Graphs in Computational Biology

EN 500.111 Week 6

Class Presentation

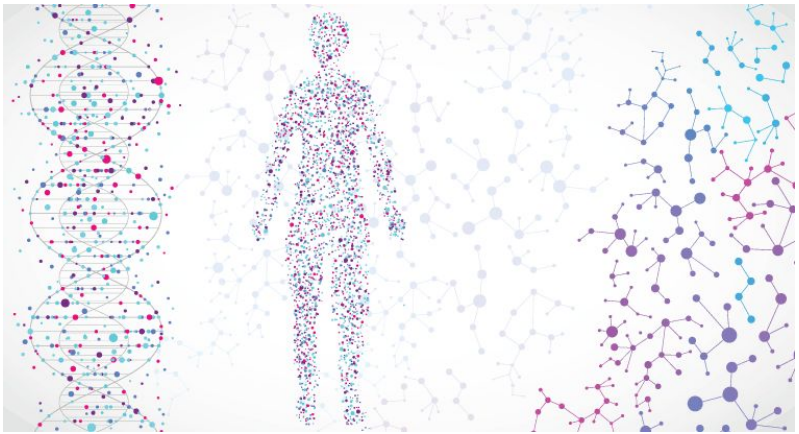
- Each group please let me know the following:
  - Group members
  - Group name
  - Presentation topic
- Guidelines here: https://github.com/mkirsche/TAG2020/blob/master/presentation/guidelines.md

Outline

- Basics of genomics and cell biology
- Genomic sequencing
- Genome assembly
  - De Bruijn graphs
  - Overlap graphs
- Gene interaction networks

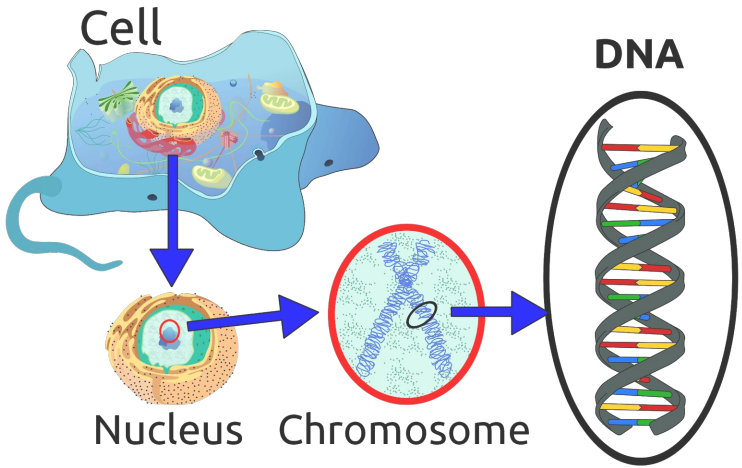
What is genomics?

- ➔ “Genomics is an interdisciplinary field of biology focusing on the structure, function, evolution, mapping, and editing of genomes.” -Wikipedia



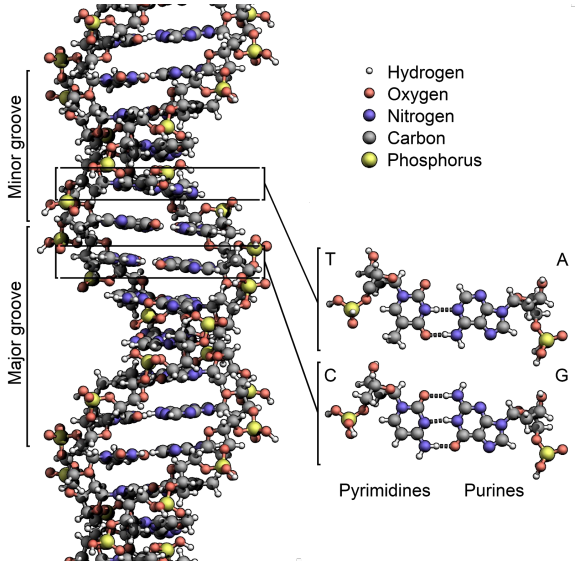
What is a genome?

