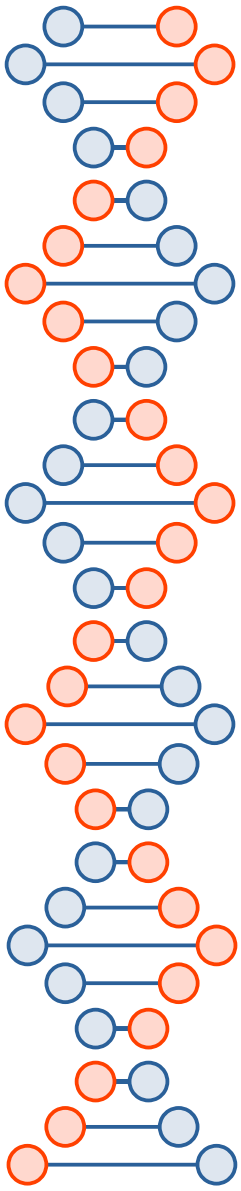
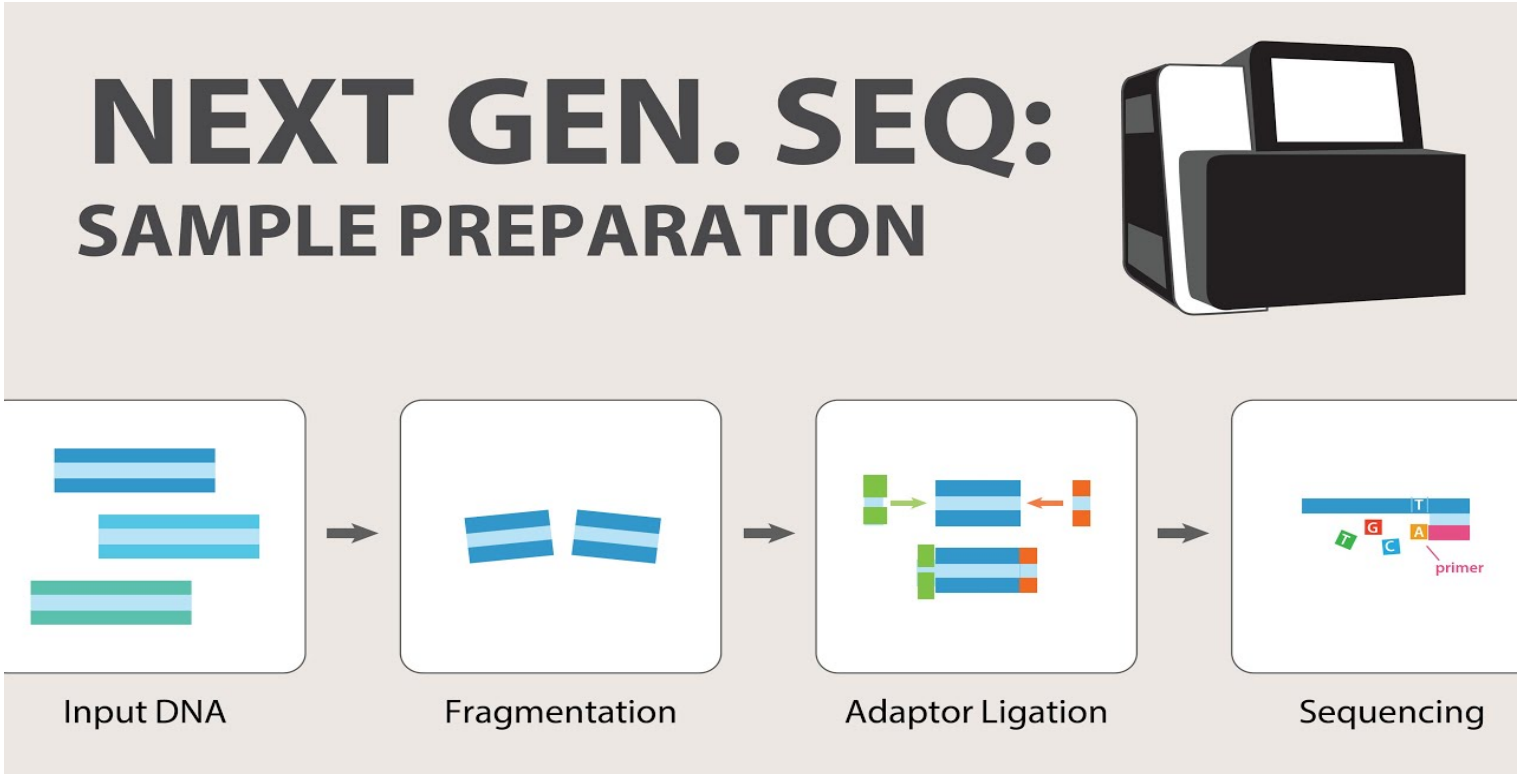


