Arizona 85721

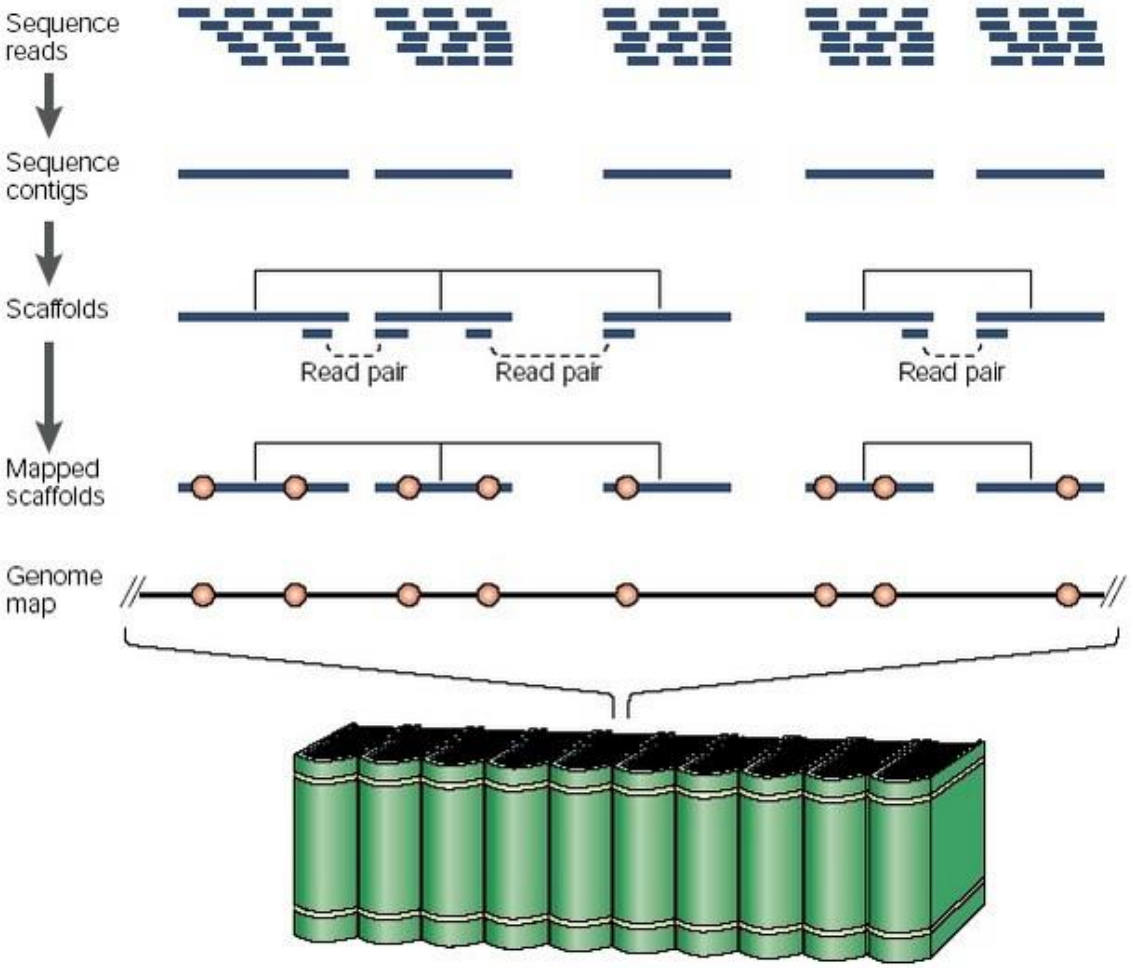
Large-scale sequencing of the human genome is now under way (Boguski et al. 1996; Marshall and Pennisi 1996). Although at the beginning of the Genome Project, many doubted the scientific value of sequencing the entire human genome, these doubts have evaporated almost entirely (Gibbs 1995; Olson 1995). Primary reasons for generating the human genomic sequence are listed in Table 1.

Table 5. Costs of Human Genomic Sequencing

Clone by clone
$0.30 per finished base
$130 million per year for 7 years
Total $900 million spent by end of 2003
Shotgun
$0.01 per raw base
$130 million for 3 years would provide 10× coverage plus an additional $90 million for informatics

Weber et al, Genome Research, 1997

How whole-genome shotgun sequencing works



Green, Nat Review Genet, 2001

- Two thoughts: whole-genome shotgun versus physical mapping followed by BAC-by-BAC sequencing

PERSPECTIVE

Against a Whole-Genome Shotgun

Philip Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

The human genome project is entering its decisive final phase, in which the genome sequence will be determined in large-scale efforts in multiple laboratories worldwide. A number of sequencing groups are in the process of scaling up their throughput; over the next few years they will need to attain a collective capacity approaching half a gigabase per year to complete the 3-Gb genome sequence by the target date of 2005. At present, all contributing groups are using a clone-by-clone approach, in MIT Center for Genome Research, http://www-genome.wi.mit.edu], with several intensively mapped chromosomes already exceeding it (Nagaraja et al. 1997, Bouffard et al. 1997), and BACs average 130 kb or more in size in current libraries (Kim et al. 1996), this STS density should be adequate to obtain contiguous clone coverage of much of the genome; most gaps that remain should be closable by developing new STSs directly from the sequence adjacent to the gap and rescreening the library.

At present, all contributing groups are using a clone-by-clone approach, in which mapped bacterial clones (typically 40–400 kb in size) from known chromosomal locations are sequenced to completion. Among other advantages, this permits a variety of alternative sequencing strategies and methods to be explored independently without redundancy of effort.

Green et al, Genome Research, 1997

Draft human genome

- Public effort: BAC based sequencing (Lander et al., Nature, Feb. 15, 2001). Performed by a consortium of government labs and universities.
  - Break genome into ~100-300kb pieces
  - Create Bacterial Artificial Chromosomes (BACs) from each piece
  - Assemble each piece by Sanger sequencing
  - Then stitch overlapping BACs together to draft genome.

- Private effort: whole genome random shotgun sequencing (Venter et al., Science, Feb. 16, 2001). Performed by Celera, which is a company.
  - In 1998, Craig Venter announced that he was forming Celera that within three years would sequence human genome —seven years before the projected finish of the U.S. government's Human Genome Project.

On the sequencing of the human genome

Robert H. Waterston*†, Eric S. Lander‡, and John E. Sulston§

*Genome Sequencing Center, Washington University, Saint Louis, MO 63108; ‡Whitehead Institute/Massachusetts Institut
Genome Research, Cambridge, MA 02142; and §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxtc
United Kingdom

Communicated by Aaron Klug, Medical Research Council, Cambridge, United Kingdom, December 21, 2001 (received for r

Two recent papers using different approaches reported draft sequences of the human genome. The international Human Genome Project (HGP) used the hierarchical shotgun approach, whereas Celera Genomics adopted the whole-genome shotgun (WGS) approach. Here, we analyze whether the latter paper provides a meaningful test of the WGS approach on a mammalian genome. In the Celera paper, the authors did not analyze their own WGS data. Instead, they decomposed the HGP's assembled sequence into a "perfect tiling path", combined it with their WGS data, and assembled the merged data set. To study the implications of this approach, we perform computational analysis and find that a perfect tiling path with 2-fold coverage is sufficient to recover virtually the entirety of a genome assembly. We also examine the manner in which the assembly was anchored to the human genome and conclude that the process primarily depended on the HGP's sequence-tagged site maps, BAC maps, and clone-based sequences. Our analysis indicates that the Celera paper provides neither a meaningful test of the WGS approach nor an independent sequence of the human genome. Our analysis does not imply that a WGS approach could not be successfully applied to assemble a draft sequence of a large mammalian genome, but merely that the