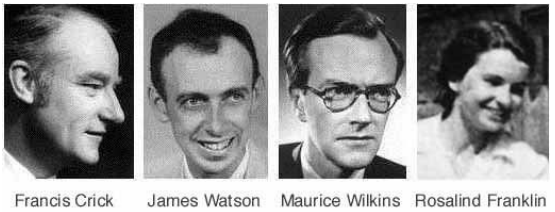
- All organisms are made up of cells (humans have ~37 trillion)
- Each cell has a copy of that organism's DNA
- The genome is the genetic material making up the DNA



DNA is a sequence of nucleotides

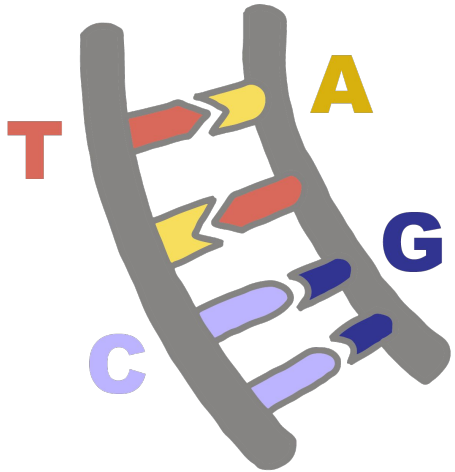
- DNA is a molecule consisting of a smaller unit called nucleotides
- These nucleotides are formed into two chains, or strands, which are wrapped around each other in a spiral
- There are 4 different nucleotides:
  - Adenine (A)
  - Cytosine (C)
  - Guanine (G)
  - Thymine (T)
- The sequence of nucleotides, or bases, in DNA can be thought of as a code for the organism's genetic information





Complementary Bases

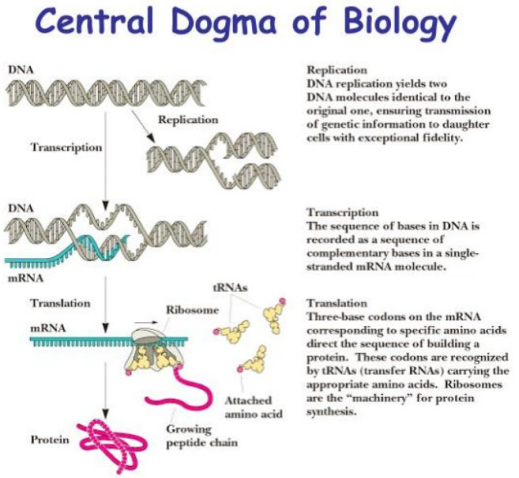
- Thymine (T) and adenine (A) are called complementary nucleotides, meaning that if one of them occurs on one strand, it's always linked to the other one on the opposite strand
- The same is true for cytosine (C) and guanine (G)
- Because of this, if we know the sequence of nucleotides on one strand, the other strand's sequence can be inferred



https://innovativegenomics.org/glossary/complementary/

DNA codes for specific proteins

- ➔ The sequence of bases in DNA is transcribed into RNA, a similar but single-stranded molecule.
- ➔ Specific sections of the RNA are translated into proteins, which are responsible for most of the cell's structure, function, and regulation
- ➔ A gene is the portion of DNA which encodes a single protein



https://www.slideserve.com/shen/central-dogma-of-biology

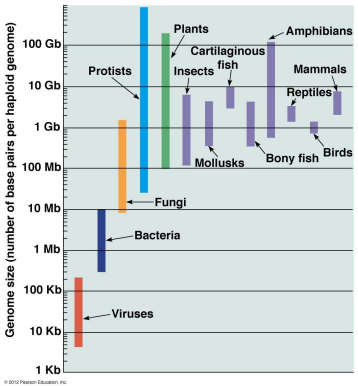
Back to the definition of genomics

- ➔ A genome can be thought of as the sequence of nucleotides making up the DNA of a cell
- ➔ It is typically the same for all cells in an individual, but the genome is different for different individuals (identical twins are an exception to this) and for different species
- ➔ In genomics we try to study these sequences and understand what they do and how they evolve

https://medium.com/@Genomesio/the-importance-of-whole-genome-sequencing-c7f9271a9c74

How big are these nucleotide sequences?

- ➔ Usually about the same for all individuals within a species
- ➔ The smallest known genome is a bacteria with ~159,000 nucleotides
- ➔ The human genome is about 3,000,000,000 (3 billion) nucleotides
- ➔ The largest known genome is a Japanese flower with ~150 billion nucleotides
- ➔ Largely unrelated to the size of the organism



https://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-18/CB18.html

The cat genome is about 2.7 billion basepairs.

How much do they differ?

The DNA sequence that can be directly **compared** between the two **genomes** is almost 99 percent identical. When DNA insertions and deletions are taken into account, **humans** and **chimps** still share 96 percent of their sequence. At the protein level, 29 percent of genes code for the same amino sequences in **chimps** and **humans**. Aug 31, 2005

www.genome.gov › 2005-release-new-genome-comparis...

New Genome Comparison Finds Chimps, Humans Very ...

