Debruijn Graph and its Application in **Genome Assembly** and **Variant Calling**

Bahar Alipanahi

UF UNIVERSITY of FLORIDA

Genome Assembly

Genome Assembly

Sequencing

Reads

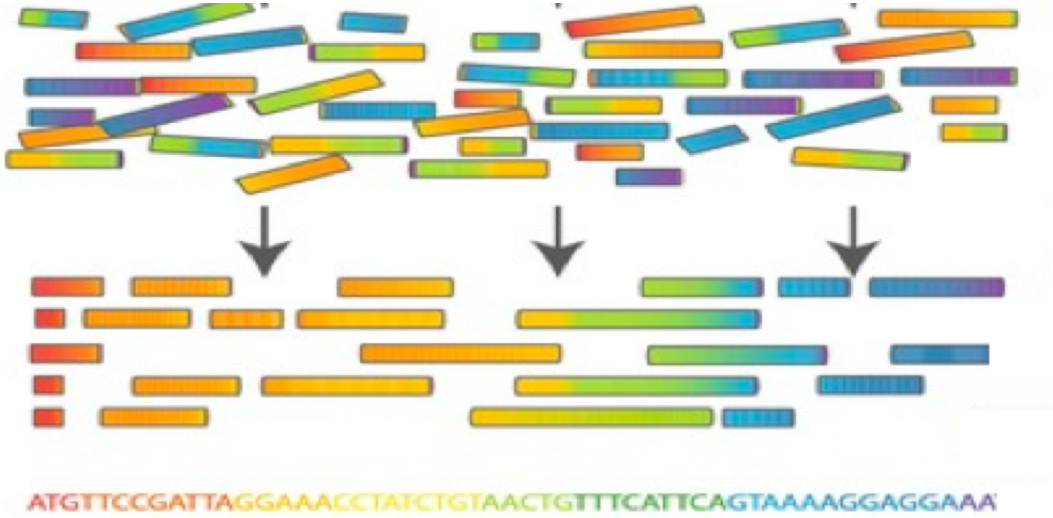
Assembly

ACGTAGAATACGTAGAA
ACGTAGAATCGACCATG
GGGACGTAGAATACGAC

ACGTAGAATACGTAGAAACAGATTAGAGAG...