The present and future of de novo whole-genome assembly

Jang-il Sohn, Jin-Wu Nam

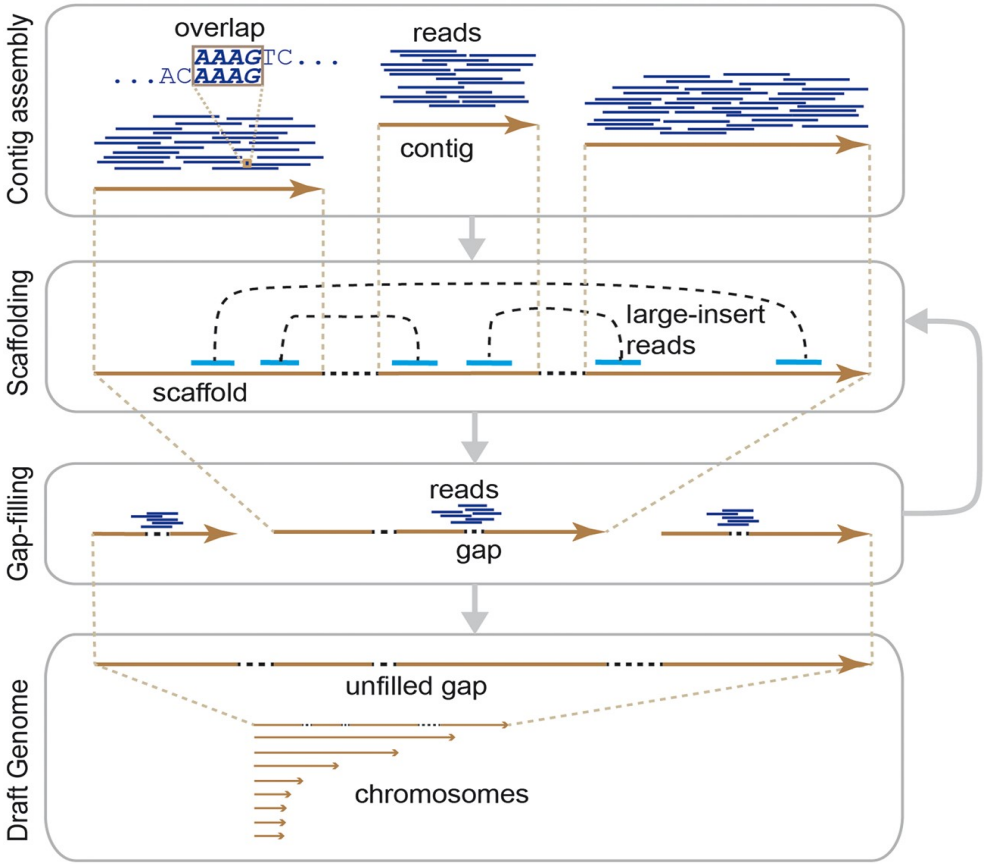
Presented by:
Yousef Alayoubi
Anastasiya Stepanenko

NGS made our lives easier



https://youtu.be/-kTcFZxP6kM
2) Next Generation Sequencing (NGS) - Sample Preparation

