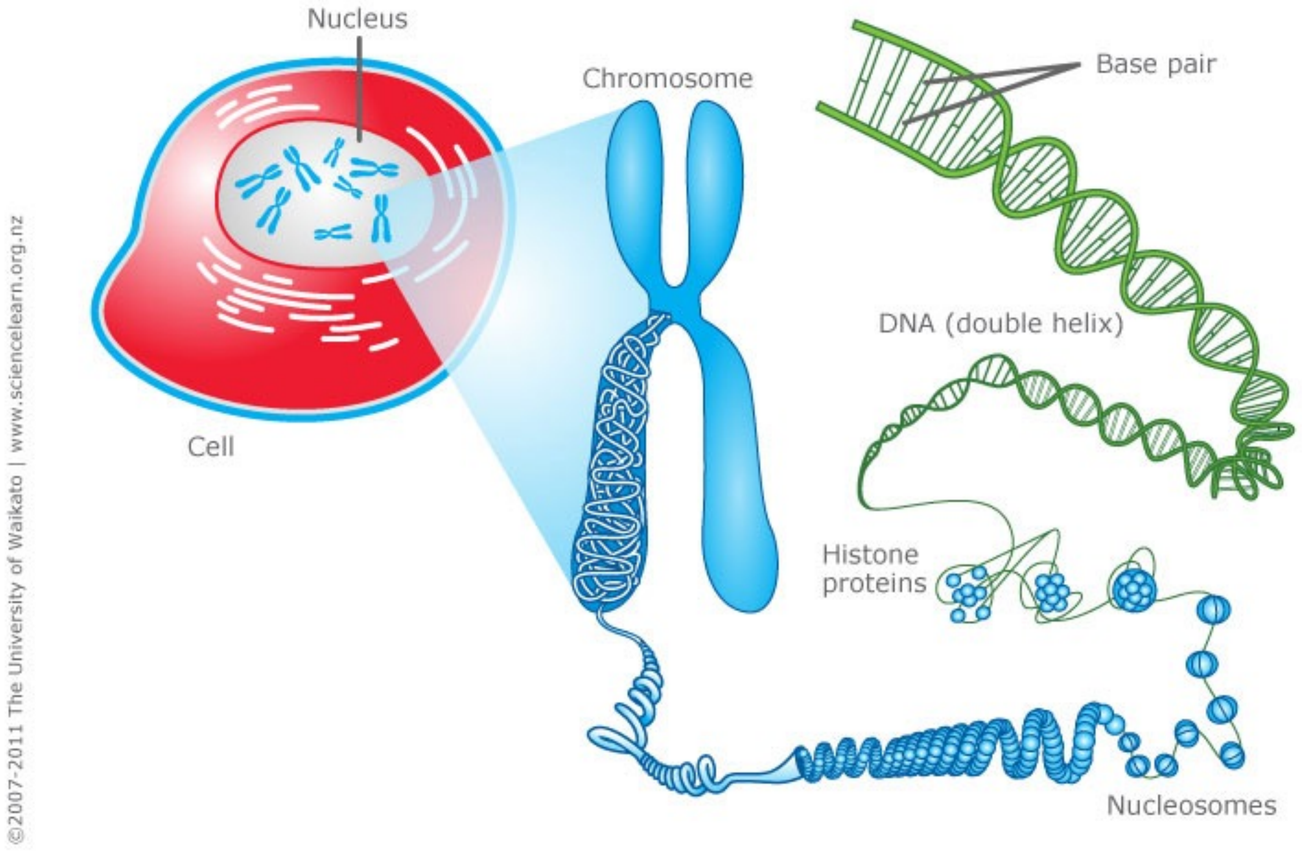
GENOME ASSEMBLY AND ANNOTATION

Having good quality fastq data is important!!

