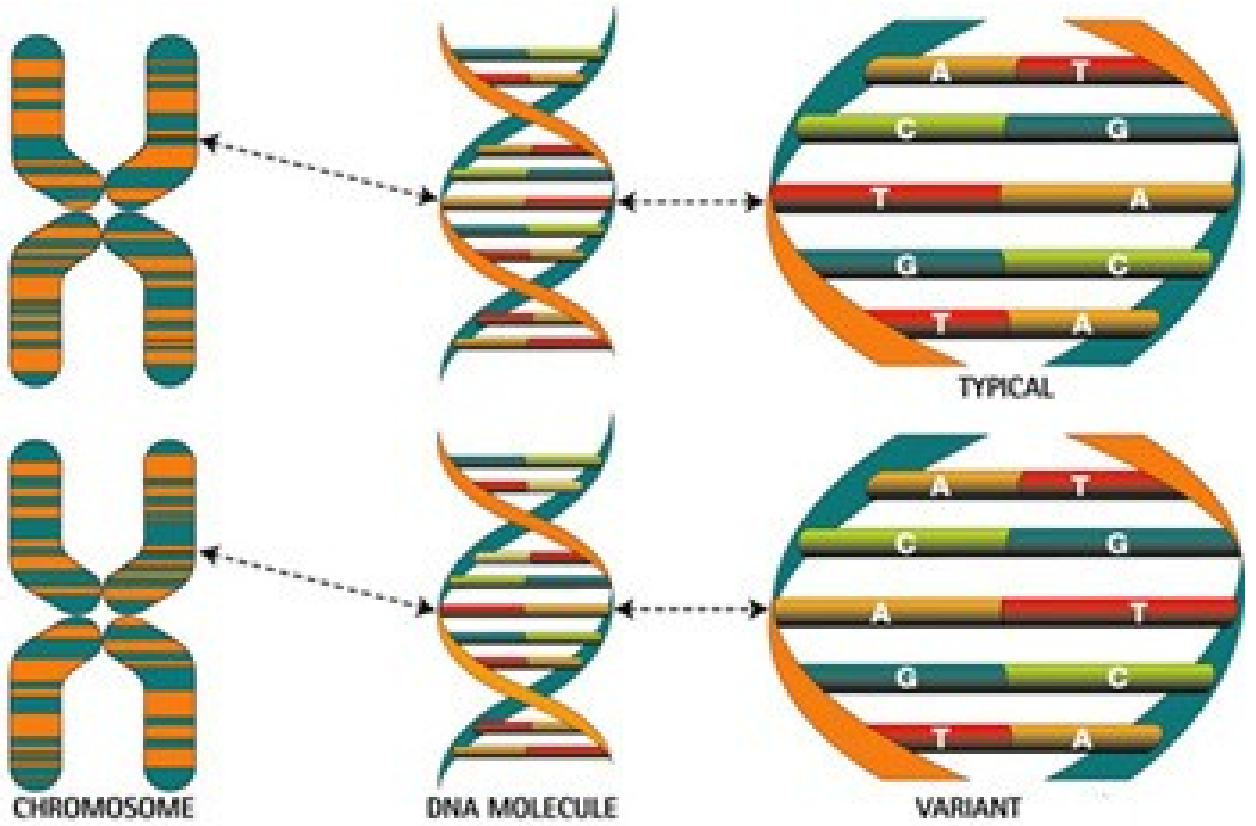
Identifying Differences Between Sequencing Data Sets