Assembly, de novo assembly

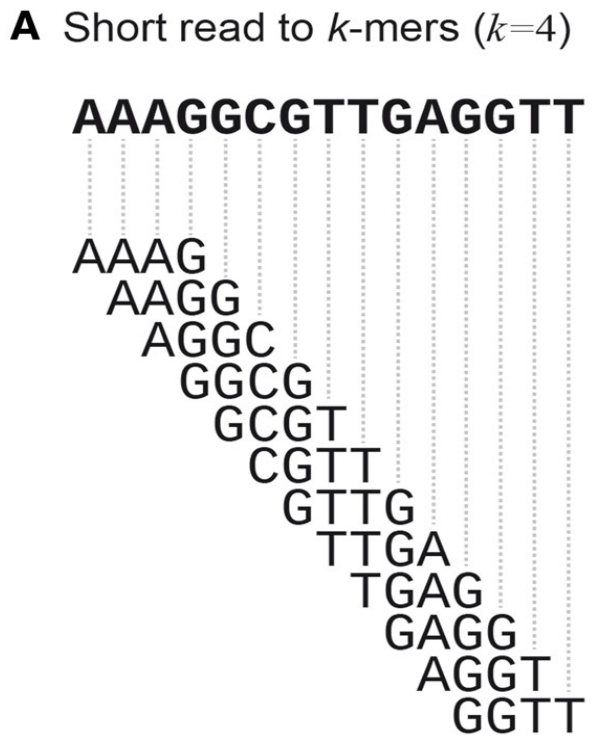


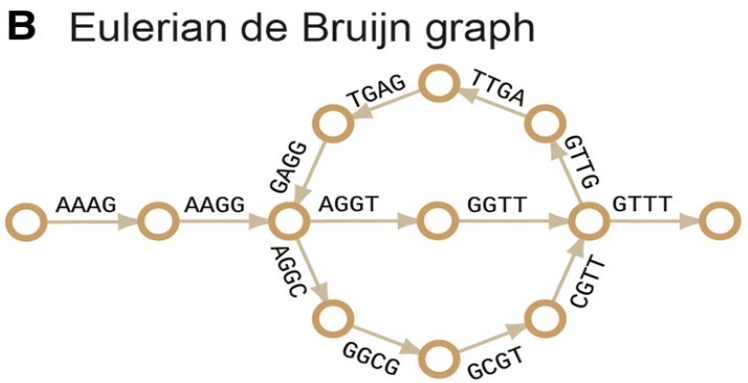
Short-read assemblers

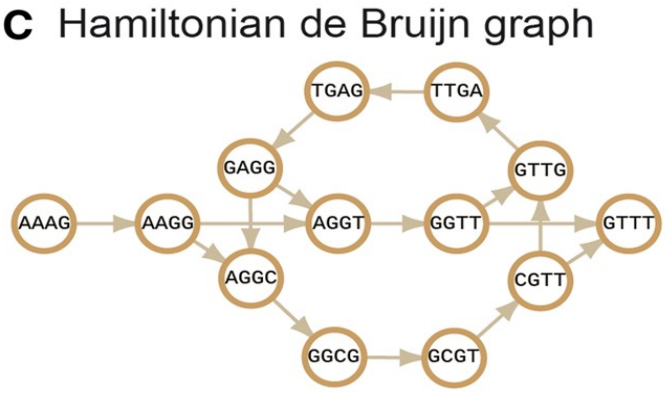
Assembler	Speed[a]	Memory efficiency[a]	N50 length[b]	Input data type	Assembly steps
Celera	+	+	+++	S,P,Li,L	C,S,G
ALLPATHS-LG	+	+	+++	P,Li (L[c])	E,C,S,G
ABySS	++	+++	++	S,P,Li	E,C,S
Velvet	++	++	+	S,P,Li	C,S
SPAdes	++	+++	++	P,Li	E,C,S
SOAPdenovo	+++	++	++	S,P,Li	C,S,G
SparseAssembler	++	+++	++	S,P,Li	C,S
SGA	+++	++	+	S,P,Li	E,C,S
MaSuRCA	+	+	+++	S,P,Li,L	C,S,G
Meraculous	++	++	++	P,Li	C,S,G
JR-Assembler	+	+	+++	S,P,Li	E,C,S,G

Note: +++: high; ++: medium; +:low.
In the 'Data Type' column, the symbols S, P, M and L refer to Single-end reads, Paired-end reads, Large-insert reads and Long reads, respectively.
In the 'Assembly steps' column, the symbols E, C, S and G refer to Error-correction, Contig assembly, Scaffolding and Gap-filling steps, respectively.