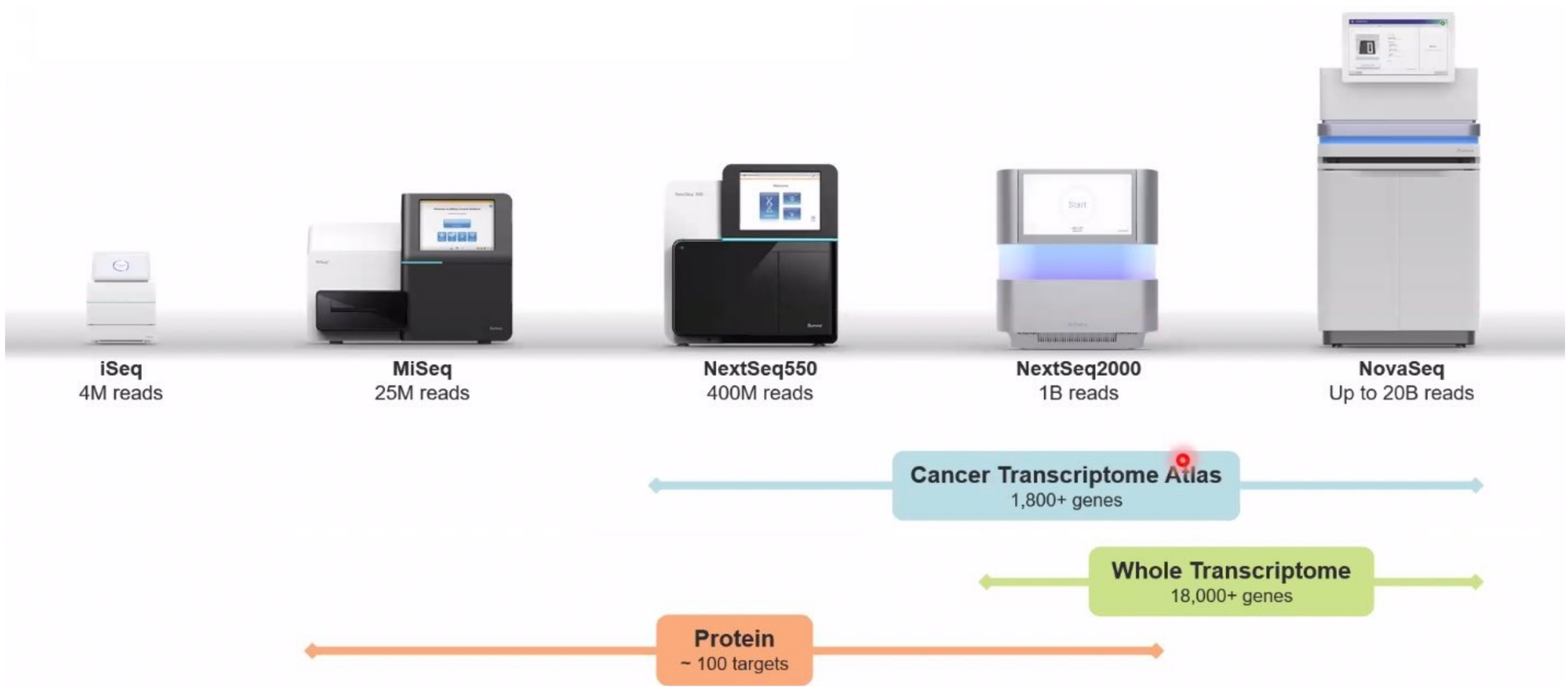
Genome sequencing is conceptually straight forward




```
CCACAAGTTCCTGACTGCCTGACTTCCCTTCACCGACTGGCACTTTCCACTCGGATCGCC
AGCGACCGTTACTAAAAAAACAAACATCGAATACTGTCTGCAAGACAGTGCAATAAAGCA
AATGAAAATAATTAGAATAAGAATAATGTTAATAATGATACCAAAAATTCTTCGGCTGGA
ACTGATGTGACTCTATGCATAATGTGAAATTTCCATGACGAACGAACACGCATCCTACAC
CAGATTTTGAGTAATGTTCCTCTATATATGCATCTAATTCCTAAGATAAATGGGTGTGAG
CAGCAACTAGAATTGGAAAGAACCACTGAACAGCTTGGTACCTTTTGCAGAGCTACACGC
CACGATTCTAAAGCGCTCGATTCTGCTGGATACGGCTGGATGTCACTGGCCTCCTTCCGA
AGATGATCATACTTGAAGGATTTCTTCGACGCATCATCCACGTAGCGACCTCCTCCAATA
TCGTTGAATGGCGTCACAAGTGTCTAAGCGACATTTTTTGAAAGTGTAAGCAATACTTGA
TCAACTTCTTTTCCGAATCCTCGATGAATAATCATTAGAACAATACGTTTCCATTTTTTA
CTGCACCCTTTTCCATACAAATAGTTACAAGTTAGATTGGACAATGACAACAAATGAACA
CCGAGTTCTATGGTGAGAAAACATGCACTAGGGAATCGACCGCCTTGTGCAGCAGCATTC
ATGGTGAAAAGACAAGATCTGTCATAGATGAGTTTAAGTTTGAACAGTTCCCCTTCAAAT
TCCACACTATCACAGACCATTCCCCGAAGAATGTTCGACCTCTAGACCTGACCTCTACGA
```

