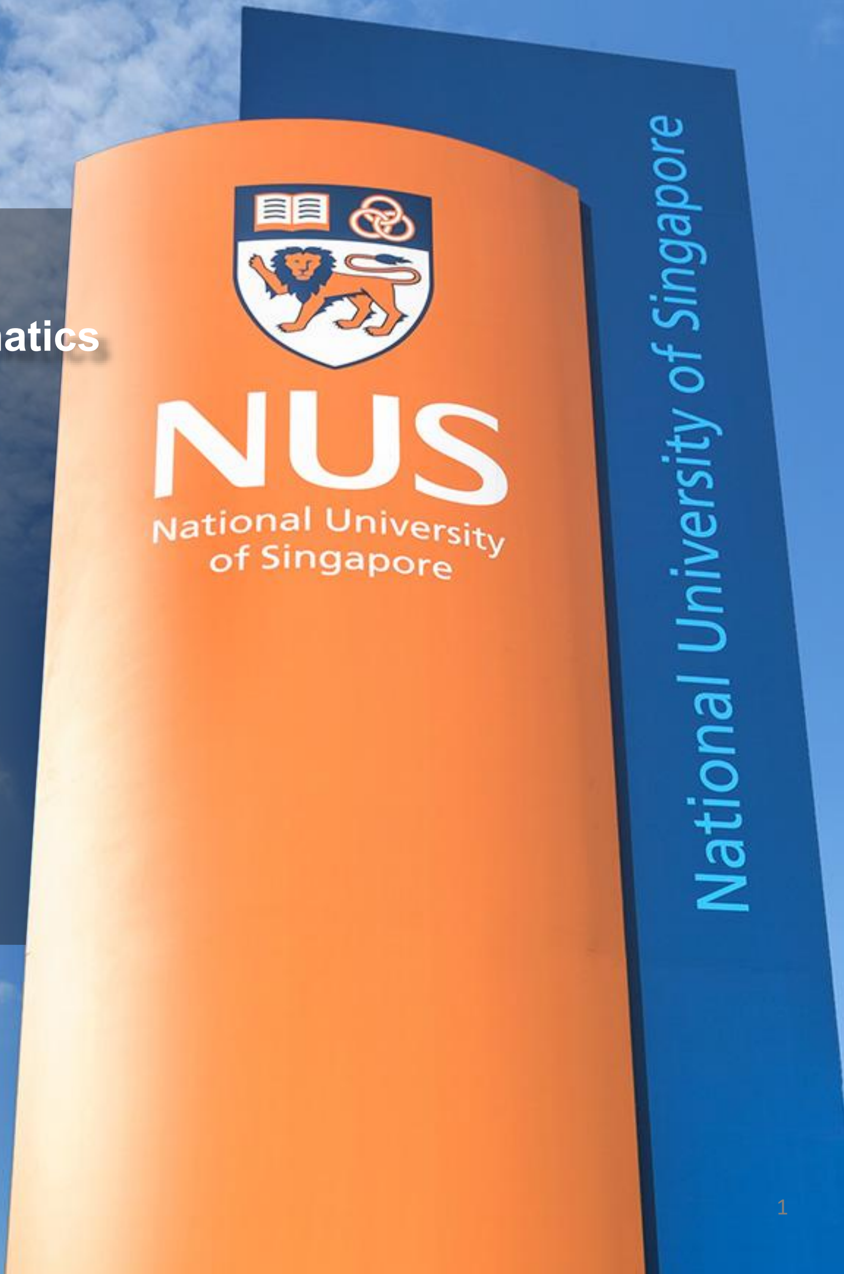
CS4330: Combinatorial Methods in Bioinformatics