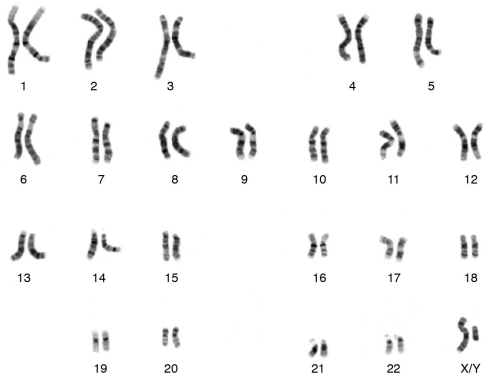
There are more than three million **differences** between your genome and anyone else's. On the other hand, we are all 99.9 percent the same, **DNA**-wise. (By contrast, we are only about 99 percent the same as our closest relatives, chimpanzees.) Jan 15, 2003

www.genomenewsnetwork.org › resources › Chp4_1

human genome variation - What's a Genome?

How is the genome passed on?

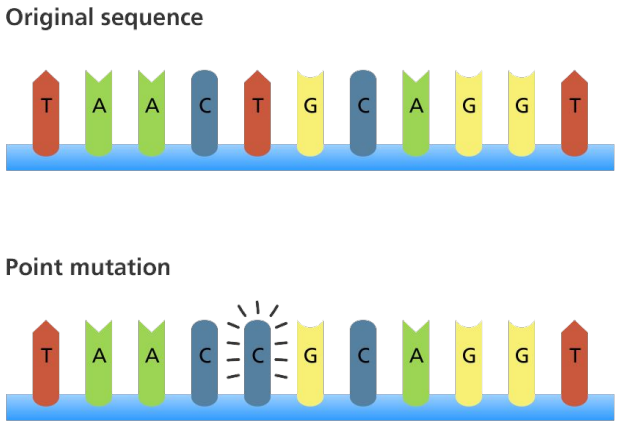
- The genome is segmented into pieces called chromosomes (humans have 23 pairs)
- Each individual has 2 copies of each chromosome, one acquired from each parent
- A small number of mutations, or changes in the nucleotide sequence, occur from generation to generation



https://www.genome.gov/genetics-glossary/Karyotype

Effects of genetic variation

- ➔ Genome differences are responsible for many different traits such as eye color, freckles, and inherited diseases
- ➔ Some traits are affected by a very specific known mutation, others are affected by combinations of mutations, and still others are known to be genetically linked but with unknown causes



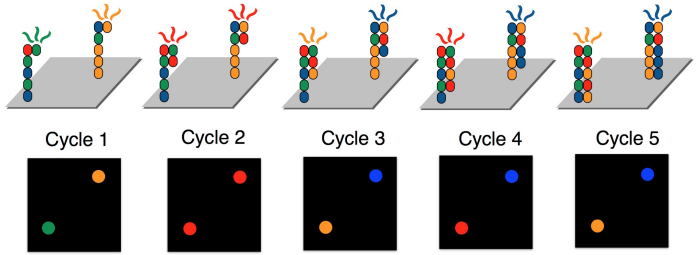
https://www.yourgenome.org/facts/what-types-of-mutation-are-there

Genome sequencing

- ➔ To study genetic variation, we want to know what these sequences are
- ➔ But DNA is much too small and fragile to just read this sequence
- ➔ Decades of research have gone into techniques to read even small pieces of the sequence

Illumina Sequencing

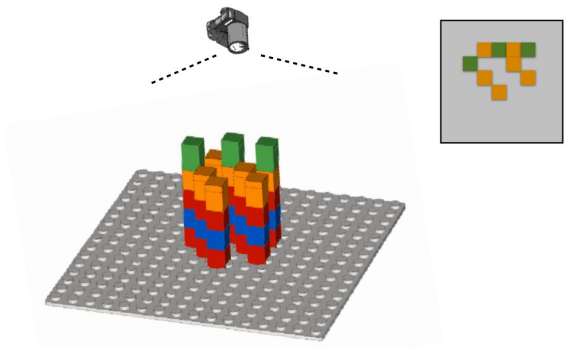
- ➔ The most commonly used protocol for DNA sequencing
- ➔ Relies on DNA's mechanism for synthesizing a second strand from the first one
- ➔ Procedure
    1. Break DNA into small pieces (~500 bp)
    2. Make a large number of copies of each DNA fragment
    3. Separate the two strands and attach the end of one strand to a plate
    4. Add modified nucleotides to the plate which are tagged with "colored lights" with different colors for A/C/G/T
    5. Take pictures of the plate as each nucleotide in the second strand is synthesized and record the color