https://www.sciencelearn.org.nz/images/198-cell-chromosomes-and-dna

Sequencing genomes is easy…



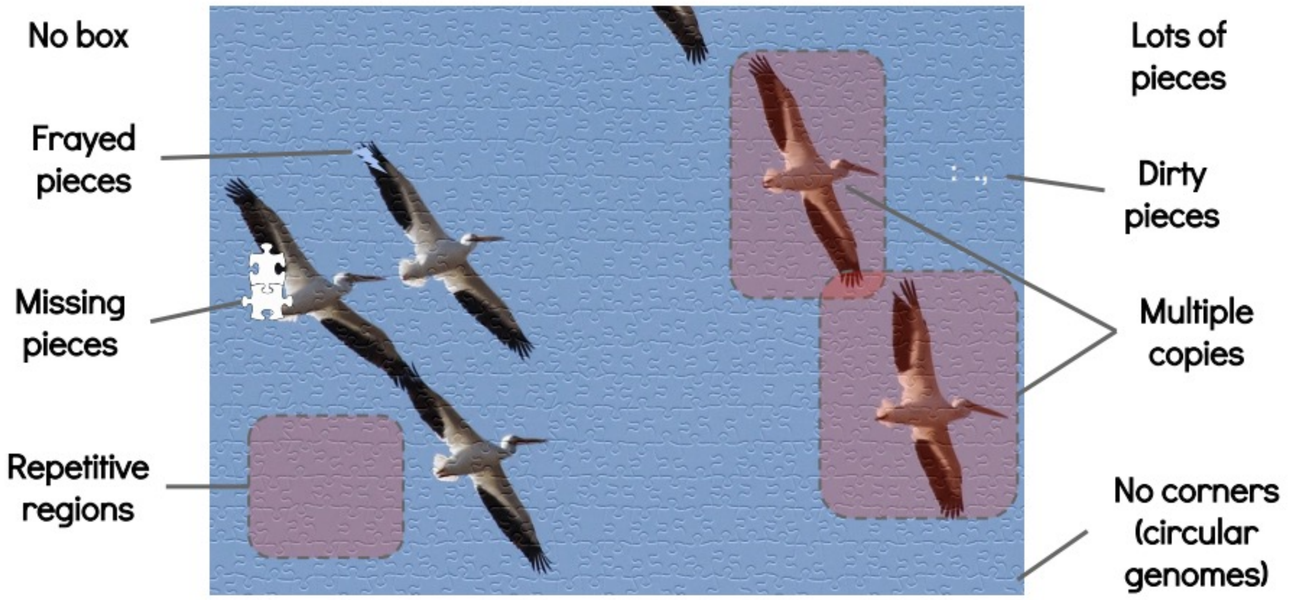
Sequencing is easy, constructing good genomes is not