Primer on de novo genome assembly

Wong Limsoon


NUS
National University of Singapore
National University of Singapore

Types of genome assembly

Sequencing produces reads which are quite short

These reads need to be assembled into a genome

De novo assembly

Mapping or reference-based assembly

Reference-guided assembly for long reads

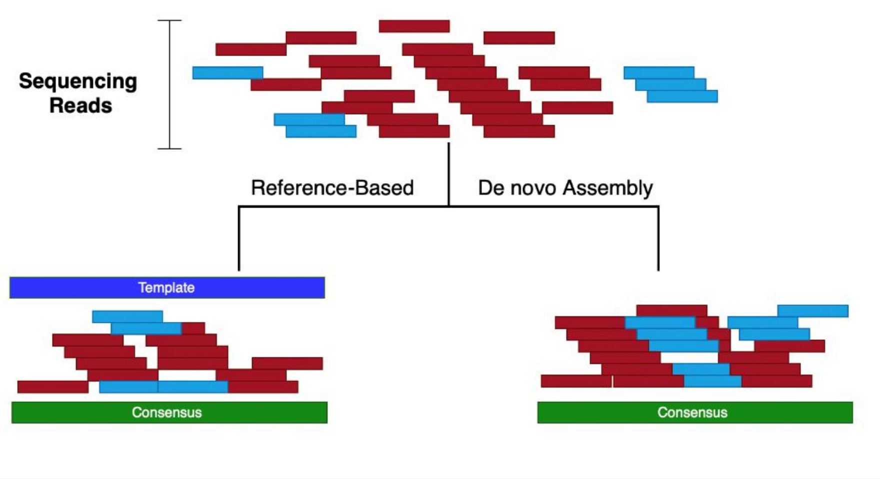
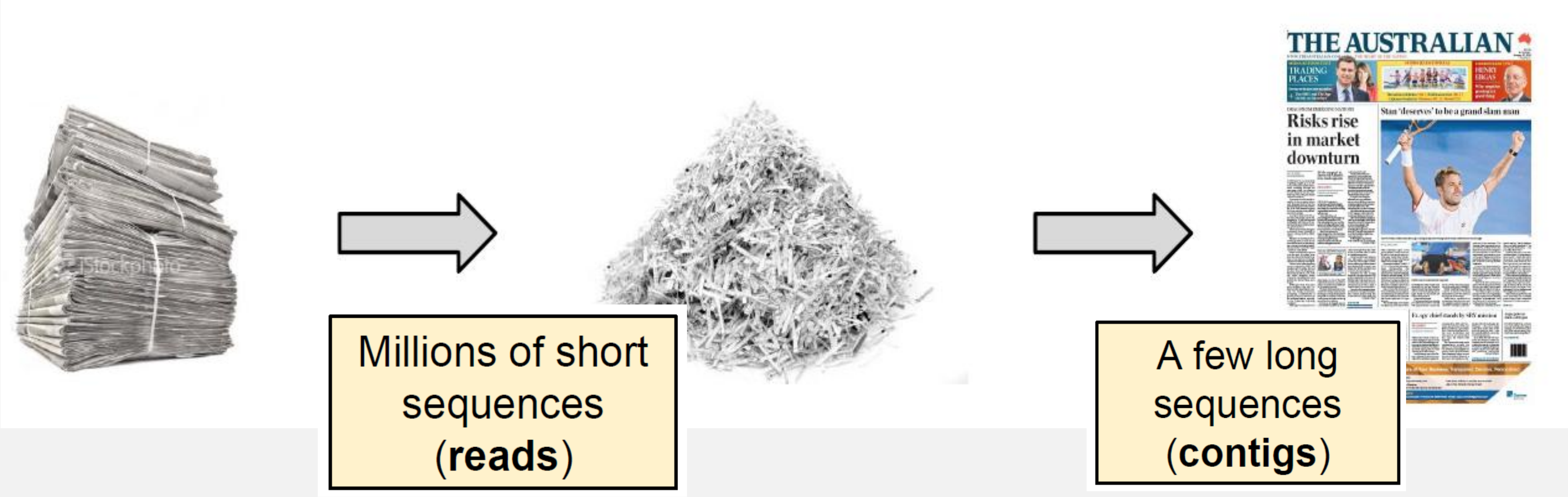


Image credit: Wikipedia

De novo genome assembly



Millions of short sequences (**reads**)