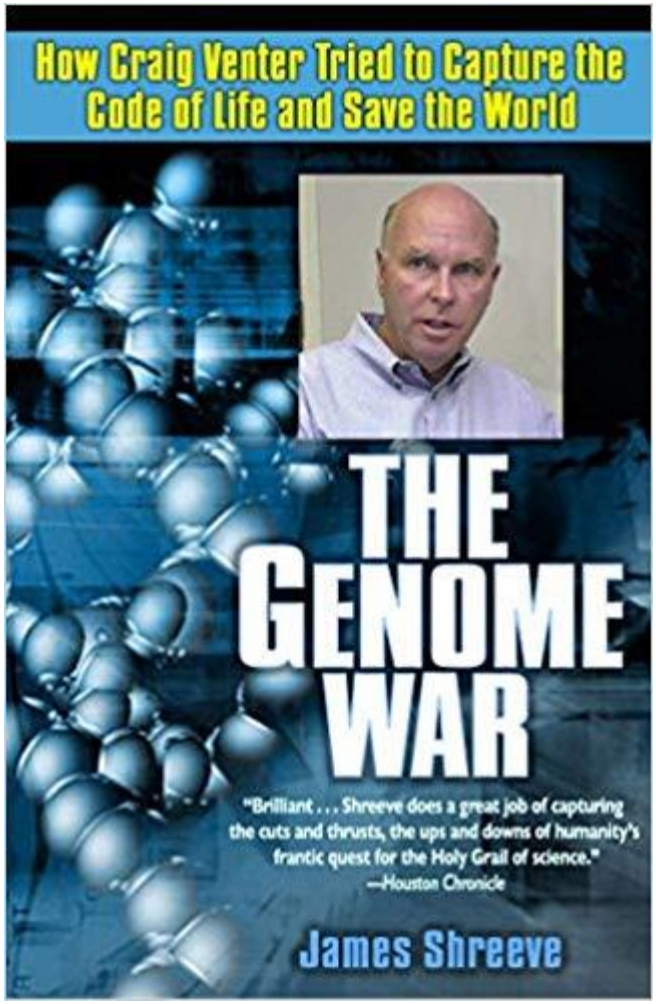
problem is straightforward, beca
sequence can typically be merg
misassembly. The relatively few ga
to produce complete sequences. T
successfully to produce complete
such as plasmids, viruses, organ
genome shotgun data alone also ha
15-fold redundancy (5) to produ
euchromatic portion of the *Dro*
content), although a clone-based
convert this to a finished sequenc

A greater challenge arises in ta
a large proportion of repeat seq
misassembly. Two alternative appr

Hierarchical shotgun (HS) assembl
nome is first broken up into a
intermediate clones such as bact
(BACs). The sequence of each BA
sequencing, and the sequence of
merging the sequences of the BAC
a guaranteed route for producing

What happened from Craig's perspective

- Calling his company Celera (from the Latin for "speed"), he assembled a small group of scientists in an empty building in Rockville, Maryland, and set to work.
- At the same time, the leaders of the government program began to mobilize an unexpectedly unified effort to beat Venter to the prize—knowledge that had the potential to revolutionize medicine and society.
- It is also the story of how one man's ambition created a scientific Camelot where, for a moment, it seemed that the competing interests of pure science and commercial profit might be gloriously reconciled—and the national repercussions that resulted when that dream went awry.



How Perl Saved the Human Genome Project

How Perl Saved the Human Genome Project

by Lincoln Stein

Reprinted courtesy of the Perl Journal, http://www.tpj.com

Lincoln Stein's website is http://stein.cshl.org

DATE: Early February, 1996

LOCATION: Cambridge, England, in the conference room of the largest DNA sequencing center in Europe.

OCCASION: A high level meeting between the computer scientists of this center and the largest DNA sequencing center in the United States.

THE PROBLEM: Although the two centers use almost identical laboratory techniques, almost identical databases, and almost identical data analysis tools, they still can't interchange data or meaningfully compare results.

THE SOLUTION: Perl.

The human genome project was inaugurated at the beginning of the decade as an ambitious international effort to determine the complete DNA sequence of human beings and several experimental animals. The justification for this undertaking is both scientific and medical. By understanding the genetic makeup of an organism in excruciating detail, it is hoped that we will be better able to understand how organisms develop from single eggs into complex multicellular beings, how food is metabolized and transformed into the constituents of the body, how the nervous system assembles itself into a smoothly functioning ensemble. From the medical point of view, the wealth of knowledge that will come from knowing the complete DNA sequence will greatly accelerate the process of finding the causes of (and potential cures for) human diseases.

Six years after its birth, the genome project is ahead of schedule. Detailed maps of the human and all the experimental animals have been completed (mapping out the DNA using a series of landmarks is an obligatory first step before determining the complete DNA sequence). The sequence of the smallest model organism, yeast, is nearly completed, and the sequence of the next smallest, a tiny soil-dwelling worm, isn't far behind. Large scale sequencing efforts for human DNA started at several centers a number of months ago and will be in full swing within the year.

The scale of the human DNA sequencing project is enough to send your average Unix system administrator running for cover. From the information-handling point of view, DNA is a very long string consisting of the four letters G, A, T and C (the letters are abbreviations for the four chemical units that form the "rungs" of the DNA double helix ladder). The goal of the project is to determine the order of letters in the string. The size of the string is impressive but not particularly mind-boggling: 3 x 10^9 letters long, or some 3 gigabytes of storage space if you use 1 byte to store each letter and don't use any compression techniques.

Random web quotes: “Perl and the human genome are almost perfectly matched; both are almost incomprehensible, with no central design, accreted haphazardly over a long time.”

https://web.stanford.edu/class/gene211/handouts/How_Perl_HGP.html

Back to today

- Gigabase genomes can be easily sequenced and assembled using PacBio/Nanopore long-read assemblers
  - Often with a N50 value over 10Mb
  - Are orders of magnitude higher than conventional approaches that generate sub-kb reads
- Whole-genome shotgun is the predominant approach to be used today for assembling these large genomes
  - But optical mapping or physical map of specific genetic markers can still provide clue in placing large contigs/scaffolds in genomic regions.

Lander-Waterman statistics in genome assembly

- Originally developed to guide the planning of a genome assembly project (how to select random clones and declare overlap of clones)
- Provide guidelines on expected number of contigs

GENOMICS **2**, 231–239 (1988)

Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis

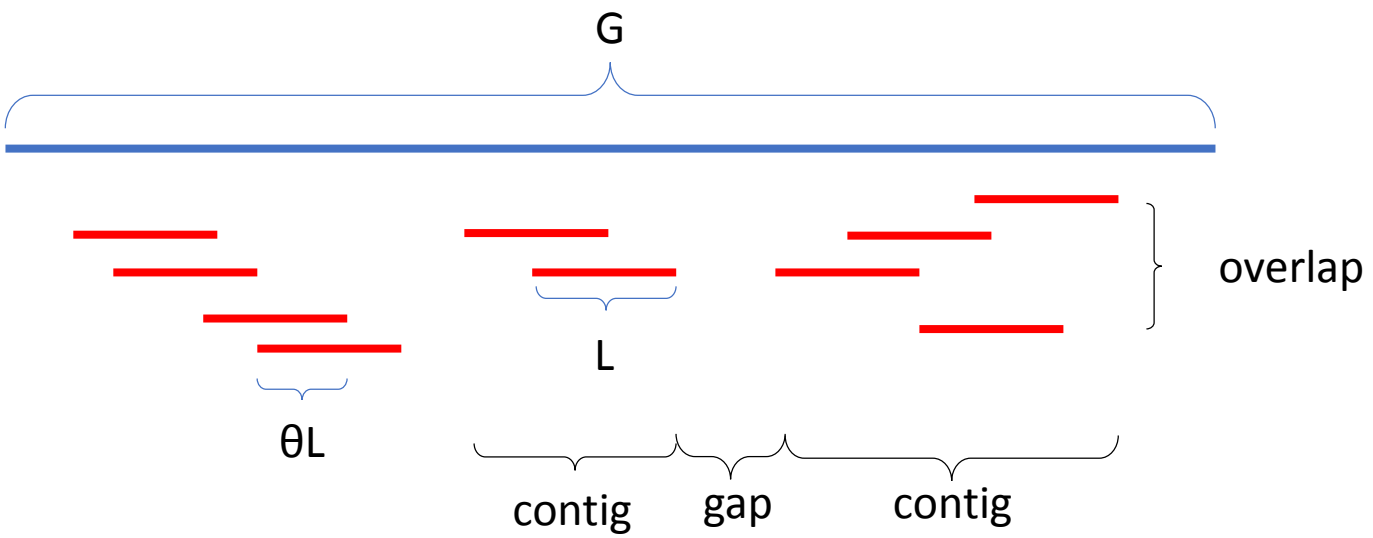
ERIC S. LANDER*,† AND MICHAEL S. WATERMAN‡

*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; †Harvard University, Cambridge, Massachusetts 02138; and ‡Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089

Received January 13, 1988; revised March 31, 1988

Lander-Waterman statistics

- Let N = # of reads, L = length of a read (average length), G = genome length, c = coverage = N∗L/G
- Main questions to address:
  - What's the fraction of genome that are covered by reads?
  - How many contigs are generated?