- **Genome: *biologically***
  - “the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism”
  - “the complete set of genes or genetic material present in a cell or organism”

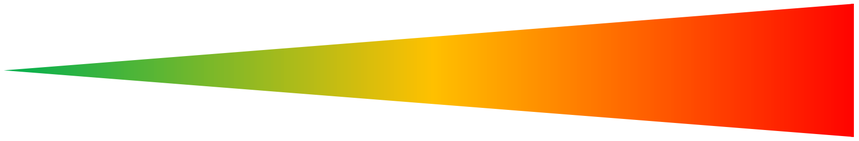
Adapted from Torsten Seemann presentation: “De novo genome assembly”
https://www.slideshare.net/torstenseemann/de-novo-genome-assembly-tseemann-imb-winter-school-2016-brisbane-au-4-july-2016

Sequencing genomes is easy, constructing good genomes is not

- **Genome: *bioinformatically***
  - Best guess, but often:
    - highly fragmented
    - misassembled to some degree
    - contaminated
    - duplicated or missing



Draft genomes
"manageable"(?)



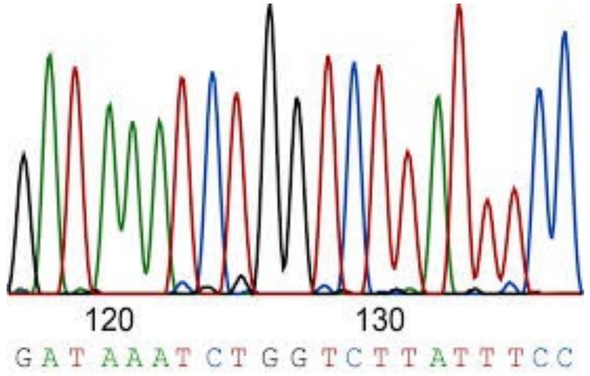
Chromosome-scale genomes
HARD

Time, money, expertise

Adapted from Torsten Seemann presentation: "De novo genome assembly"
https://www.slideshare.net/torstenseemann/de-novo-genome-assembly-tseemann-imb-winter-school-2016-brisbane-au-4-july-2016

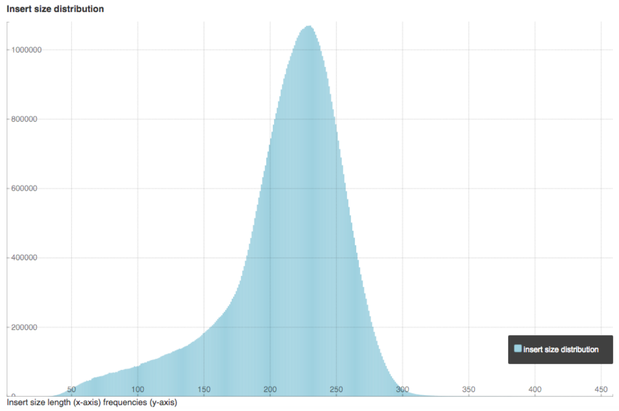
New technologies are making genome assembly easier