A few long sequences (**contigs**)

iStockphoto
THE AUSTRALIAN
Risks rise in market downturn
Stan 'deserves' to be a grand slam man

Image credit: Torsten Seemann

Reconstruct original genome from sequence reads only

“Shakespearomics”

Original text

“Friends, Romans, countrymen, lend me your ears”

- ***Reads***

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me
```

- ***Overlaps***

```
Friends, Rom
     ds, Romans, court
             ns, courtrymen, le
                     crymen, lend me
                             send me your ears;
```

- ***Majority consensus***

```
Friends, Romans, courtrymen, lend me your ears;
```

(1 contig)

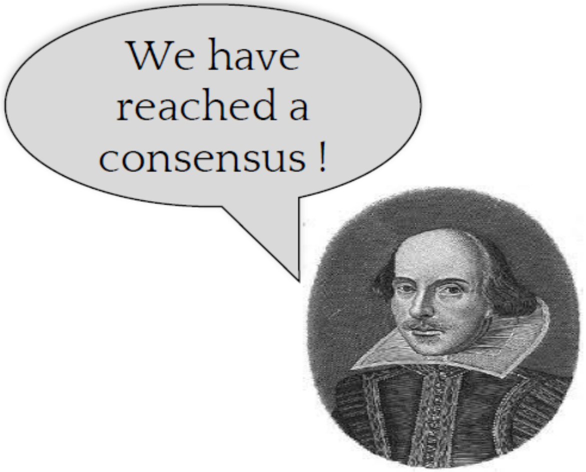
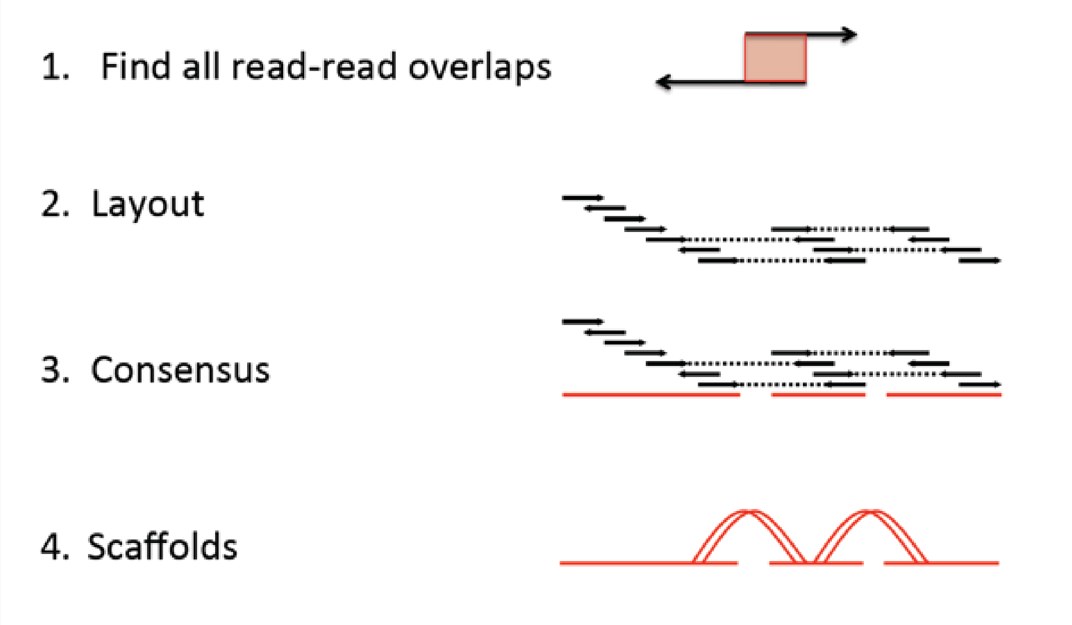


Image credit: Torsten Seemann

Overlap layout consensus

Overlap layout consensus is an assembly method that takes all reads and finds overlaps between them, then builds a consensus sequence from the aligned overlapping reads



1. Find all read-read overlaps
2. Layout
3. Consensus
4. Scaffolds

https://bioinformaticsworkbook.org/dataAnalysis/GenomeAssembly/Intro_GenomeAssembly.html#gsc.tab=0

De Bruijn graph (DBG) & k-mers

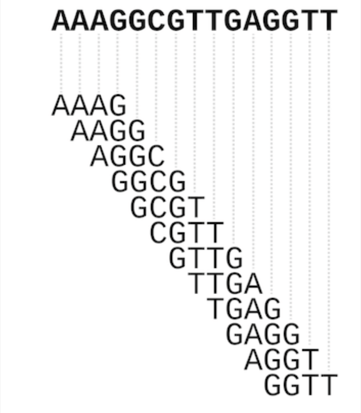
Chop the reads into k-mers

Construct DBG from k-mers