Coverage at a position can be modelled by Poisson distribution

- c = coverage = N∗L/G
- Use Poisson distribution

$$P(k, \lambda) = e^{-\lambda} \times \frac{\lambda^k}{k!}$$

- Probability that a base is NOT covered = P(0,c) = e^{-c}
- For genome size G, # of uncovered bases = G* P(0,c) = G * e^{-c}

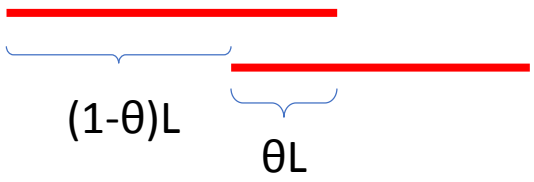
What's the fraction of genome that are covered by reads?

- This is the same question as "how many positions have coverage>0"?

- For genome size G, number of positions with read coverage: G * (1-P(0,c)) = G * (1-e^{-c}).

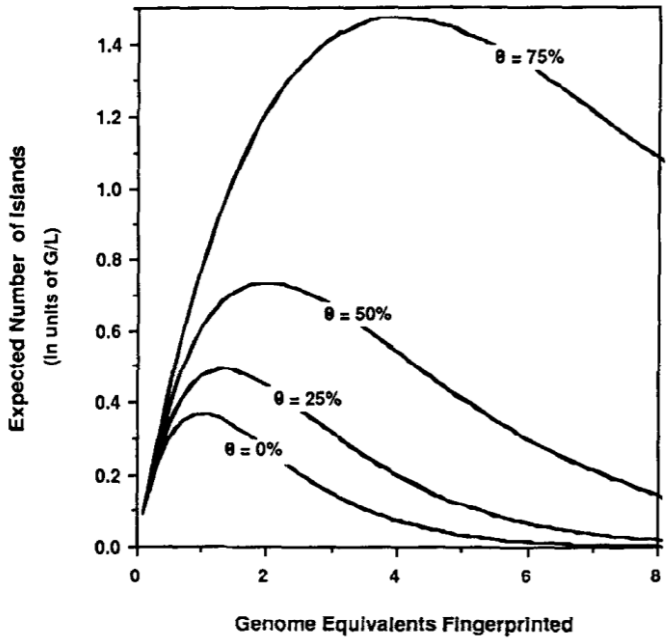
How many contigs are generated?

- How many contigs are generated?
- The same question: how many gaps are there? Or, how many reads do not have overlap > θ with any other reads (Remember that we declare two reads as overlap only if the fraction of overlap > θ)
- For each read, we calculate the probability that zero reads start at (1-θ)L: $e^{-(1-\theta)c} = e^{-(1-\theta)(NL/G)}$
- For all N reads, number of contigs are: $N * e^{-(1-\theta)c}$



Theory is different from reality !

- If there is no repeat, no polymorphism, no region bias, and no sequence error, sequence assembly could be very easy when >8X coverage can be generated even with θ =25%



	Approximate value of G/L		
	Phage (15kb)	Cosmid (40kb)	Yeast (1Mb)
E. coli	267	100	4
S. cerevisiae	1333	500	20
C. elegans	5,667	2,125	85
Human	200,000	75,000	3,000

Lander et al, Genomics, 1988

Assembling large genomes today

- Since 2013, *de novo* assembly of large genomes has shifted from short-read sequencing to synthetic or true long-read sequencing.

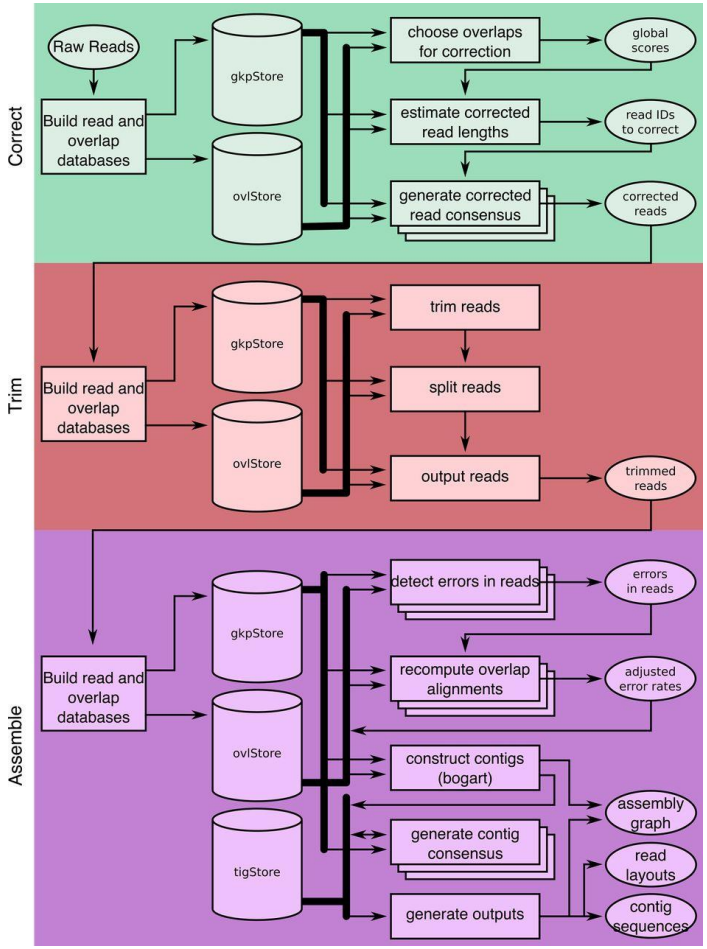
Table 1. Summary of Selected Long-Read Sequencers for *De Novo* Assemblies of Large Eukaryotic Genomes[a,b]

Pros and Cons	10X Genomics Chromium[c] (HiSeq 4000)	Pacific Biosciences (SEQUEL/Cell)	Oxford Nanopore (MinION)	BioNano (Saphyr/Chip)	Dovetail[c] (HiSeq 4000)
Compatible platforms	Illumina	RS II	GridION[d] and PromethION[d]	Irys	Illumina
Minimum input	~3 ng	~20 µg	~1 µg	~200 ng	~5 µg
Long-read	Synthetic	True	True	True	Synthetic
Average/maximum read length	~300 bp (PE)/ ~150 kb (LLR)	~12 kb/~150 kb	~12 kb/~2 Mb	~350 kb/~1 Mb	~150 kb/~1 Mb (SLR)
Throughput	~1500 Gb	0.7 Gb–20 Gb (SEQUEL)	50 Gb–15 Tb (PromethION)	~640 Gb	~1500 Gb
Reads	~5 Billion (B)	0.07 million (M)–2 M	1.5–5 M	~2 M (image file)	~5 B
Runtime	~3 Days	6–10 h	2 h to 6 days	~1 day	~3 days
Quality scores	>30	>10	>10	NA (only nonsequence based method)	>30
Error profile	<1% (GC/AT biased and substitutions)	5–10% (indels)	5–15% (indels and substitutions)	Sizing error, false sites, and missing sites	<1% (GC/AT-biased and substitutions)
Output format	Fasta Fastq	Bam Fasta Fastq Hdf5 (RS II)	Fast5	BNX C/S/XMAP SVMerge TIFF	Fasta Fastq
General assembly software	Supernova	CANU Falcon/Falcon-Unzip Flye HGAP Minimap/Miniasm	CANU Minimap/Miniasm TULIP	RefAligner	3D-DNA HiRise LACHESIS Meraculous SALSA