Sanger Sequecing: ABi



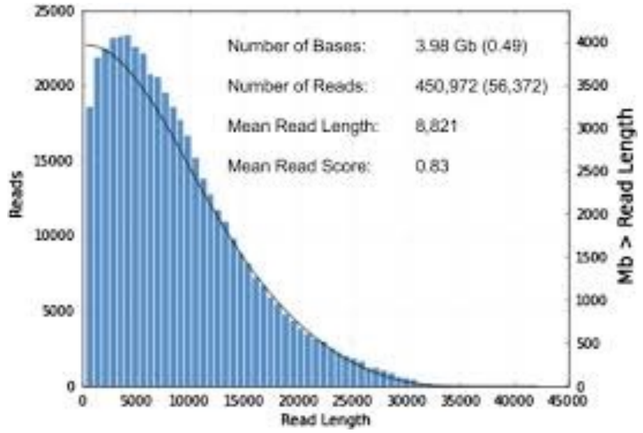
Read length:500-1000 bp

High Throughput Sequencing: Illumina



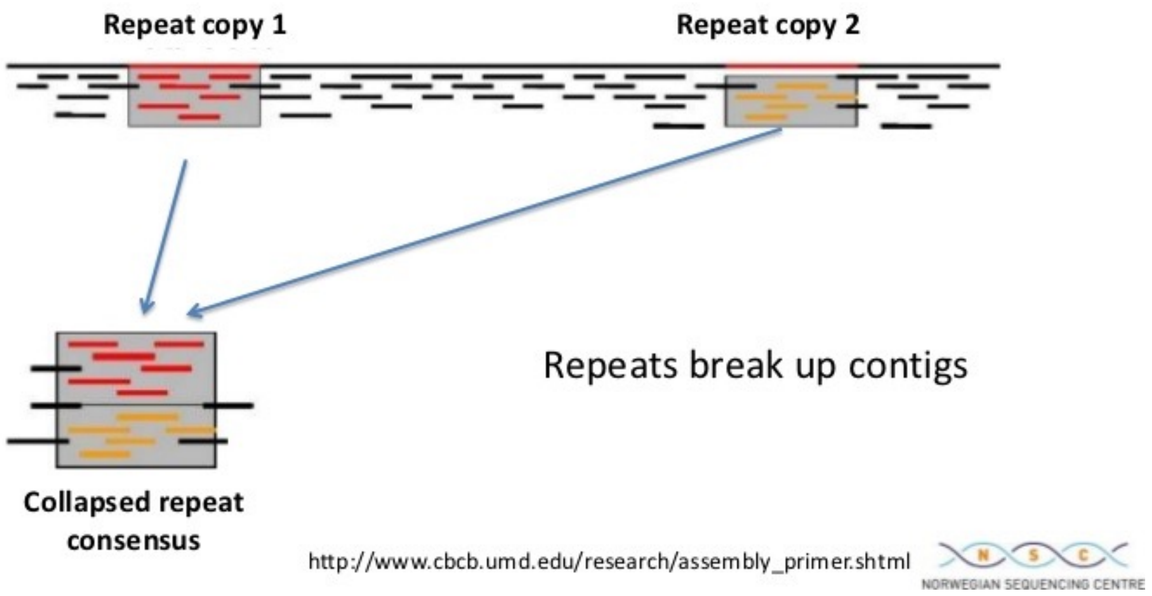
Read lengths: 100-300 bp
Insert lengths: ave 300-500 bp