DBG represents a sequence in terms of its k-mer components

Find Eulerian/Hamiltonian path in graph

Derive the genome sequence from path



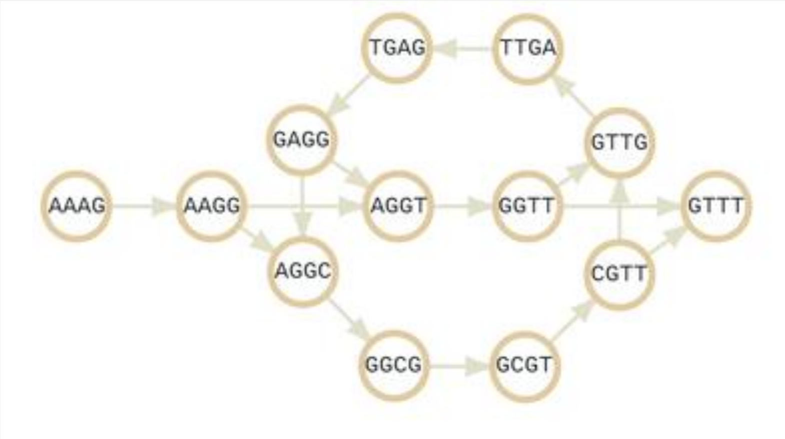


Image credit: Jang Il Sohn

Genome assembly is hard

A puzzle with millions to billions of pieces (N)

Read length is short relative to genome (10^2 vs 10^9 bp)

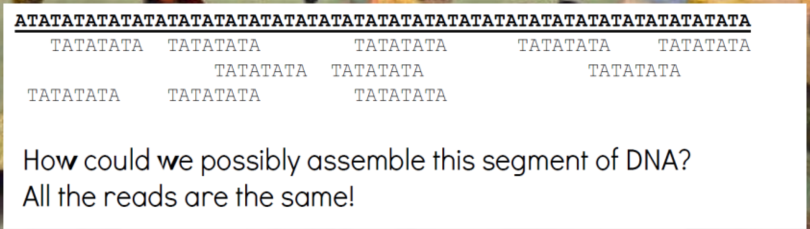
Lots of overlap to check

Compare each pair of pieces to find overlap (N^2)

Lots of sky; i.e. short repeats

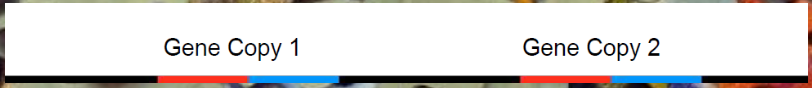
Lots of dirty pieces; i.e. sequencing errors

Multiple copies; i.e. long repeats

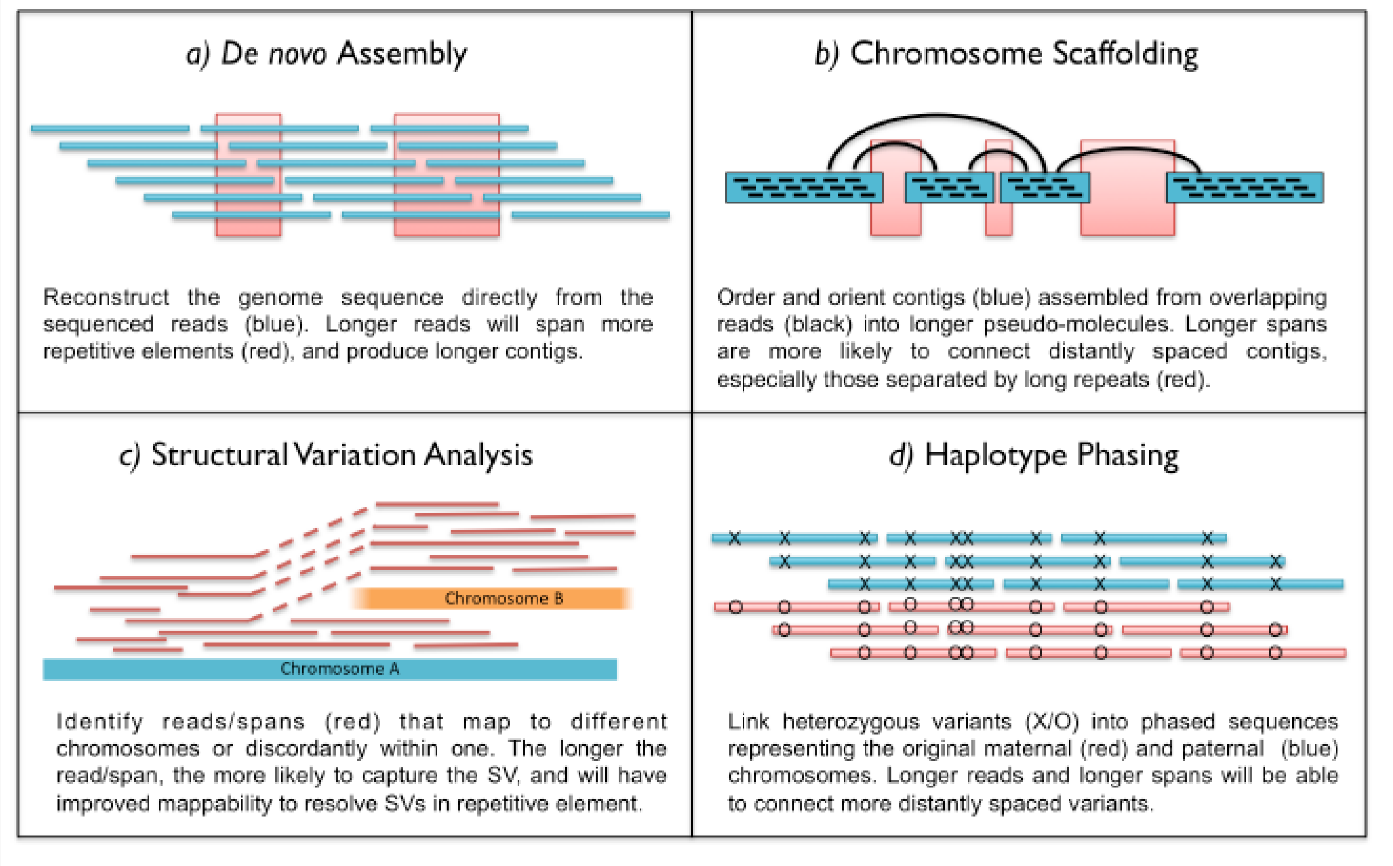