Jung et al, Trends in Plant Science, 2019

Canu assembler for long-read assembly

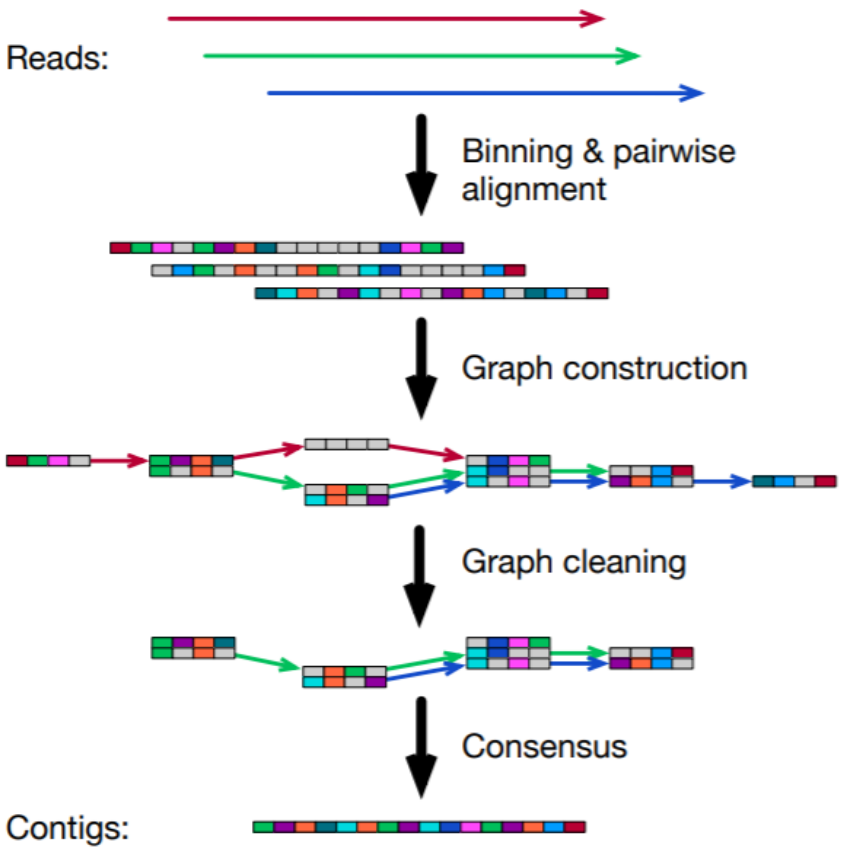
- Canu was branched from Celera Assembler in 2015, and specialized for long-read sequencing
- A full Canu run includes three stages: correction (green), trimming (red), and assembly (purple).



Koren et al, Genome Research, 2017

Wtdbg2 assembler for long-read assembly

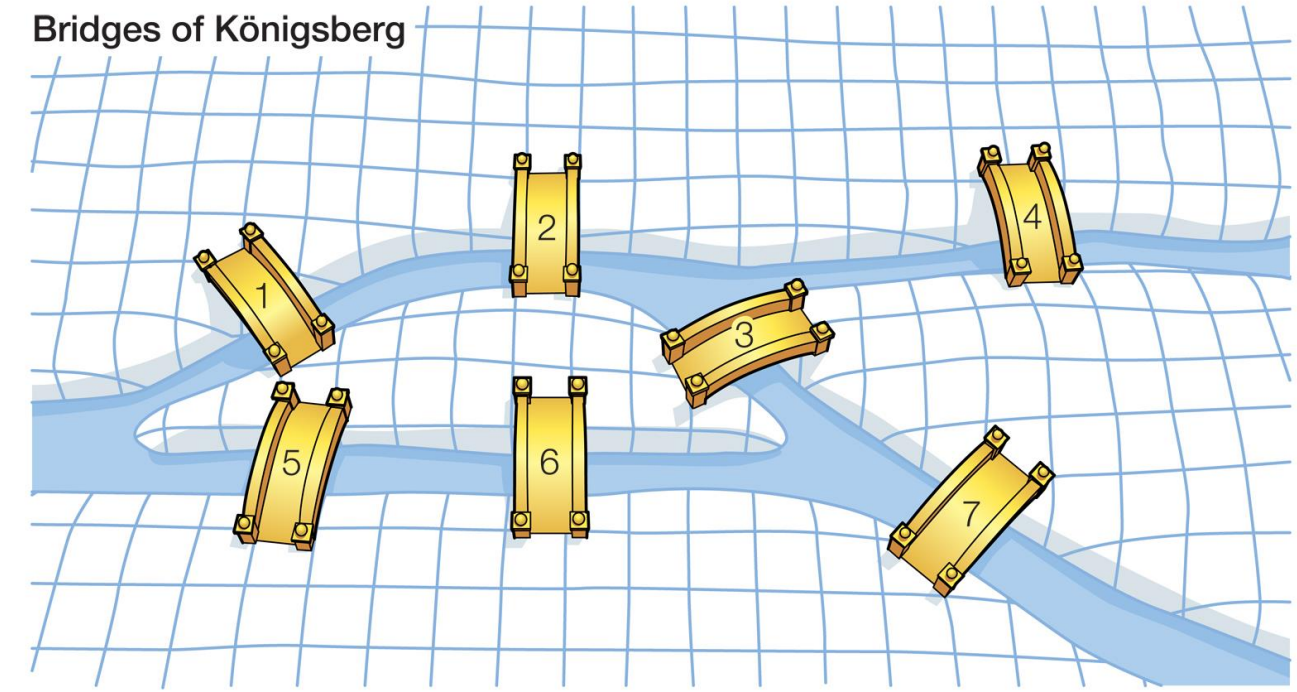
- Wtdbg2 groups 256 base pairs into a bin (bins with the same color suggest they share k-mers)
- Wtdbg2 performs all-vs-all alignment between binned reads, ignoring detailed base sequences.
- In the fuzzy-Bruijn assembly graph, a vertex is a 4-bin segment.
- Much faster than other assemblers for long reads



Ruan et al, BioRxiv, 2019

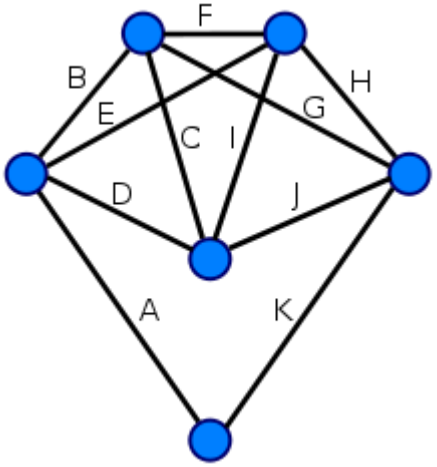
Seven Bridges of Königsberg Problem

- **Königsberg bridge problem**:
  - Whether a citizen could take a walk through the town in such a way that each bridge would be crossed exactly once.



Seven Bridges of Königsberg Problem

- In 1736, the Swiss mathematician Leonhard Euler demonstrated that the answer is no
- In graph theory
  - An **Eulerian trail** (or **Eulerian path**) is a trail in a finite graph which visits every edge exactly once (allowing for revisiting vertices).
  - Similarly, an Eulerian circuit or Eulerian cycle is an Eulerian trail which starts and ends on the same vertex.



Following the edges in alphabetical order gives an Eulerian circuit/cycle.

https://en.wikipedia.org/wiki/Eulerian_path

Euler's Theorem on directed graphs

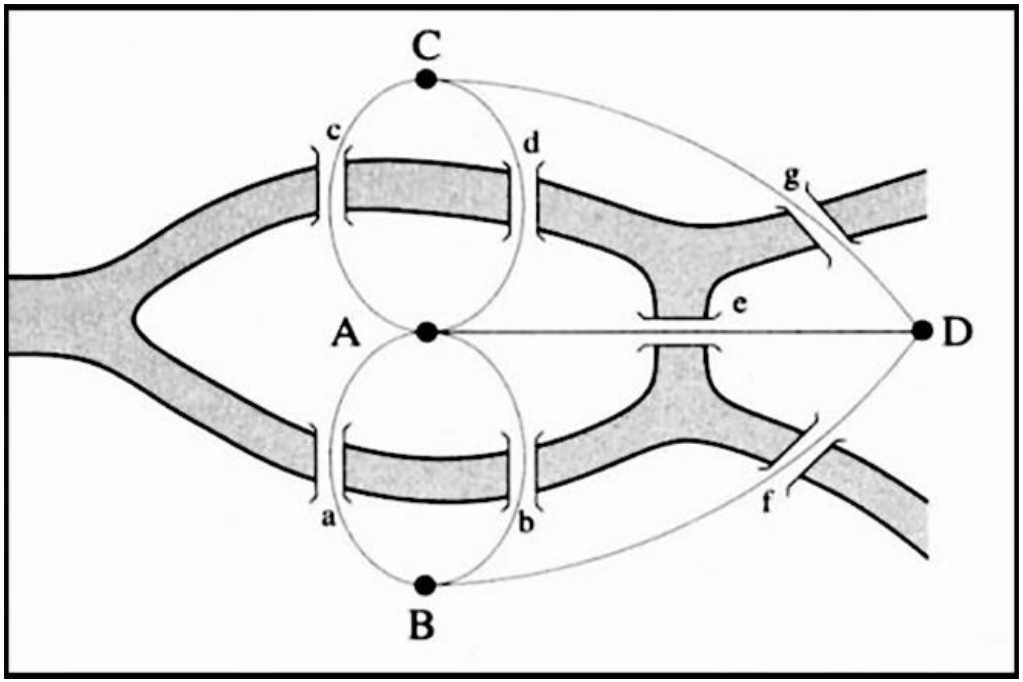
- For directed graphs, the cycle will need to follow the direction of the edges.
  - Indegree(v) = # edges coming into v
  - Outdegree(v) = # edges leaving v
- A directed graph has an Eulerian path if and only if indegree(v)=outdegree(v) for all but 2 nodes (x and y), where indegree(x)=outdegree(x)+1, and indegree(y)=outdegree(y)-1.
- A directed graph has an Eulerian cycle if and only if indegree(v)=outdegree(v) for all nodes