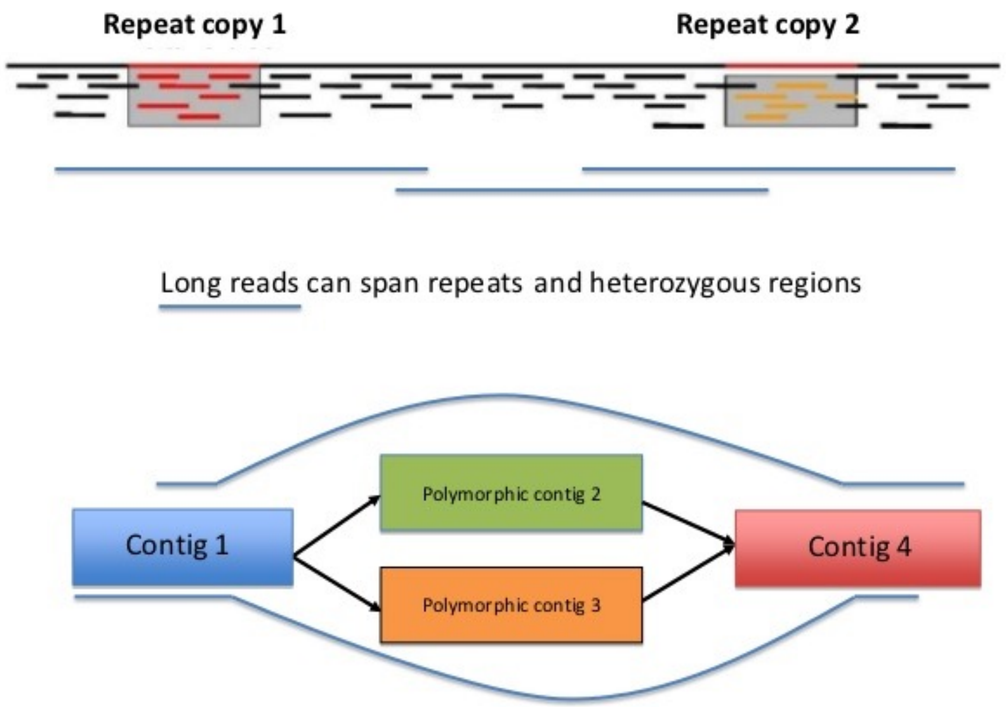
Long read sequencing: Pacbio & Nanopore



Read lengths: 5-10 kb
- Pacbio: up to 60 kb
- Nanopore: up to 1Mb

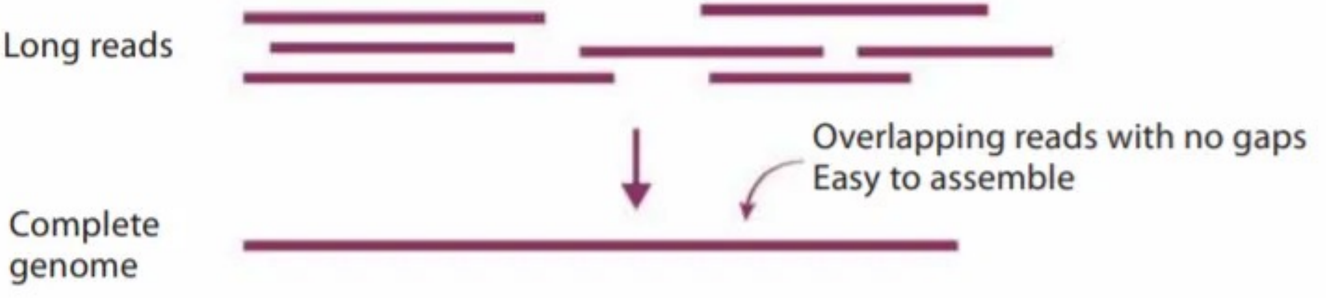
Repeats / polymorphic loci can break genomes





https://www.slideshare.net/flxlex/combining-pacbio-with-short-read-technology-for-improved-de-novo-genome-assembly

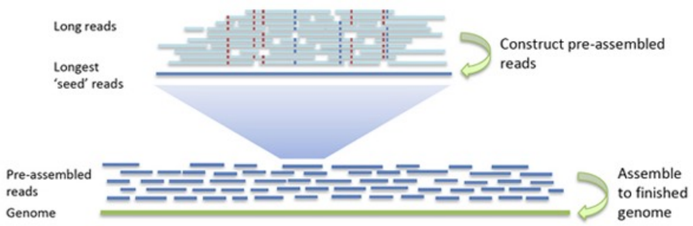
Missing sequence leads to missed genes and limits biological interpretation