Vladimír Macko
Mgr. Tomáš Vinař, PhD

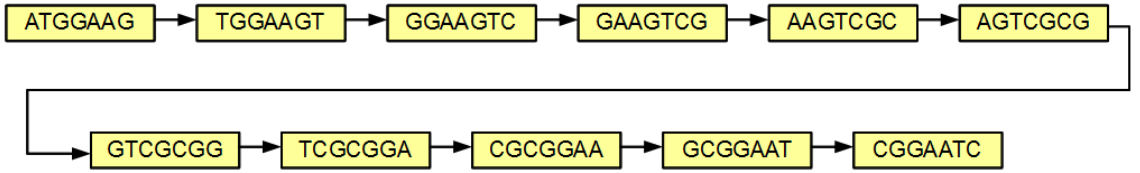

A
T
C
G
T
A
G
C
T
A
TYPICAL
A
T
C
G
A
T
G
C
T
A
CHROMOSOME
DNA MOLECULE
VARIANT

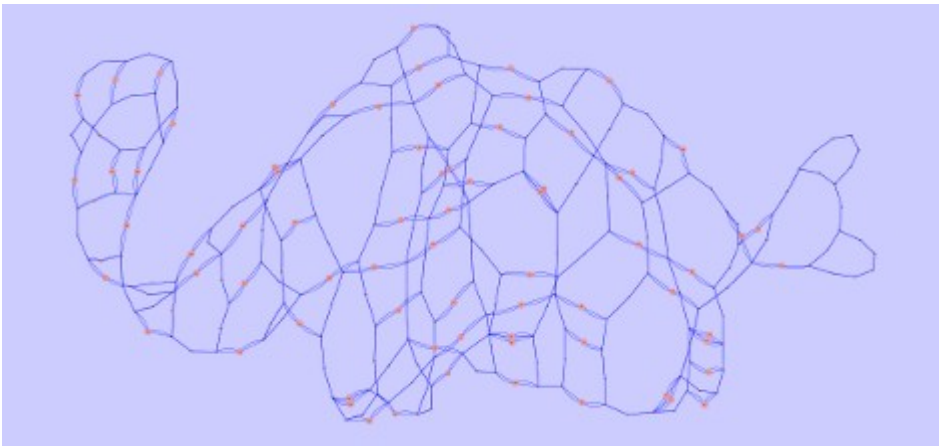
sequence **ATGGAAGTCGCGGAATC**

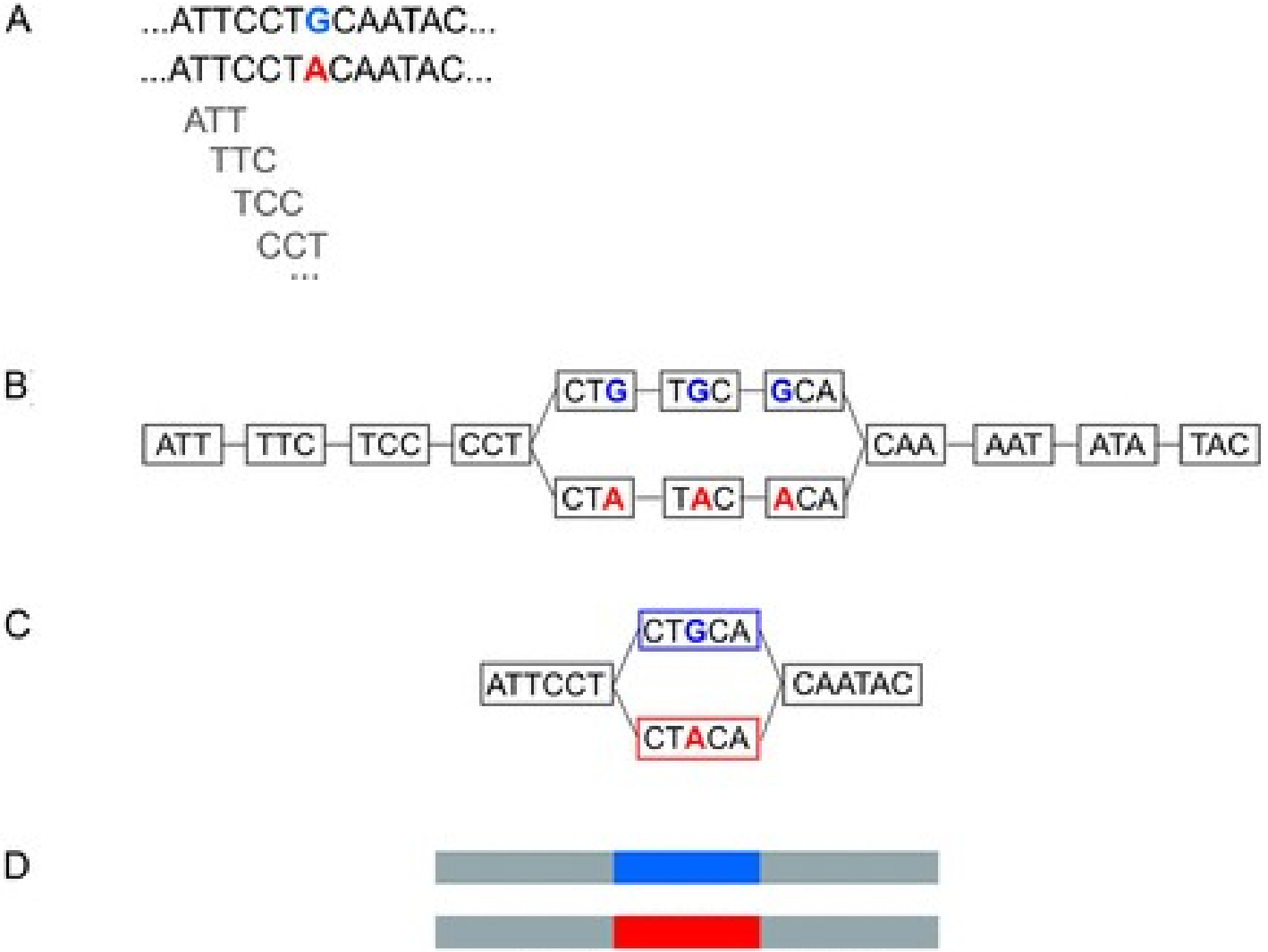
7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph






A
...ATTCCTGCAATAC...
...ATTCCTACAATAC...
ATT
TTC
TCC
CCT
...
B
ATT
TTC
TCC
CCT
CTG
TGC
GCA
CTA
TAC
ACA
CAA
AAT
ATA
TAC
C
ATTCCT
CTGCA
CTACA
CAATAC
D

Overview

- Introduction
- 1 Problem statement and related work
  - Biological motivation
  - Genome sequencing
  - Genome alignment and assembly
  - Problem statement
  - Related work
    - Bidirected de Bruijn graph
    - Bubble calling
    - Variant Calling
    - Probabilistic model for sequence assembly
    - Indexing techniques
    - A*
- 2 Proposed probabilistic approach
- 3 Catchy name for our tool that will be implemented
  - Implementation problems
  - Data
    - Variant simulation
    - Read simulation
    - Paired read simulation
    - Real data
  - Metrics
  - Benchmarks
    - Genome alignments
    - Without paired reads
    - With paired reads
  - Results
  - Future improvements
- 4 Discussion

Sequencing

- Overview of technologies

Technology	Read length (bp)	Paired reads	Error rate (%)	Cost per million bases ($)
454 GS FLX Titanium XLR70	700	Yes	0.1	10
Illumina HiSeq 2000	150	Yes	0.8	502
Illumina GAIIx	150	Yes	0.76	148
Illumina MiSeq	1500	Yes	12.86	2000
Ion Torrent PGM	200	Yes	1.71	1000
PacBio RS	1500	No	12.86	2000
Sanger 3730xl	400-900	No	0.001	2400
SOLiDv4	50	Yes	0.04	0.13
Oxford Nanopore MK 1 MinION	200000	No	12	750

Genome assembly

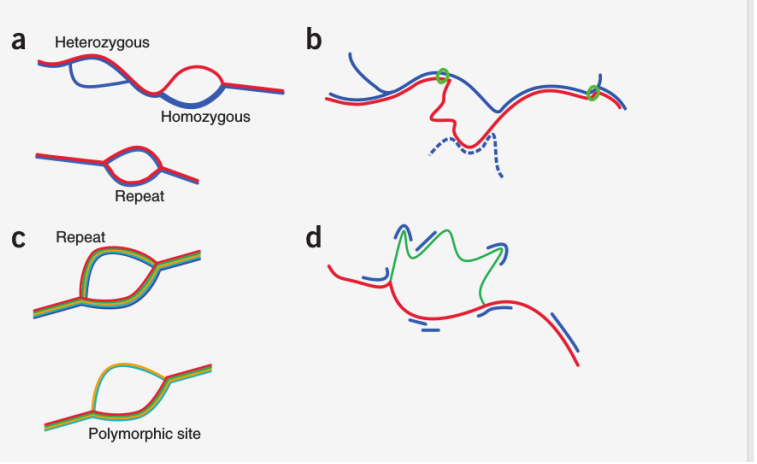
- Based on heuristics
- Paired reads are usually used only during post-processing

Variant Calling

- Heng Li,
  - Not De Novo

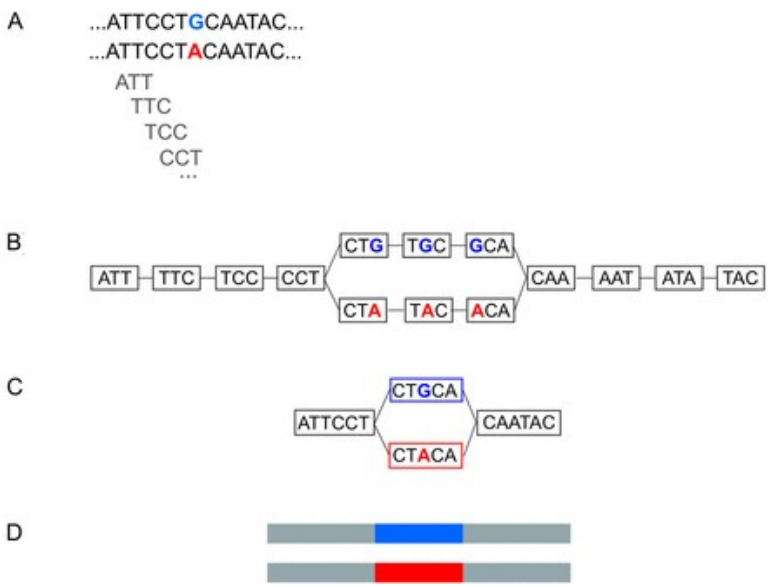
$$Q_s = \min\{q_2 - q_1 - 4.343\log(n_2), 4 + (3-k^*)(\hat{q}-14) - 4.343\log(P_1(3-k^*, 28)\}$$

- Zamin Iqbal, Mario Caccamo
  - No use of paired reads
  - Basically bubble enumeration



Bubble calling

- Enumeration is NP-hard
  - Chung-Lun Li, S.Thomas McCormick
- Polynomial delayed algorithm
  - Sacomoto



Indexing techniques

- K-ANNS with hierarchical navigable small world
- Logarithmic complexity