Euler's Theorem on undirected graphs

- For undirected graphs, the degree of a vertex is the number of edges that are incident to the vertex
- An undirected graph has an Eulerian cycle if and only if every vertex has even degree
- An undirected graph has an Eulerian trail if and only if exactly zero or two vertices have odd degree

The Königsberg graph is not Eulerian

- For the existence of Eulerian trails
  - It is necessary that zero or two vertices have an odd degree.
  - If there are no vertices of odd degree, all Eulerian trails are circuits.
  - If there are exactly two vertices of odd degree, all Eulerian trails start at one of them and end at the other.



Hamiltonian vs Eulerian path

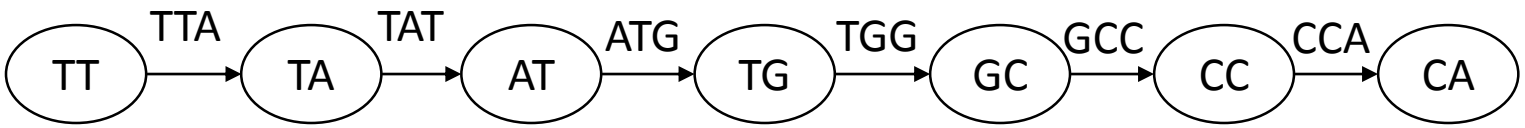
- Hamiltonian path: a path that travels to every node exactly once and ends at the starting node.
- Eulerian path: a path that visits all edges of a graph exactly once.
- When analyzing sequencing data, instead of assigning each k-mer contained in some read to a node, we will assign each such k-mer to an edge. This allows the construction of a **'de Bruijn graph**'.
- The vertices are (k-1)-mers that appear in some read, and edges defined by overlap of k-2 nucleotides.

Constructing the de Bruijn Graph

- Create the de Bruijn graph for the following string, using k=3
  - TT**ATG**CC**ATG**GG**ATG**AA
- Note that TT, CC, GG and AA are separated by 3 identical ATG
- Vertices are 2-mers
- Edges are 3-mers

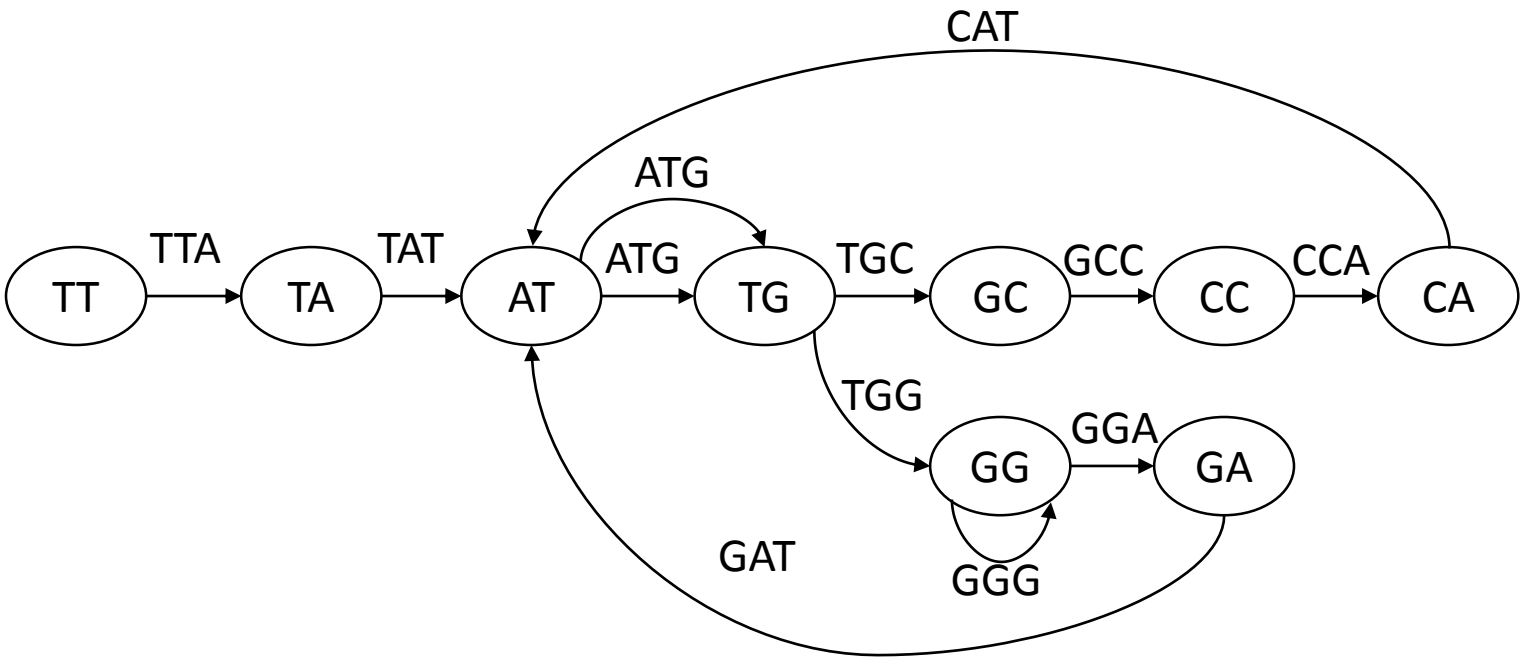
Constructing the de Bruijn Graph

- Create the de Bruijn graph for the following string, using k=3
  - TT**ATG**CC**ATG**GG**ATG**AA