http://data-science-sequencing.github.io/Win2018/lectures/lecture2/

Other properties of Illumina sequencing

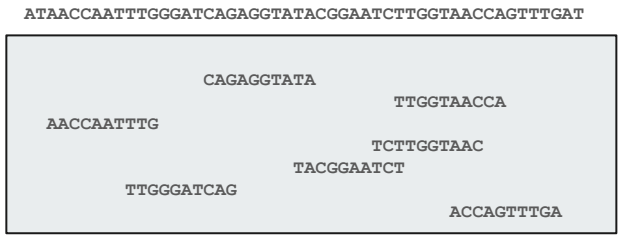
- ➔ Highly parallelizable - possible to read from **millions of fragments at once** in a single plate
- ➔ Highly accurate - identifies the nucleotides with **>99.9% accuracy**
- ➔ But **accuracy gets worse closer to the end of the read** and this is the main reason the fragments are only 500 bp long



http://www.cs.jhu.edu/~langmea/resources/lecture_notes/02_dna_sequencing_v2.pdf

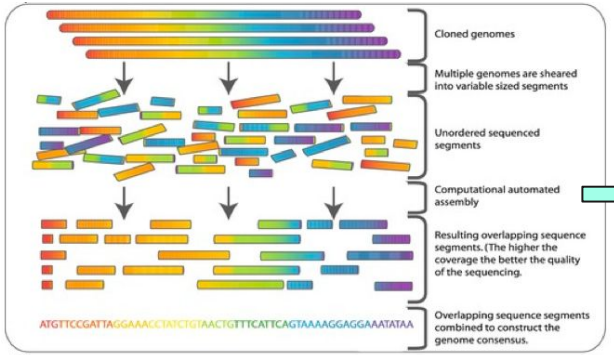
Sequencing Coverage

- ➔ When we sequence the genome, there's no way to tell which segments of the genome came from where, and they're essentially random
- ➔ If we sequence 3 billion bp total, there's no guarantee we'll hit every basepair exactly once (i.e., reads may overlap)
- ➔ So typically people sequence much more than this - e.g., 90 billion bp or 30x coverage

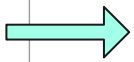


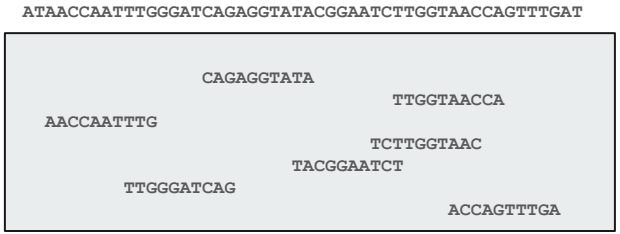
In this example we sequenced 70 total basepairs (7 reads of length 10), but still didn't cover the entire 50 bp genome

Genome Assembly



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





Genome: 3 billion basepairs (bp)

Sequencing reads: 500 bp each
(1 / 6,000,000 of the genome)

Goal: Figure out the full genome sequence from the sequenced fragments

https://www.centerforhealingneurology.com/2018/03/23/focus-on-epigenetics/

Genome Assembly Example

Given the following reads, find an example of a genome that the reads could come from.

```
GAGCG
GCGCT
CGCTA
GACTG
ATGAC
AGCGC
GCTAT
TGACT
```

Shortest common superstring

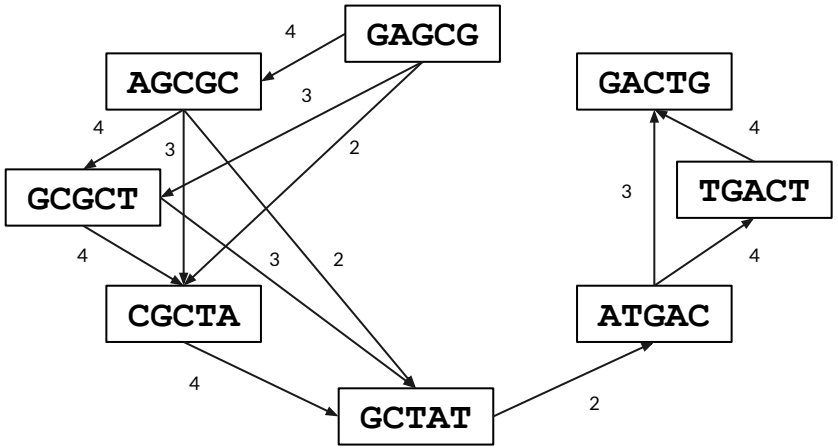
- ➔ Usually in genome assembly we favor the shorter solutions because we assume we have high coverage and that the reads will overlap
- ➔ This can be stated as the shortest common superstring problem, or that of finding the shortest genome (string of characters) that contains all of the reads
- ➔ A few different graph-based approaches have been taken to this, and they both revolve around the idea of using as many overlaps as possible.