Contigs

Genome Assembly

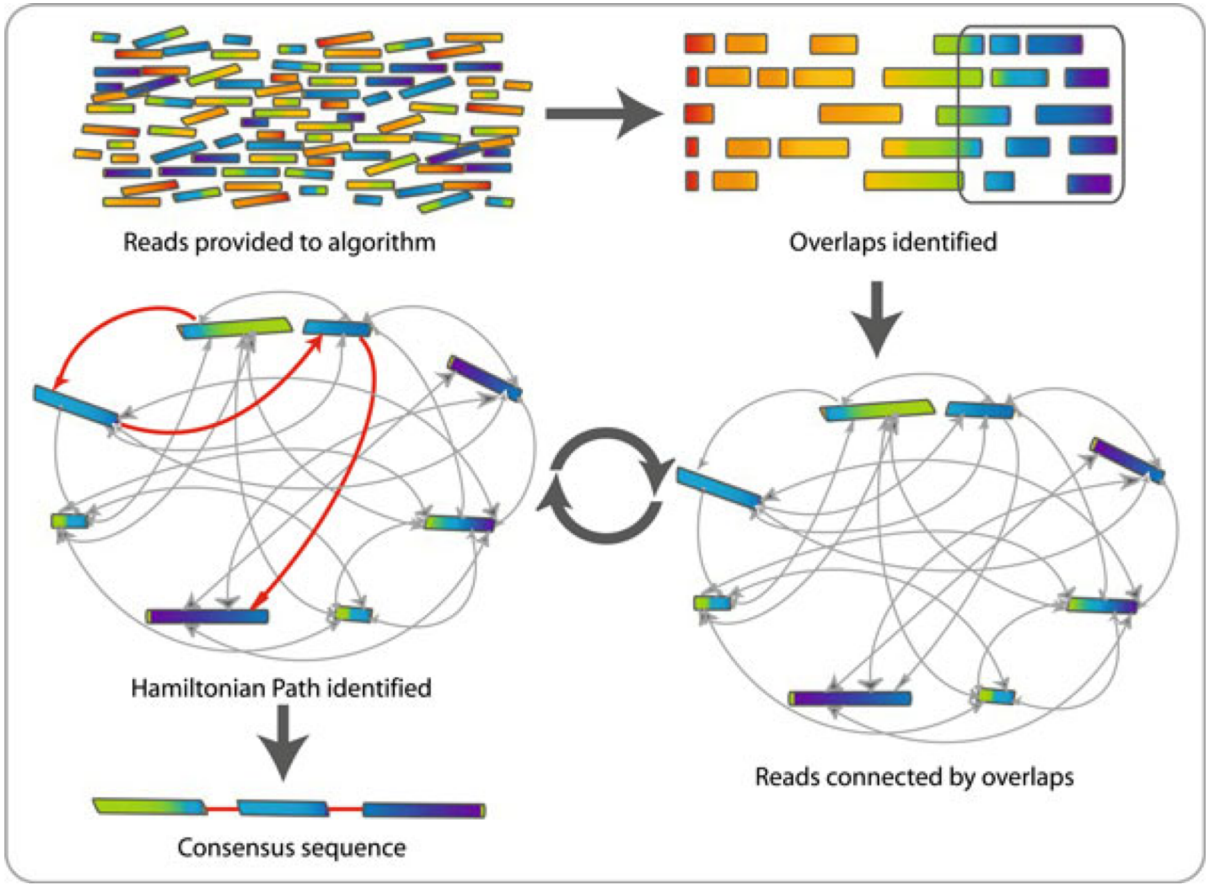


http://people.mpi-inf.mpg.de/~sven/images/assembly.png

Well-Known Assembly Approaches

	Overlap Layout Consensus (OLC)	**De Bruijn Graph (DBG)**
Advantage	Read coherency	Lack of coherency
Disadvantage	Computationally intensive	Computationally tractable

Overlap Layout Consensus (OLC)



De Bruijn Graph Algorithm: K-mers

A substring of length K

S = A C G T T C G A

All 4 mers:

A C G T

C G T T

G T T C

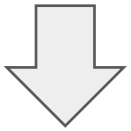
T T C G

T C G A

De Bruijn Graph Algorithm

Choose a value of k.

For each k-mer that exists in any sequence create an edge with one vertex labeled as the prefix and one vertex labeled as the suffix.



Glue all vertices that have the same label.

(Pevzner, Tang & Tesler, 2004)

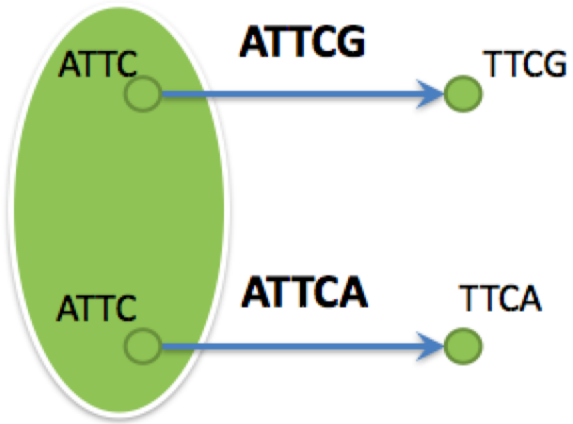
De Bruijn Graphs Construction

GTCT**ATTCG**CTA**ATTCA**CTA

(Pevzner, Tang & Tesler, 2004)

De Bruijn Graphs Construction

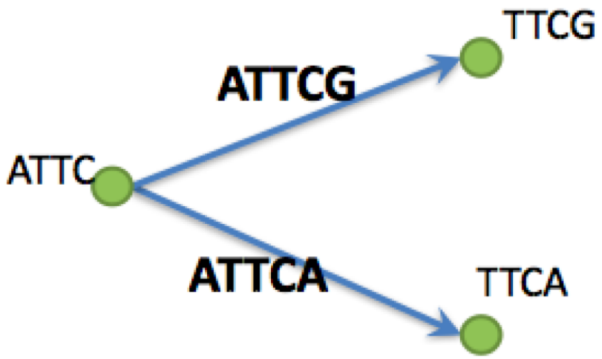
GTCT**ATTCG**CTA**ATTCA**CTA



(Pevzner, Tang & Tesler, 2004)

De Bruijn Graphs Construction

GTCT**ATTCG**CTA**ATTCA**CTA



(Pevzner, Tang & Tesler, 2004)