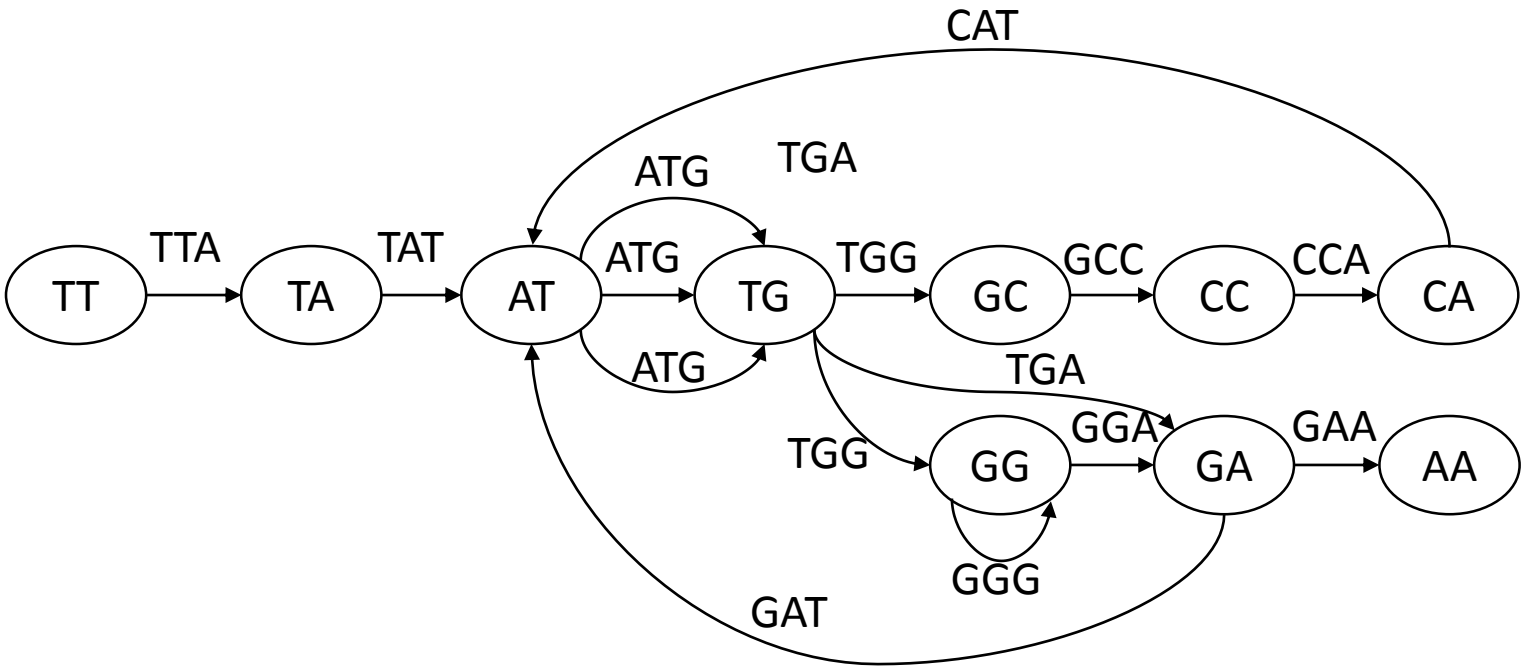
Constructing the de Bruijn Graph

- Create the de Bruijn graph for the following string, using k=3
  - TT**ATG**CC**ATG**GG**ATG**AA



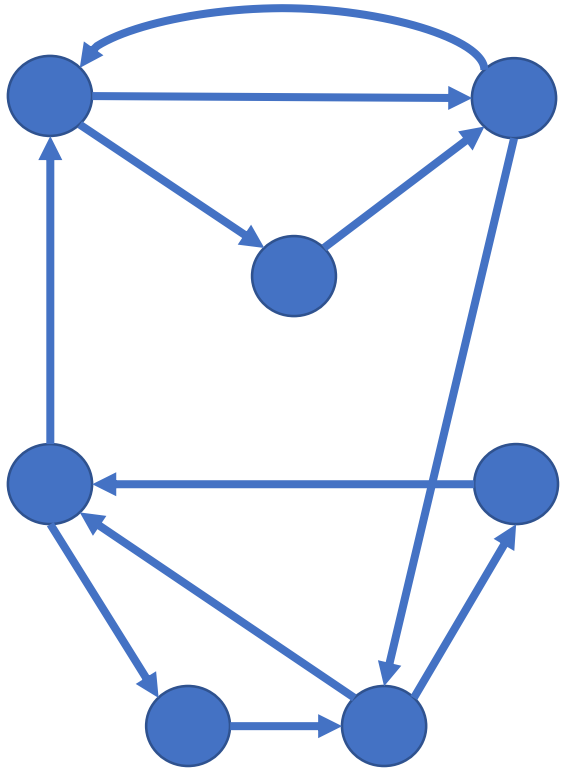
Constructing the de Bruijn Graph

- Create the de Bruijn graph for the following string, using k=3
  - TT**ATG**CC**ATG**GG**ATG**AA



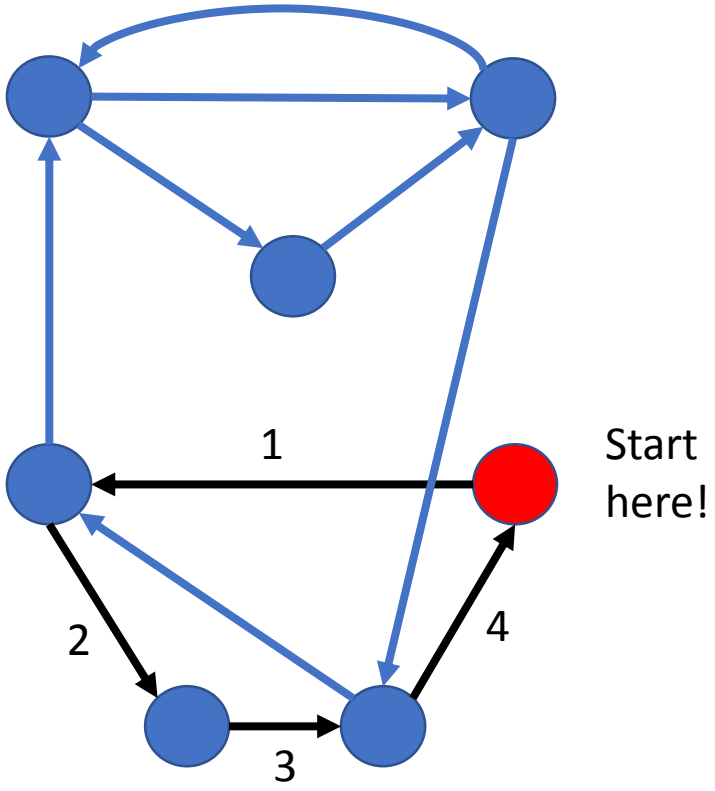
How to find a Eulerian cycle?

- Random walking of an ant in the graph



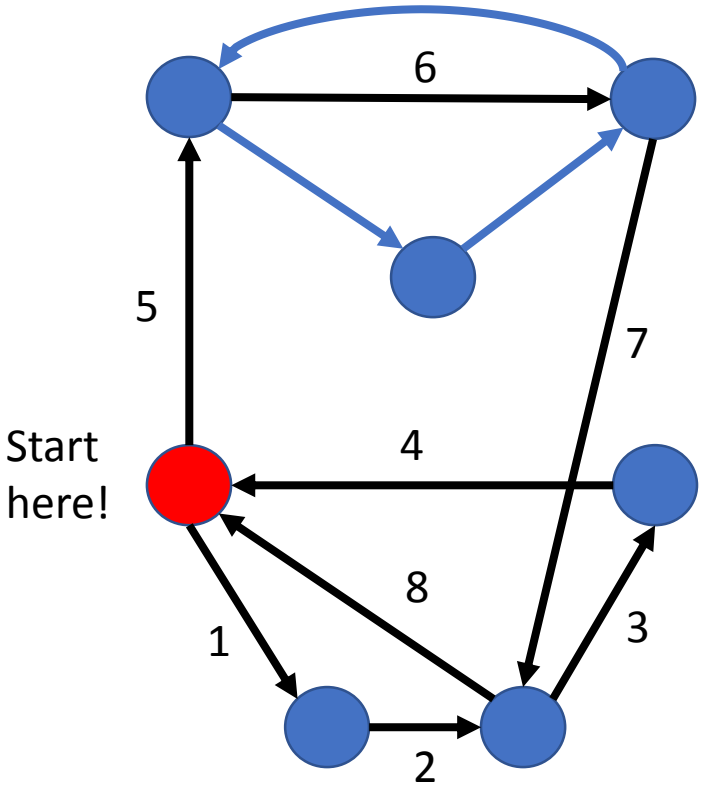
How to generate a Eulerian cycle?

- Random walking of an ant: get stuck in the starting vertex



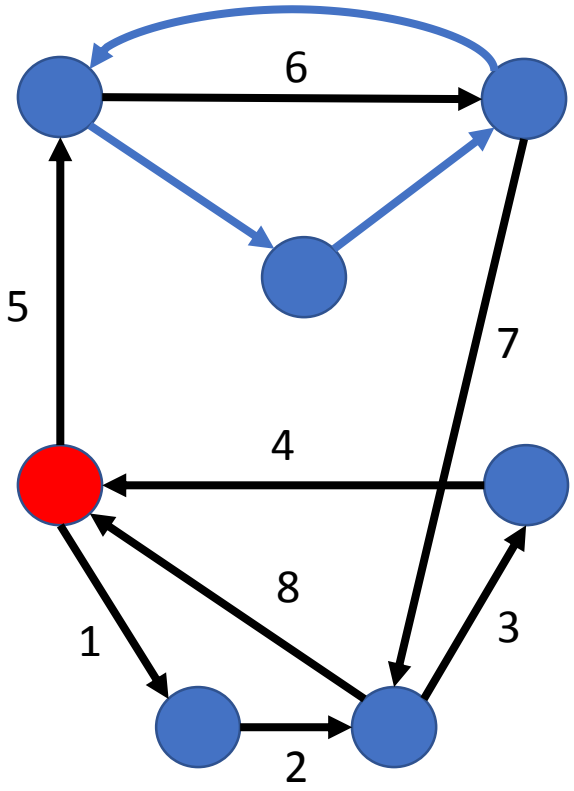
How to generate a Eulerian cycle?