Squishing overlaps

- ➔ The main idea in solving the shortest superstring problem is compressing overlaps
- ➔ For example, suppose we had only two reads: **GCGCT** and **CGCTA**
- ➔ The first read ends with “CGCT” and the second read starts with “CGCT”, so we can line those up and only have that portion once: GCGCTA
- ➔ By chaining together overlaps like this we can make the solution much shorter

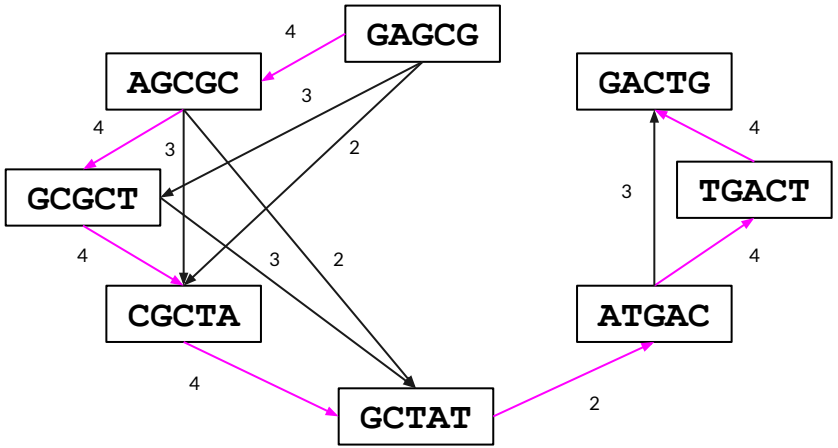
Approach 1: String overlap graph

- ➔ The more straightforward but less efficient approach
- ➔ Nodes are reads
- ➔ Edges between pairs of reads if the end of one read matches the start of another (weighted according to overlap length)
- ➔ An assembly is represented as a path through the graph which visits every single node
- ➔ How would you reconstruct an assembly (string containing all of the reads) from a path?



Assembly from overlap graph path

➔ How would you reconstruct an assembly (string containing all of the reads) from a path?



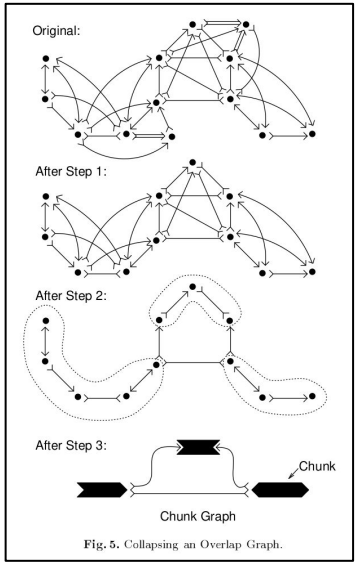
How to improve on this?

1. Apply heuristics to get a not-necessarily-optimal path
2. Reduce the size of the graph
3. Formulate the problem differently and use a different type of graph representation

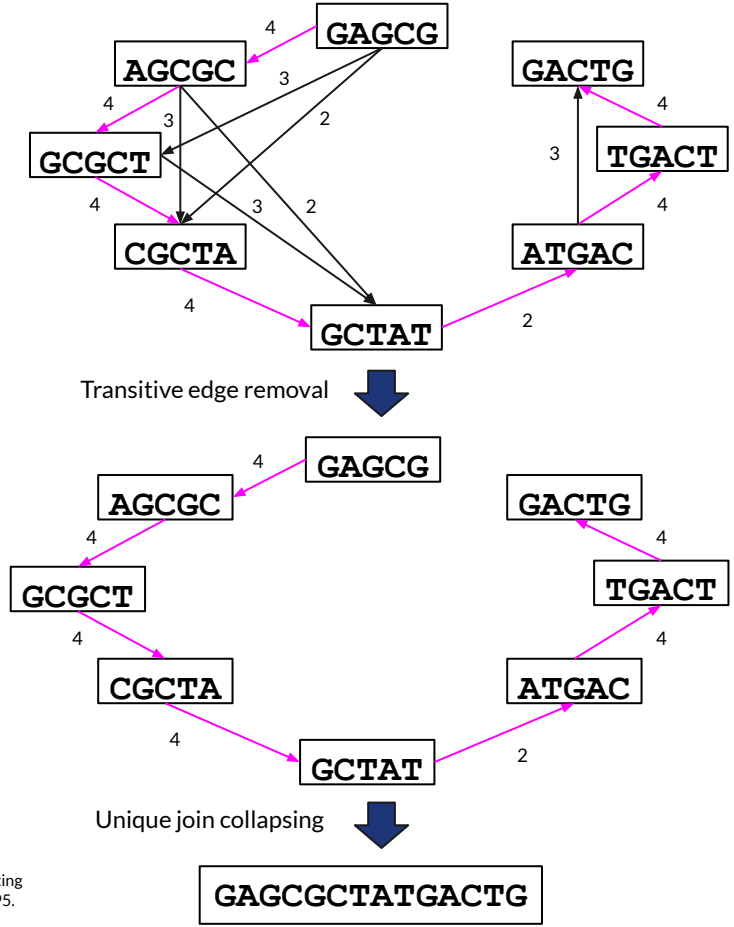
Both 2 and 3 are done in practice

Overlap graph simplification

1. Contained read removal: remove any read whose sequence is contained in another read (this only makes sense if the reads are of different lengths)
2. Remove transitive edges, or edges which can be replaced with a different path
3. Unique join collapsing: combine nodes which only have one edge going in or out