```
ATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATA
   TATATATA  TATATATA        TATATATA      TATATATA    TATATATA
                 TATATATA  TATATATA              TATATATA
 TATATATA    TATATATA        TATATATA
```

How could we possibly assemble this segment of DNA?
All the reads are the same!



Some of these difficulties are mitigated by 3rd-gen sequencing’s long read length



a) *De novo* Assembly

Reconstruct the genome sequence directly from the sequenced reads (blue). Longer reads will span more repetitive elements (red), and produce longer contigs.

b) Chromosome Scaffolding

Order and orient contigs (blue) assembled from overlapping reads (black) into longer pseudo-molecules. Longer spans are more likely to connect distantly spaced contigs, especially those separated by long repeats (red).

c) Structural Variation Analysis

Identify reads/spans (red) that map to different chromosomes or discordantly within one. The longer the read/span, the more likely to capture the SV, and will have improved mappability to resolve SVs in repetitive element.

d) Haplotype Phasing

Link heterozygous variants (X/O) into phased sequences representing the original maternal (red) and paternal (blue) chromosomes. Longer reads and longer spans will be able to connect more distantly spaced variants.

https://doi.org/10.1101/048603

To do the next exercise, you need to know a little about cell-free DNA and fragment length histogram

Normal cfDNA

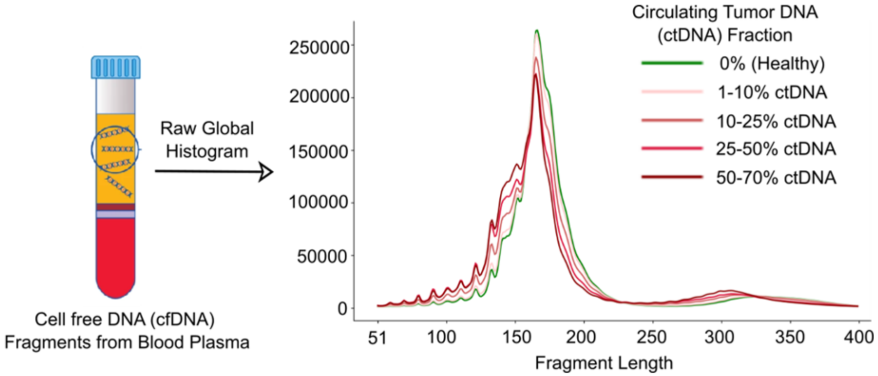
From apoptotic white blood cells