- Random walking of an ant: now start from a different vertex with unused edge, finishing the same cycle first before exploring new path



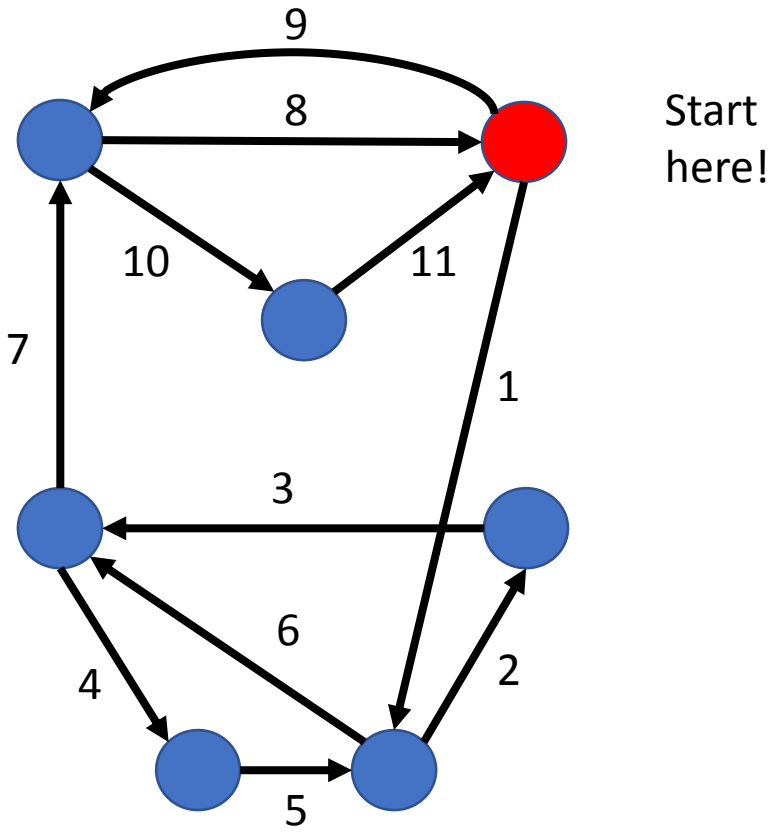
How to generate a Eulerian cycle?

- Random walking of an ant: now get stuck again without finishing all edges



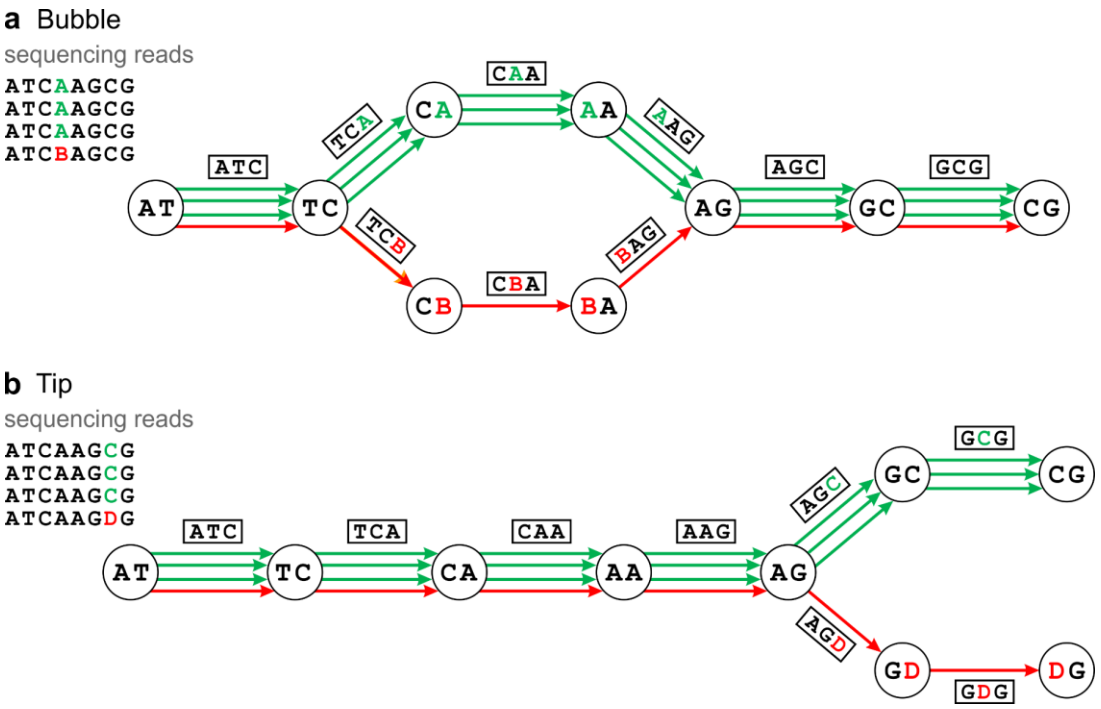
How to generate a Eulerian cycle?

- Random walking of an ant: now start from a different vertex with unused edge, finishing the same cycle first before exploring new path



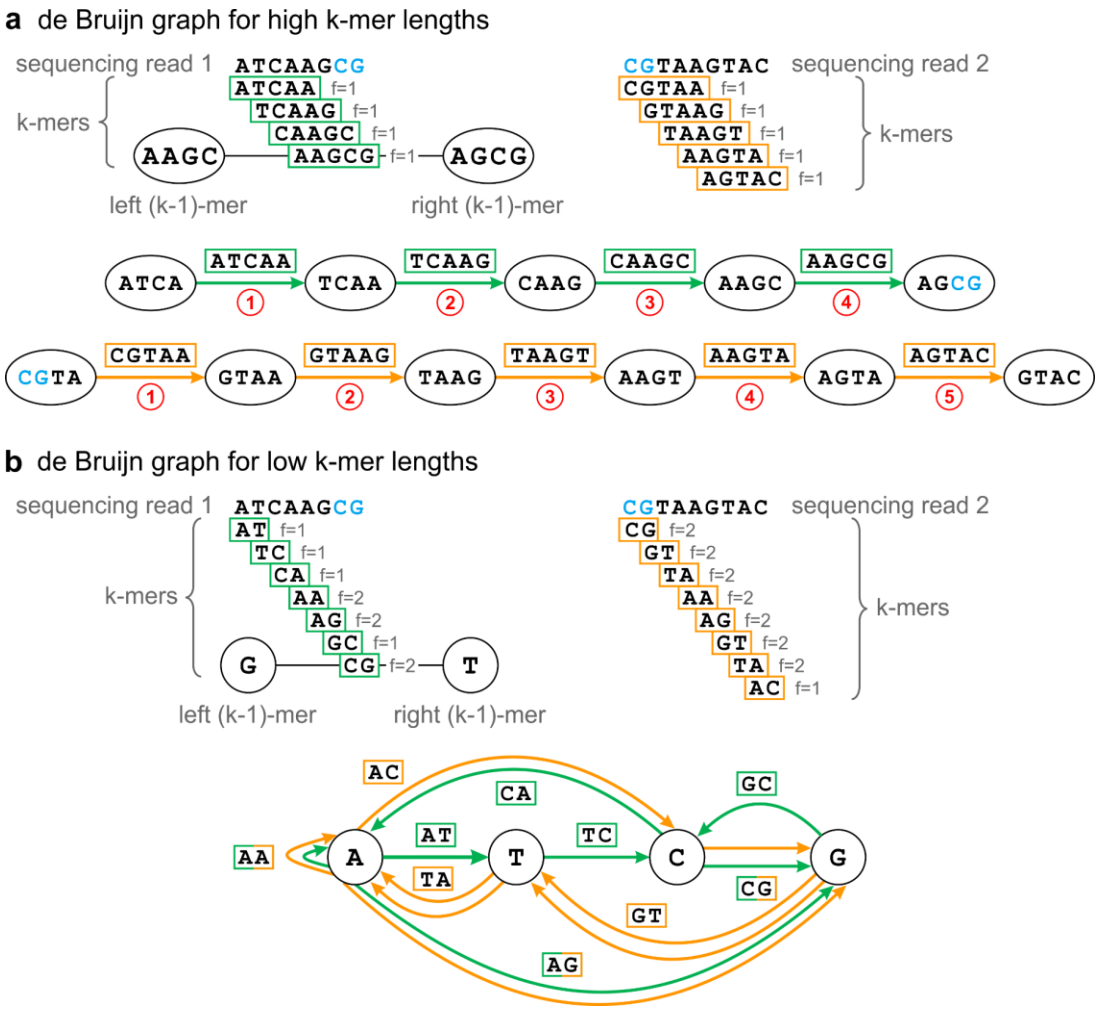
Error correction in de bruijn graph

- Erroneous reads (marked in red) introduce *k* false k-mers to the graph, resulting in additional spurious branches, so-called "bubbles" and "tips"