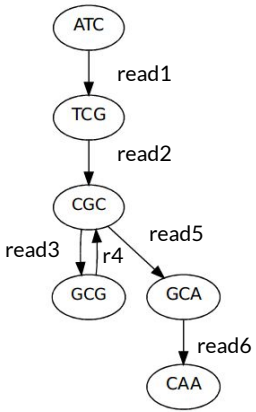


Myers, Eugene. Toward Simplifying and Accurately Formulating Fragment Assembly. Journal of Computational Biology, 1995.



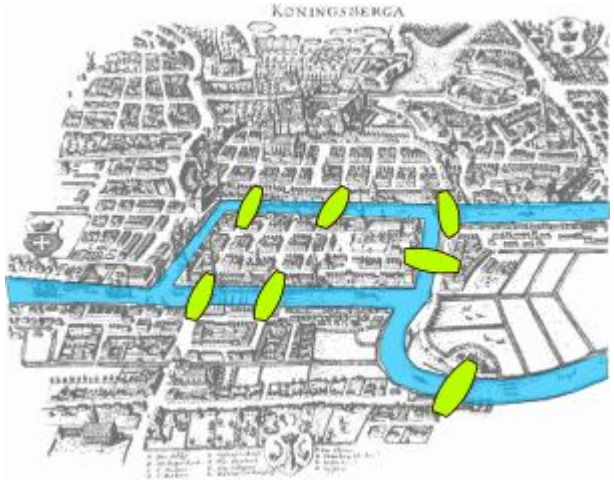
De Bruijn Graphs

- What if we reverse the definitions of nodes and edges
- In overlap graphs, nodes were reads and edges were overlaps, so we'll try making the edges reads and the nodes overlapping sequences
- Suppose we have the following reads:
  - ATCG
  - TCGC
  - CGCG
  - GCGC
  - CGCA
  - GCAA
- How can we translate paths in this graph into genomes?
- What properties would be want to look for in a path representing the shortest common superstring?



Goal: Path with all of the edges exactly once

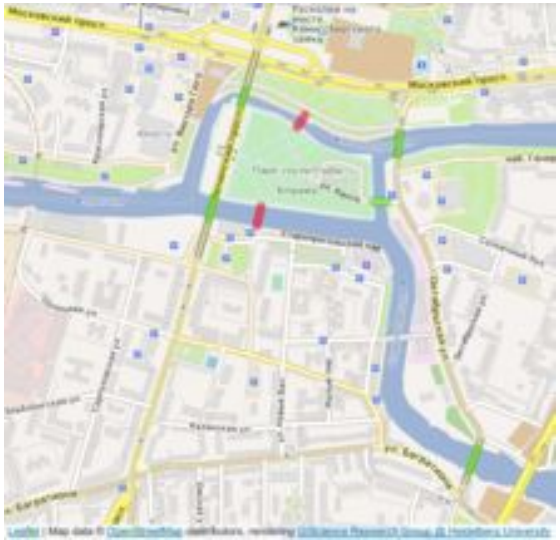
- ➔ This was presented by Euler in the 1730s as a famous mathematical problem called the “seven bridges of konigsberg”
- ➔ The goal is to find a path which walks through the city and crosses all seven bridges exactly once
- ➔ Does a solution exist? Why or why not?