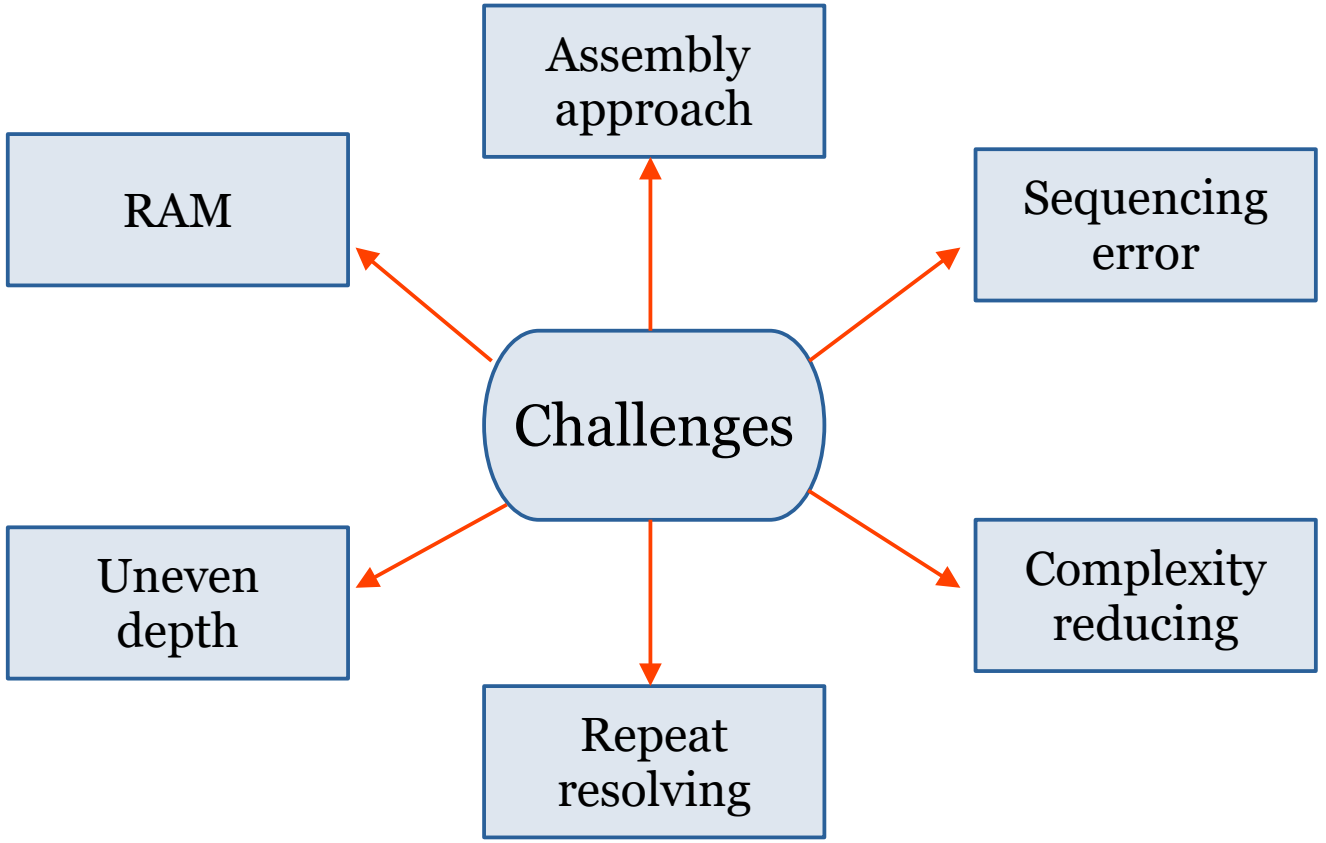
De Bruijn graph



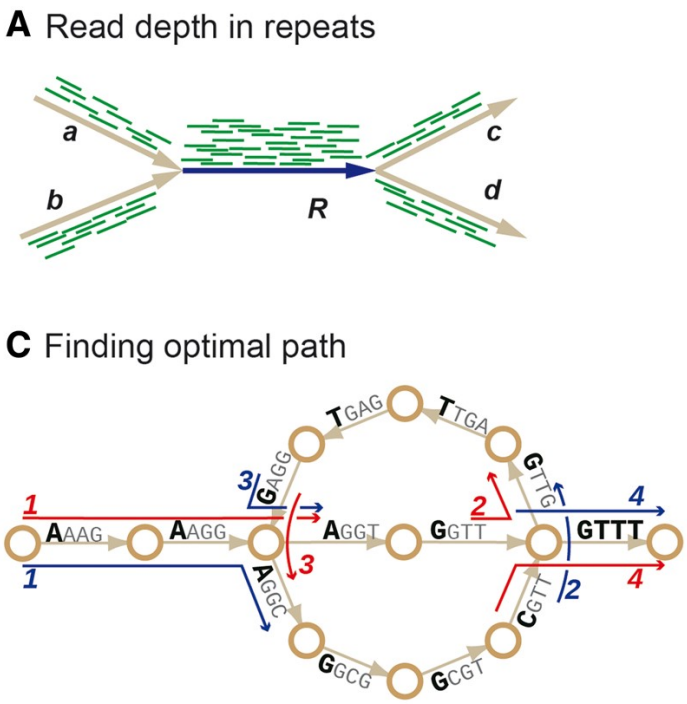


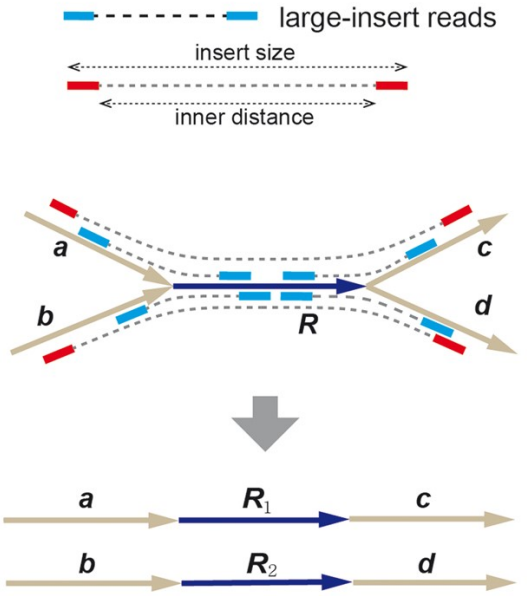