Usually 166 – 180 bp

Fetal cfDNA (cffDNA)

From apoptotic trophoblasts

Usually shorter than normal cfDNA

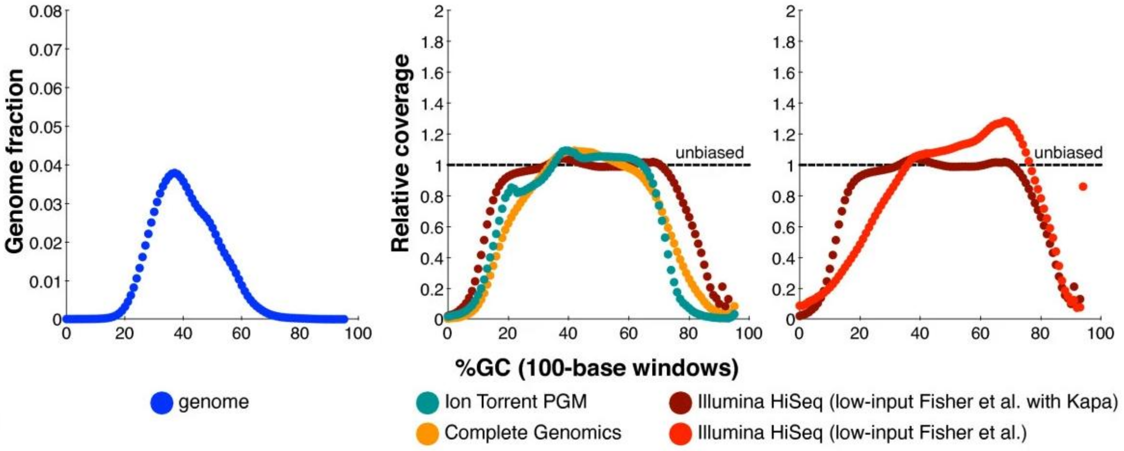


Circulating tumour DNA (ctDNA)

From tumour cells via necrosis, stress induced pathways, etc.

Usually shorter than normal cfDNA

Exercise



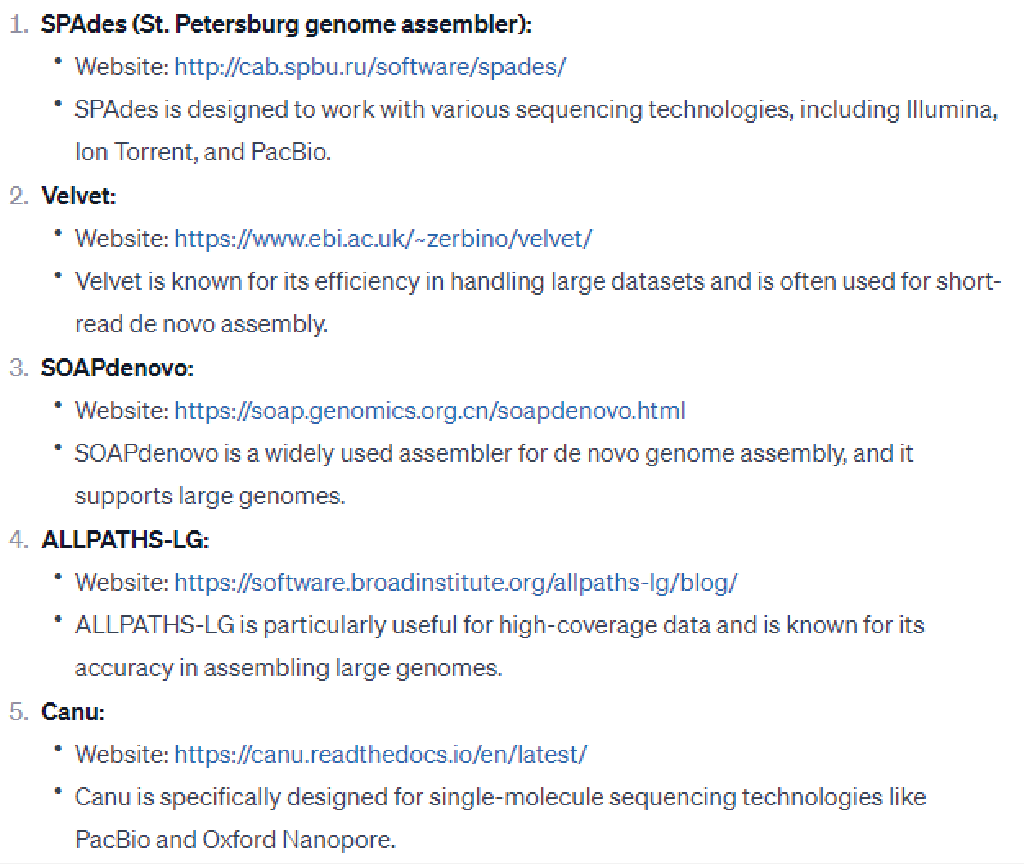
Does the GC content bias in DNA sequencing coverage have a big impact on genome assembly?

Does it have a big impact on tumour purity determination from cell-free DNA?

Does it have a big impact on non-invasive pre-natal testing using cell-free DNA?

Popular genome assembly tools

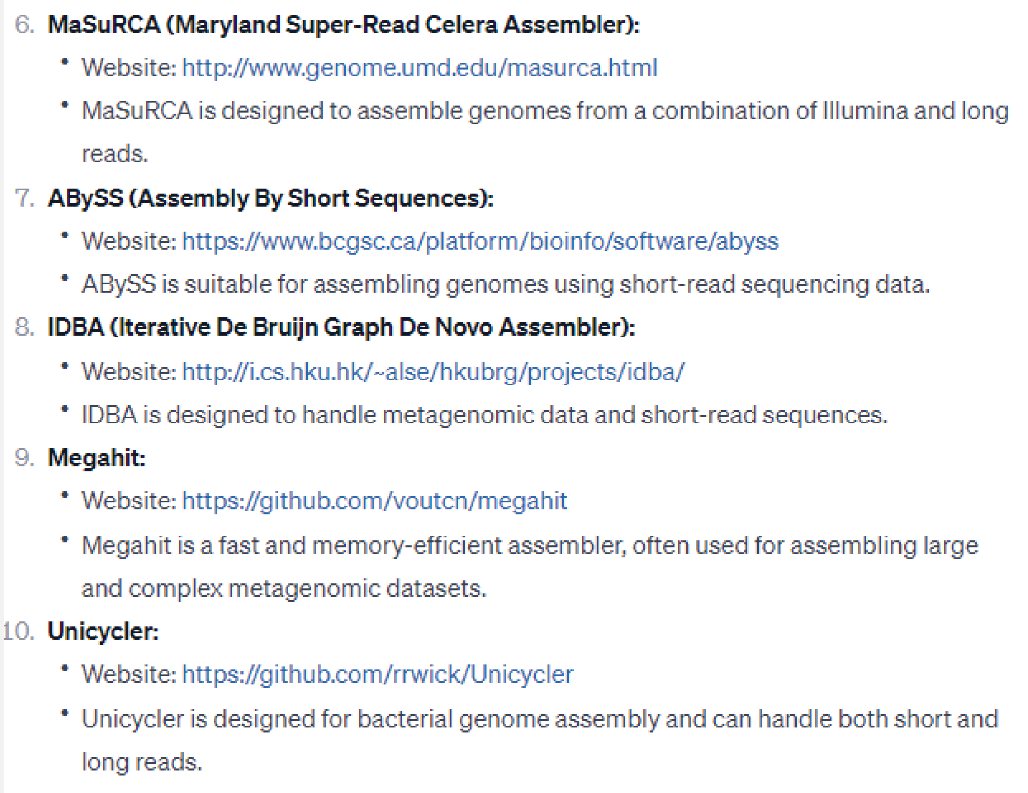
Check these out yourself



1. **SPAdes (St. Petersburg genome assembler):**
   - Website: http://cab.spbu.ru/software/spades/
   - SPAdes is designed to work with various sequencing technologies, including Illumina, Ion Torrent, and PacBio.
2. **Velvet:**
   - Website: https://www.ebi.ac.uk/~zerbino/velvet/
   - Velvet is known for its efficiency in handling large datasets and is often used for short-read de novo assembly.
3. **SOAPdenovo:**
   - Website: https://soap.genomics.org.cn/soapdenovo.html
   - SOAPdenovo is a widely used assembler for de novo genome assembly, and it supports large genomes.
4. **ALLPATHS-LG:**
   - Website: https://software.broadinstitute.org/allpaths-lg/blog/
   - ALLPATHS-LG is particularly useful for high-coverage data and is known for its accuracy in assembling large genomes.
5. **Canu:**
   - Website: https://canu.readthedocs.io/en/latest/