https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

Five bridges of Konigsberg

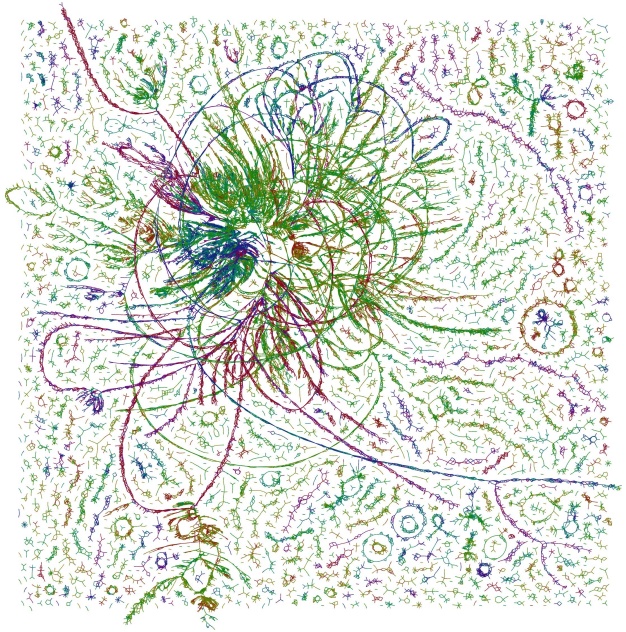
- ➔ Some of the bridges have been demolished or rebuilt since Euler's time
- ➔ Is a path which crosses each bridge once possible now? What properties must a city/graph have for this kind of path to exist?



https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg

What do these graphs look like in practice?

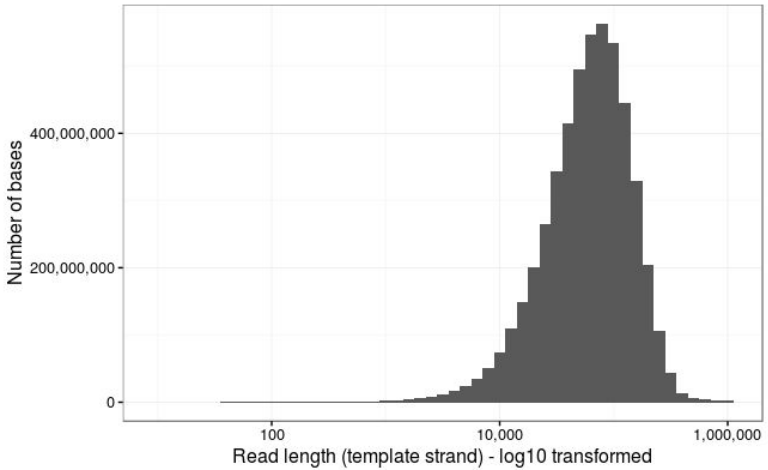
Very very messy - what might cause the weird shapes and disconnectedness?



https://armbrustlab.ocean.washington.edu/tools/seastar/

Future of genome assembly

- **Adapting to new read technology**
  - Illumina sequencing has 99.9% accuracy but length of only 500, which performs poorly in highly repetitive genomic regions
  - Newer data types such as Pacbio and Nanopore sequencing have lower accuracy (~95%) but read lengths averaging well over 10,000 bp with some reads over 1 million bp



https://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

Future of genome assembly

- **Improving the human genome**
  - Human genome project sequenced our genome back in the early 2000s, but it was full of gaps and errors and has been steadily improved since then
  - Even now there's still work to be done - last month the T2T consortium released a new build of the human genome with 100,000,000 bp of previously unresolved sequence
  - Effort involved a combination of many read technologies plus some manual assembly

The (near) complete sequence of a human genome

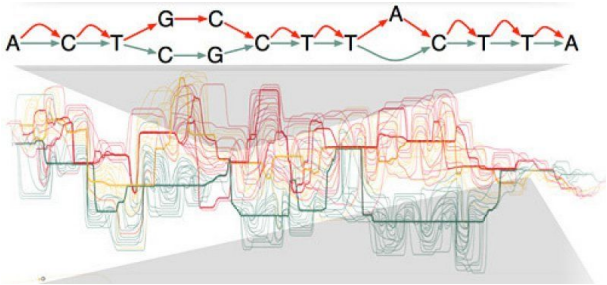
Adam Phillippy

September 22, 2020

The Telomere-to-Telomere (T2T) consortium is proud to announce our v1.0 assembly of a complete human genome. This post briefly summarizes our work over the past year, including a month-long virtual workshop in June, as we strove to complete as many human chromosomes as possible. Our progress over the summer exceeded our wildest expectations and resulted in the completion of *all* human chromosomes, with the only exception being the 5 rDNA arrays. Our v1.0 assembly includes more than 100 Mbp of novel sequence compared to GRCh38, achieves near-perfect sequence accuracy, and unlocks the most complex regions of the genome to functional study. We plan to release a series of preprints in the coming months that fully describe our methods and analyses, but due to its tremendous value, we are releasing the assembly immediately.