Challenges

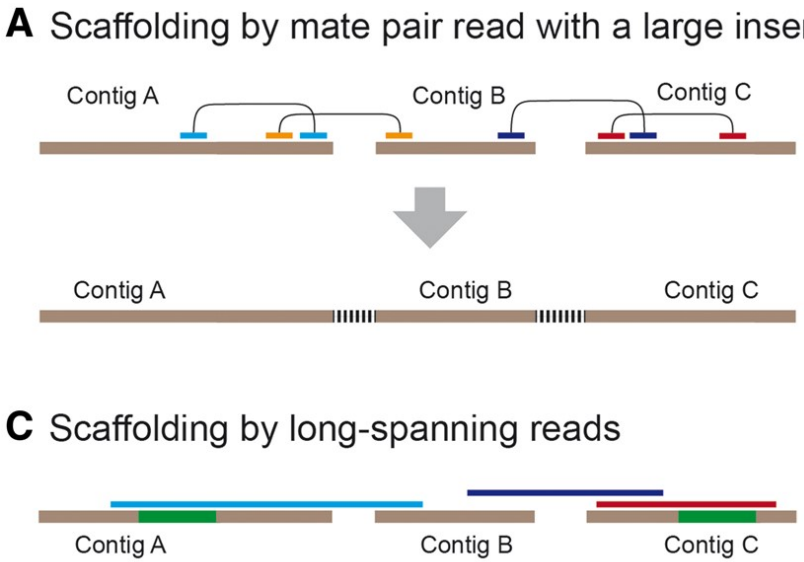

Assembly approach
RAM
Sequencing error
Challenges
Uneven depth
Complexity reducing
Repeat resolving