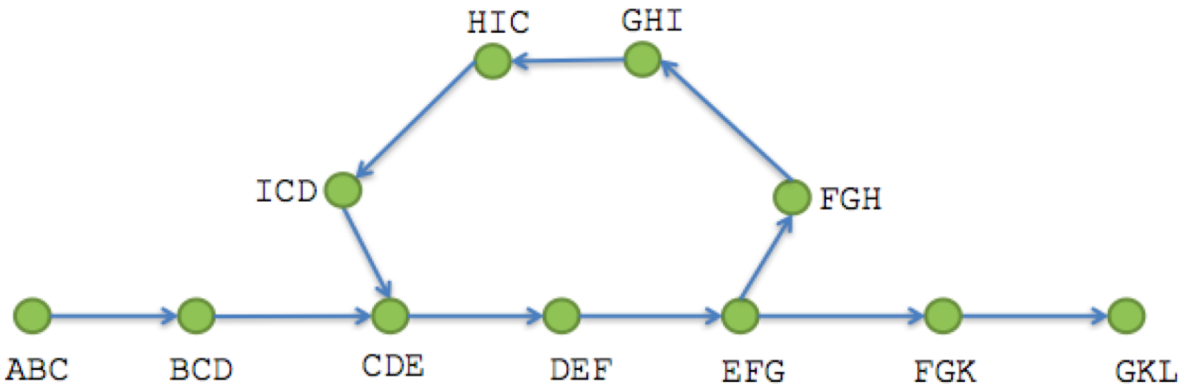
De Bruijn Graph

Sequence Read: ABCDEFGHICDEFGKL

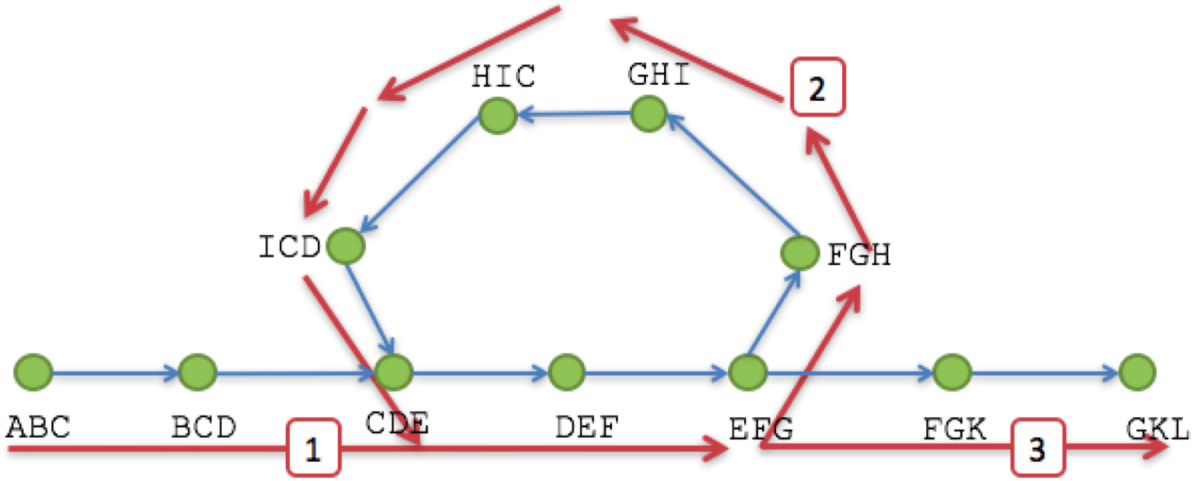
k-mers	$(k-1)$-mers	
ABCD	ABC	HIC
BCDE	BCD	ICD
CDEF	CDE	FGK
DEFG	DEF	GKL
EFGH	EFG	
GHIC	GHI	