Vollmers et al, PLoS ONE, 2017

Influence of k-mer length on assembly



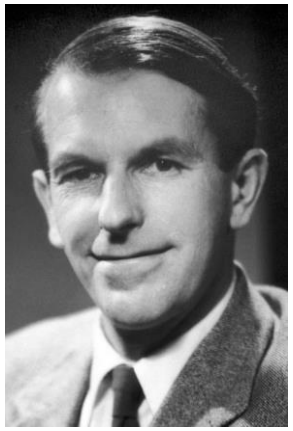
Vollmers et al, PLoS ONE, 2017

What is the appropriate k-mer?

- Tools such as kmerGenie (http://kmergenie.bx.psu.edu/) and Velvet Advisor (dna.med.monash.edu.au/~torsten/velvet_advisor) have been developed to predict the optimal value for k for a given read dataset, based on read length and k-mer frequencies.
- Other strategies may include the merging of assemblies produced with different k-mer sizes or the use of a multi k-mer assembly method.

A fun hypothetical case study: Frederick Sanger's insulin sequencing study

- Frederick Sanger (1918 – 2013) was a British biochemist who twice won the Nobel Prize in Chemistry,
- In 1958, he was awarded a Nobel Prize in Chemistry "for his work on the structure of proteins, especially that of insulin"
- In 1980, Walter Gilbert and Sanger shared half of the chemistry prize "for their contributions concerning the determination of base sequences in nucleic acids"



https://en.wikipedia.org/wiki/Frederick_Sanger

Sequencing insulins

- Sanger determined the complete amino acid sequence of the two polypeptide chains of bovine insulin, A and B, in 1952 and 1951, respectively.
- In determining these sequences, Sanger proved that proteins have a defined chemical composition

Details on sequencing insulin

FREDERICK SANGER

The chemistry of insulin

Nobel Lecture, December 11, 1958

It is great pleasure and privilege for me to give an account of my work on protein structure and I am deeply sensitive of the great honour that has been done to me in recognizing my work in this way. Since the work on insulin has extended over about 12 years it will be necessary to give a somewhat simplified account and to omit most of the work that did not contribute directly to the main problem, the determination of the structure of a protein.

https://www.nobelprize.org/uploads/2018/06/sanger-lecture.pdf

The Amino-acid Sequence in the Phenylalanyl Chain of Insulin

1. THE IDENTIFICATION OF LOWER PEPTIDES FROM PARTIAL HYDROLYSATES

By F. SANGER (Beit Memorial Fellow) and H. TUPPY*
Biochemical Laboratory, University of Cambridge

(*Received* 17 *January* 1951)

When insulin is oxidized with performic acid, the —S—S— bridges of the cystine residues are broken by conversion to —SO_3H groups (Sanger, 1949*a*) and the molecule is split into its separate polypeptide chains. From the oxidized insulin two fractions could be isolated: an acidic fraction *A*, which contained only glycyl *N*-terminal residues (see below) and no basic amino-acids, and a basic fraction *B*, having phenylalanyl *N*-terminal residues. These appeared to be the only significant fractions present.

From a study of the partial hydrolysis products of the dinitrophenyl (DNP) derivatives of the two fractions it was possible to determine the sequence of the amino-acids adjoining the *N*-terminal residues partial hydrolysis might yield considerable information about the overall amino-acid sequence in these fractions. Consden, Gordon & Martin (1947) have described a method for the fractionation of lower peptides using paper chromatography which was successfully used to determine the pentapeptide sequence of 'gramicidin S' (Consden, Gordon, Martin & Synge, 1947). The present paper describes the application of this technique to partial acid hydrolysates of fraction *B* and the determination of a number of amino-acid sequences.

Throughout this paper the abbreviations for the amino-acid residues suggested by Brand & Edsall (1947) are used. These are listed in Table 1. In

Deduction of short peptides

- Sanger partially hydrolysed the insulin into short peptides
  - Either with hydrochloric acid or using an enzyme such as trypsin.
- The mixture of peptides was fractionated in two dimensions on a sheet of filter paper
  - First by electrophoresis in one dimension and
  - Then, perpendicular to that, by chromatography in the other.
- The different peptide fragments of insulin,
  - Detected with ninhydrin, moved to different positions on the paper
  - Creating a distinct pattern that Sanger called "fingerprints".
- By repeating this type of procedure Sanger was able to
  - Determine the sequences of the many peptides generated using different methods for the initial partial hydrolysis.
  - Assembled into the longer sequences to deduce the complete structure of insulin.

The peptides that Sanger found in fraction B

- We want to do a fun exercise:
  - Only Dipeptides and Tripeptides are shown below

Table 15. *Peptides obtained from phenylalanyl chain of insulin (fraction B)*

Dipeptides	Phe.Val (*B* 4β2) Asp.Glu (*B* 1α2) His.Leu (*B* 2γ8) $CySO_3H$.Gly (*B* 1α1) Val.Asp (*B* 1β8) Glu.His (*B* 1γ1) Leu.$CySO_3H$ (*B* 1α6)	Thr.Pro (*B* 1δ8) Lys.Ala (*B* 2γ4)	Gly.Glu (*B* 1δ1) Arg.Gly (*B* 2γ5) Glu.Arg (*B* 1γ3)	Tyr.Leu (*B* 3β10) Val.$CySO_3H$ (*B* 1α4) Leu.Val (*B* 3β7) $CySO_3H$.Gly (*B* 1α1)	Ser.His (*B* 2γ2) Leu.Val (*B* 3β7) Glu.Ala (*B* 1δ4) His.Leu (*B* 2γ8) Val.Glu (*B* 1δ7)	Ala.Leu (*B* 1δ11) Gly.Phe (*B* 1γ12)
Tripeptides	Phe.Val.Asp (*B* 1β13) His.Leu.$CySO_3H$ (*B* 1γ4) Val.Asp.Glu (*B* 1β10) Leu.$CySO_3H$.Gly (*B* 1α5) Glu.His.Leu (*B* 1γ7)	Pro.Lys.Ala (*B* 5γ6)	Gly.Glu.Orn (*B* 5γ1) [Glu, Arg, Gly] (*B* 1γ2)	Tyr.Leu.Val (*B* 3β12) Leu.Val.$CySO_3H$ (*B* 1α8) Val.$CySO_3H$.Gly (*B* 1α3)	Ser.His.Leu (*B* 2γ7) Val.Glu.Ala (*B* 1γ10) Leu.Val.Glu (*B* 2β14)	Ala.[Tyr, Leu] (*B* 3β9)

Fun exercise: assemble insulin from dipeptides and tripeptides for Sanger

- What if Sanger knew de Bruijn graph in 1951? Can he use the information below to assemble isulin to a few contigs?

Dipeptide	Tripeptide
Ala.Leu	Glu.Arg.Gly
Arg.Gly	Glu.His.Leu
Asp.Glu	Gly.Glu.Arg
Cys.Gly	His.Leu.Cys
Cys.Gly	Leu.Cys.Gly
Glu.Ala	Leu.Val.Cys
Glu.Arg	Leu.Val.Glu
Glu.His	Phe.Val.Asp
Gly.Glu	Pro.Lys.Ala
Gly.Phe	Ser.His.Leu
His.Leu	Tyr.Leu.Val
His.Leu	Val.Asp.Glu
Leu.Cys	Val.Cys.Gly
Leu.Val	Val.Glu.Ala
Leu.Val	
Lys.Ala	
Phe.Val	
Ser.His	
Thr.Pro	
Tyr.Leu	
Val.Asp	
Val.Cys	
Val.Glu	