Future of genome assembly

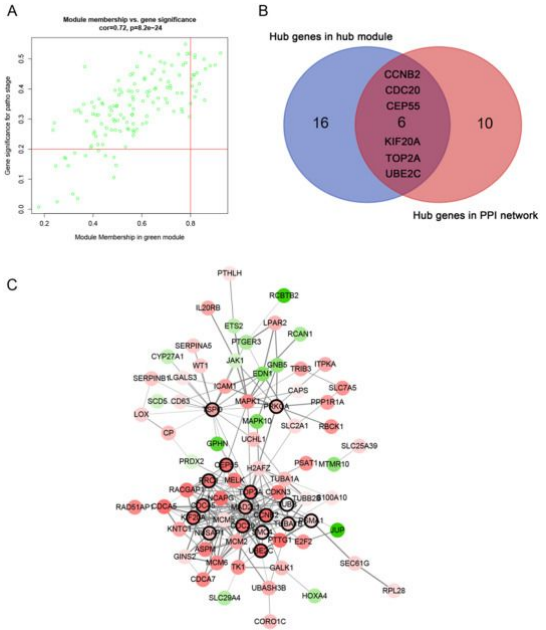
- **Population genomes**
  - Because of differences in individuals' genomes, it is often useful to have multiple genomes within a species so that our scientific knowledge is not biased towards a particular group of people
  - Ongoing field of research is finding ways to represent multiple genomes compactly
  - The current state-of-the-art is graph genomes, which we'll talk more about next week



https://geneticliteracyproject.org/2016/10/11/genome-graph-offers-way-scientists-map-human-gene-pool/

Gene Interaction Networks

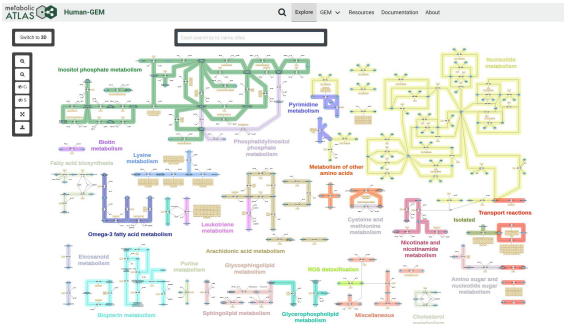
- ➔ While all genes in an individual have the same genome and code for the same proteins, environmental factors affect how much each protein is produced
- ➔ By sequencing RNA instead of DNA, it is possible to measure the extent to which a cell is translating different proteins (i.e., expressing the genes which code for them)
- ➔ By correlating these genes' expressions with one another, it's possible to infer pathways, identify functionally significant genes, and classify cell types

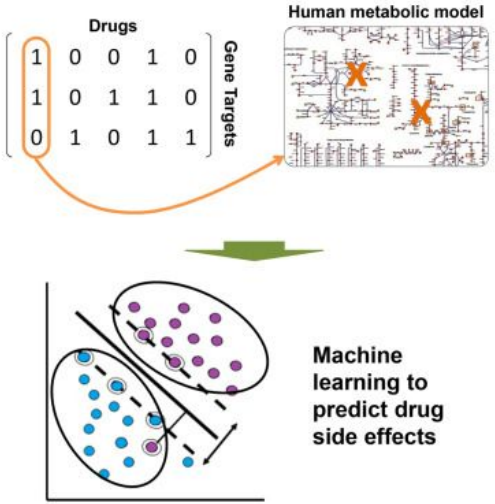


Yuan, Lushun et. al. Co-expression network analysis identified six hub genes in association with progression and prognosis in human clear cell renal cell carcinoma (ccRCC). Genomics Data, 2017.

Using metabolic pathway models

➔ Once pathways have been predicted from gene expression data or other experiments, they can be combined with machine learning techniques to help inform drug response

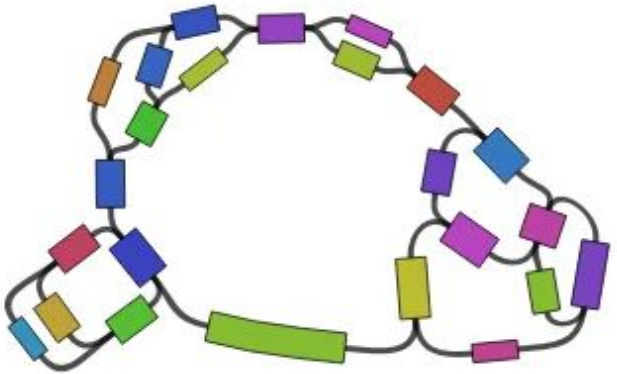




Shaked, Itay et. al. Metabolic Network Prediction of Drug Side Effects. Cell Systems, 2016.

Conclusion

- ➔ Genome assembly is a problem centered around overlaps between genomic reads
- ➔ Multiple approaches involving graph representations and algorithms are being used to work on this problem
- ➔ The field of genome assembly is constantly changing as new sequencing technologies are developed and become accessible
- ➔ Graphs can also be used to represent gene co-occurrences and predict functional pathways



https://www.molecularecologist.com/2017/12/visualize-your-genome-assemblies/

Any questions?