De Bruijn Graph

Example Genome: ABCDEFGHICDEFGKL

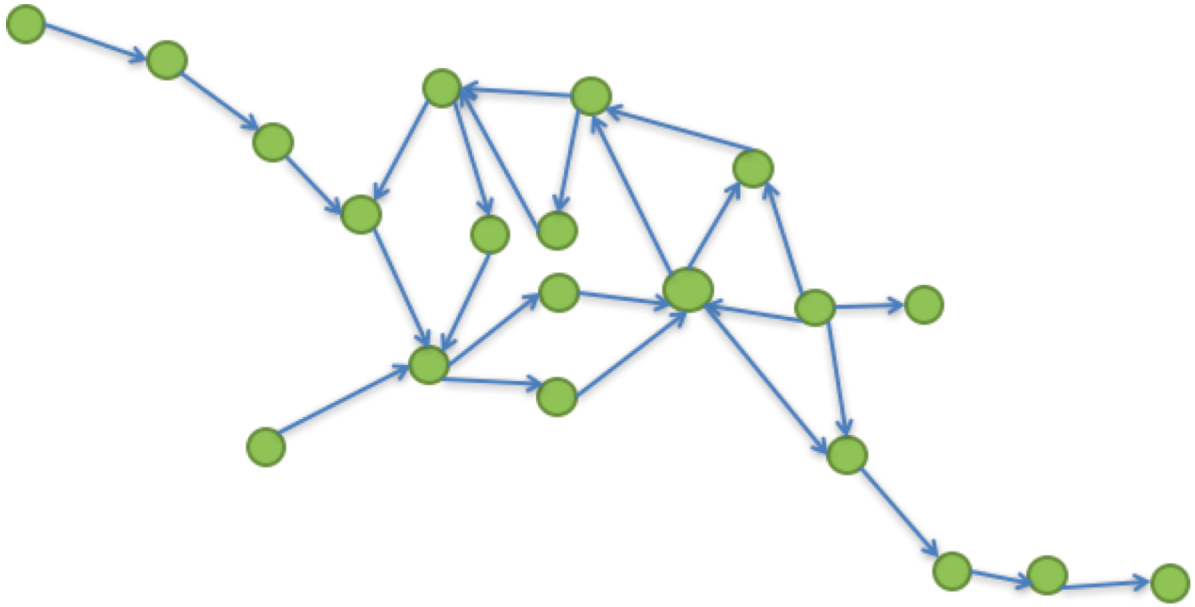


Traversing: Find walks on DBG



Contig: ABCDEFGHICDEFGKL

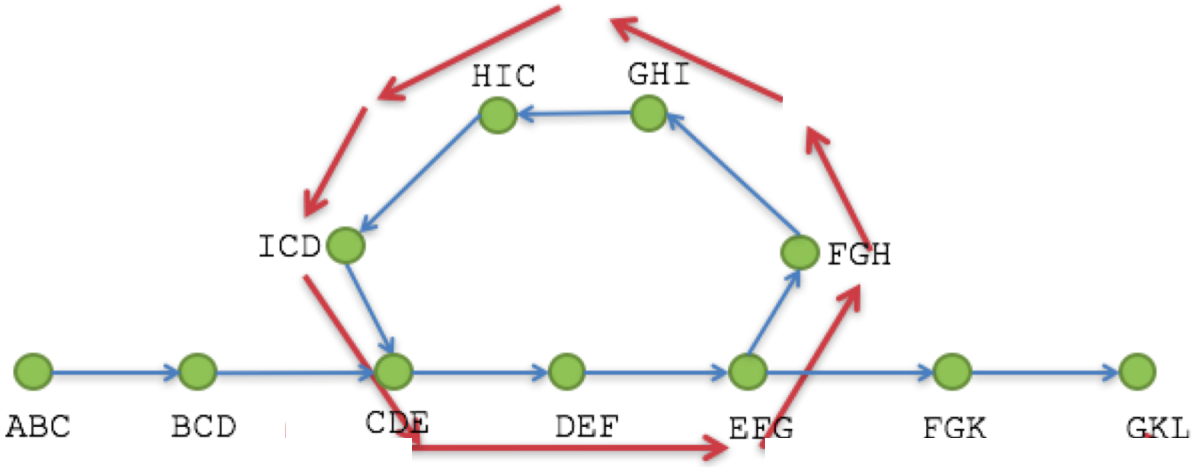
Typical De Bruijn Graph



over a **billion** nodes for a very small bacteria genome

Ambiguities in Traversing

Repeats ⟶ **Whirls** (Directed cycles)