A comprehensive structural, functional, and organisional picture of the genome

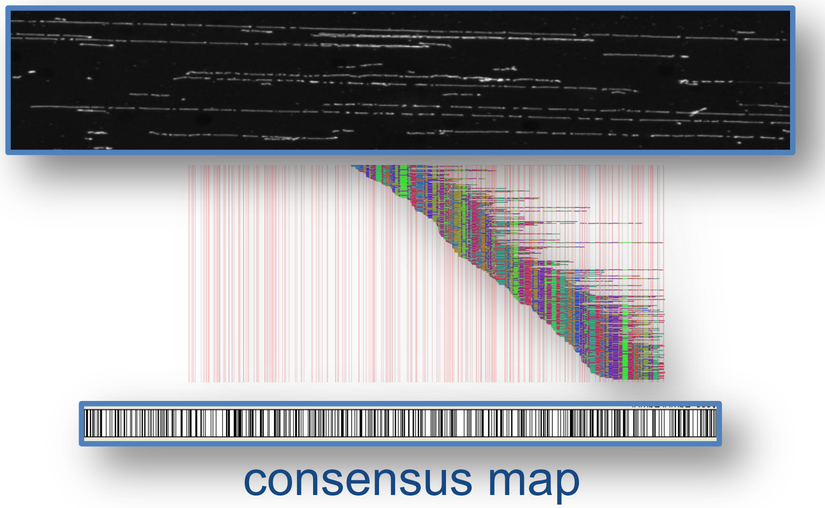
Long read / range sequencing is key to good genomes

Pacific Biosciences (PacBio)



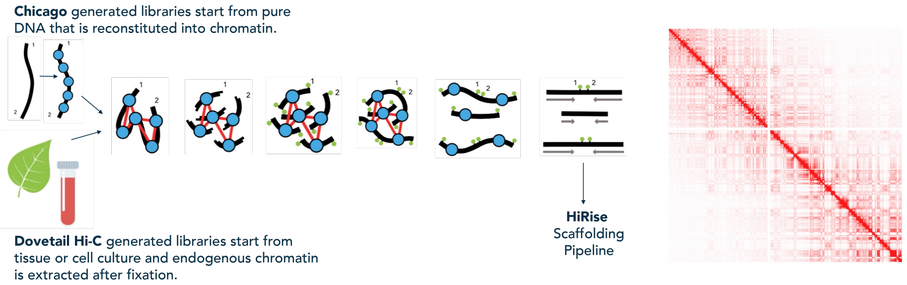
Oxford Nanopore

Optical Mapping (OpGen, Bionano genomics)



Linked reads (10X Genomics)

Chromosome confirmation capture, ie Hi-C (Dovetail Genomics)

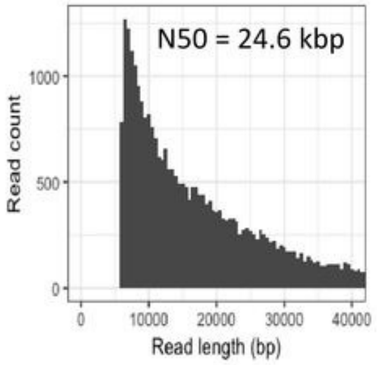


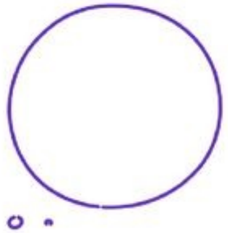
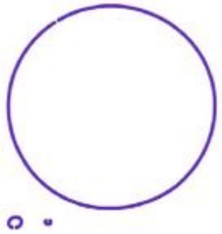
Klebsiella pneumoniae INF116