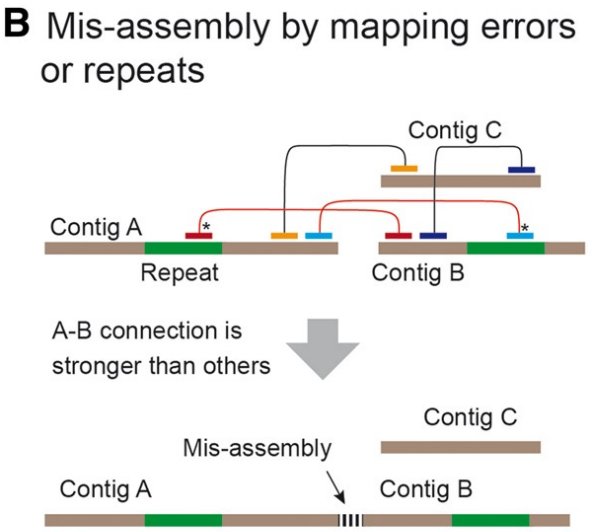
Repetitive Regions





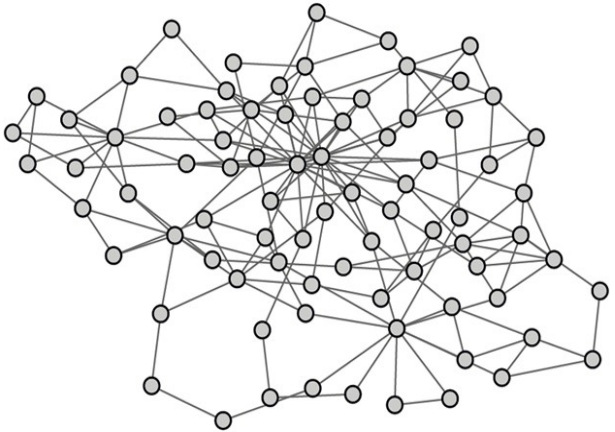
Scaffolding



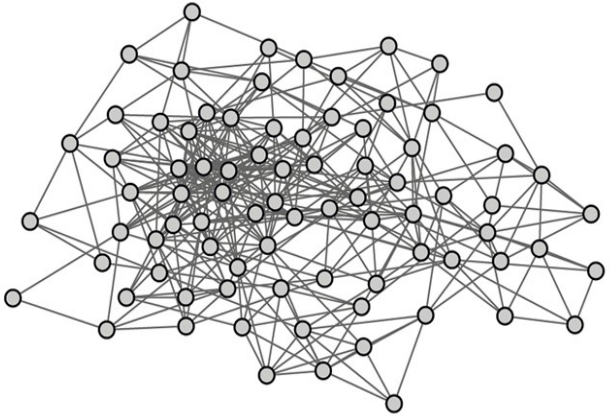


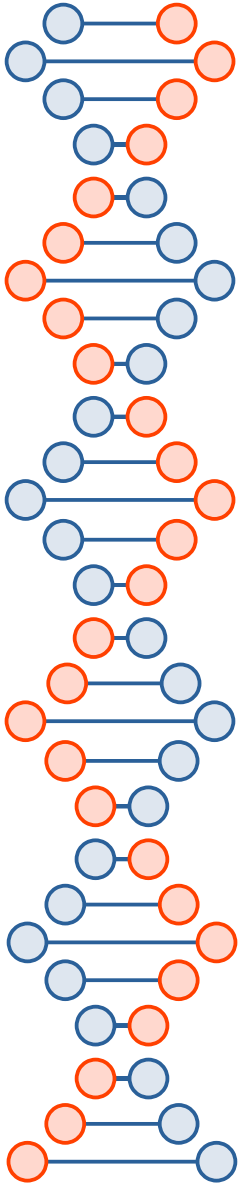
Scaffolding

D Assembly gaph by long-spanning reads



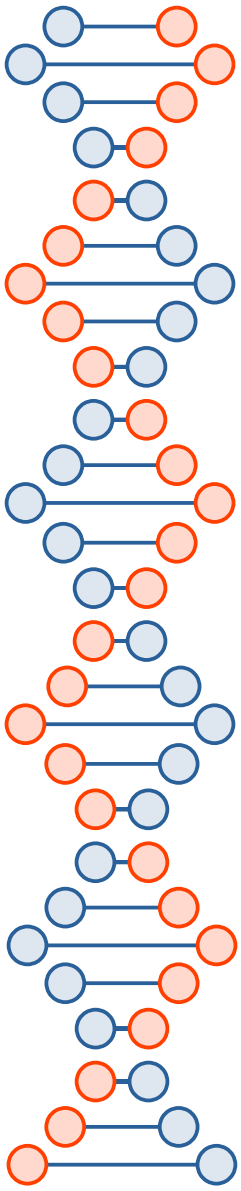
E Assembly graph by mate-pair reads





Long-read assembly

- Short-read assemblers struggle with repetitive regions and scaffolding.
- Third generation sequencing (PacBio and Nanopore) can generate reads ~ 30 kb long on average.
- Challenges:
  - High sequencing error
  - Low throughput
  - Expensive
- Hybrid methods are more cost efficient



Thank you

Any questions?