   - Canu is specifically designed for single-molecule sequencing technologies like PacBio and Oxford Nanopore.

Popular genome assembly tools

Check these out yourself



6. **MaSuRCA (Maryland Super-Read Celera Assembler):**
   - Website: http://www.genome.umd.edu/masurca.html
   - MaSuRCA is designed to assemble genomes from a combination of Illumina and long reads.
7. **ABySS (Assembly By Short Sequences):**
   - Website: https://www.bcgsc.ca/platform/bioinfo/software/abyss
   - ABySS is suitable for assembling genomes using short-read sequencing data.
8. **IDBA (Iterative De Bruijn Graph De Novo Assembler):**
   - Website: http://i.cs.hku.hk/~alse/hkubrg/projects/idba/
   - IDBA is designed to handle metagenomic data and short-read sequences.
9. **Megahit:**
   - Website: https://github.com/voutcn/megahit
   - Megahit is a fast and memory-efficient assembler, often used for assembling large and complex metagenomic datasets.
10. **Unicycler:**
    - Website: https://github.com/rrwick/Unicycler
    - Unicycler is designed for bacterial genome assembly and can handle both short and long reads.

Self-guided practice

Bilal Wajid & Erchin Serpedin, “Do it yourself guide to genome assembly”, *Briefings in Functional Genomics* 15(1):1-9, 2016. Supplementary section

Good to read

Sequence assembly. https://en.wikipedia.org/wiki/Sequence_assembly

B. Wajid & E. Serpedin, “Do it yourself guide to genome assembly”, Briefings in Functional Genomics, 15(1):1-9, 2016. https://doi.org/10.1093/bfgp/elu042