High Nanopore read depth

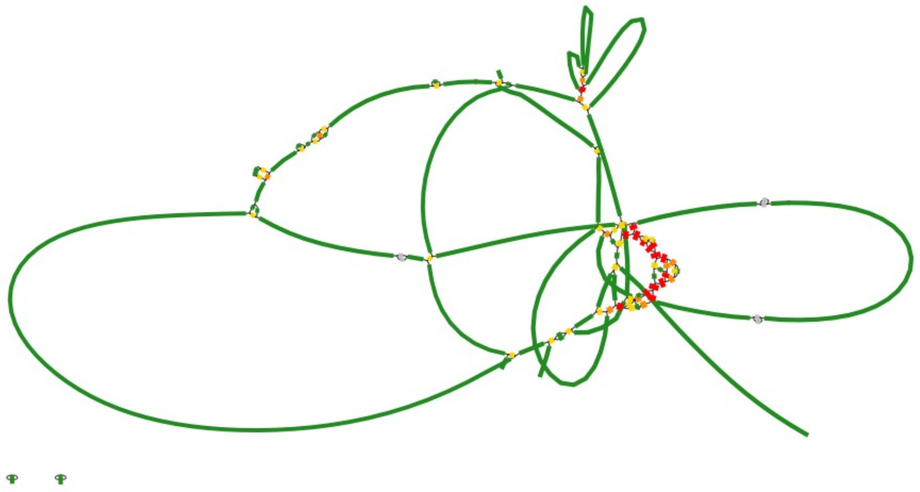
Nanopore reads:

- R9.4, whole flow cell
- Subsampled for length, quality
- 564 Mbp (>100x depth)
- Albacore 1.0.2
- Read length distribution:

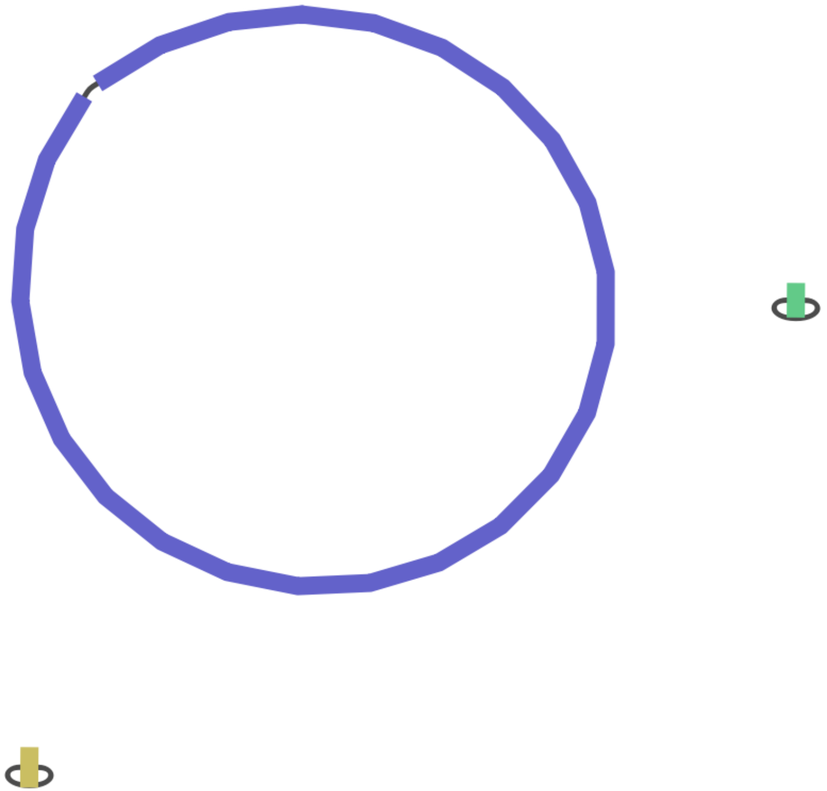


Illumina-only	Nanopore-only	+ Nanopolish	Hybrid
			
450 kbp	n/a	n/a	n/a
Accuracy:	99.15%	99.54%	100%
Bases per error:	118	218	∞

Klebsiella pneumoniae assemblies



Illumina – Spades



Illumina + Nanopore - Unicycler

Today's Agenda

- Use git command to download assemblies
- Use quast to assess various quality metrics
- Annotate genomes using prokka
- View genomes in Artemis

Questions?