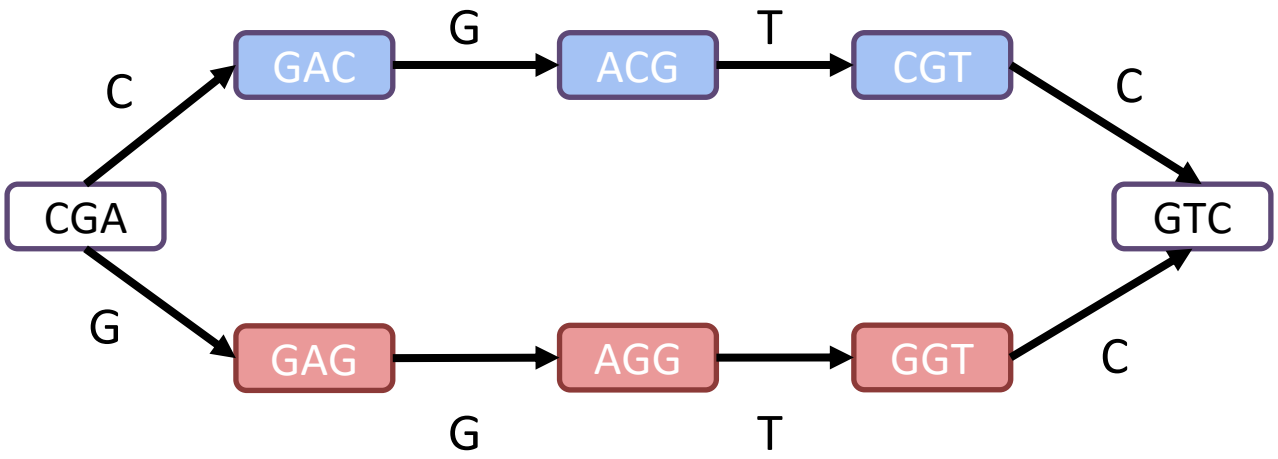
ABCDEFGHICDEFGKL

Ambiguities in Traversing

Variants → **Bulges** (Undirected cycles)

Read 1 = CGACGTC

Read 2 = CGAGGTC



Tackling the Ambiguities in Traversing

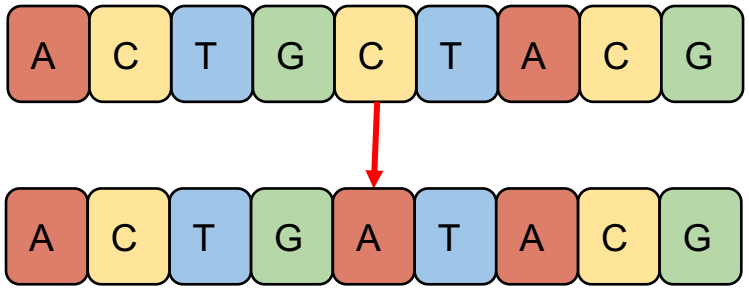
Auxiliary information to guied the traversing

- Positional de Bruijn graph
- Type of colored de Bruijn graph (read-colored de bruijn graph)

Genome Variant Calling

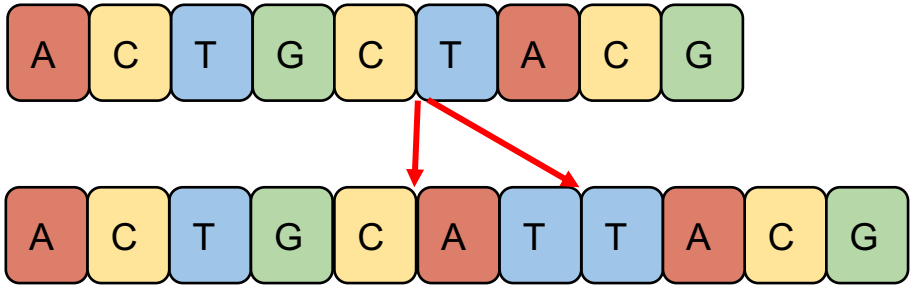
Genome Variants

Single Nucleotide Polymorphism (SNP)

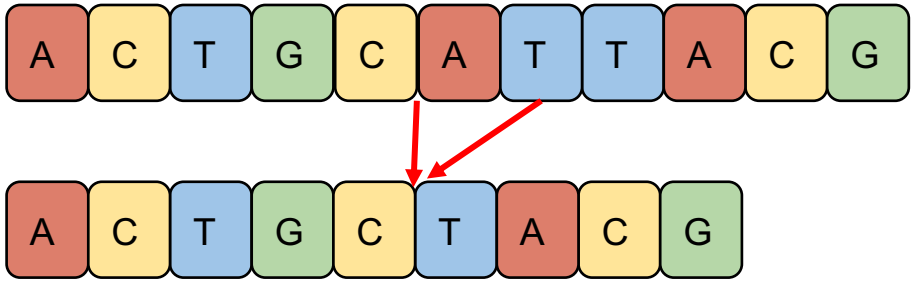


Genome Variants

Insertion



Deletion



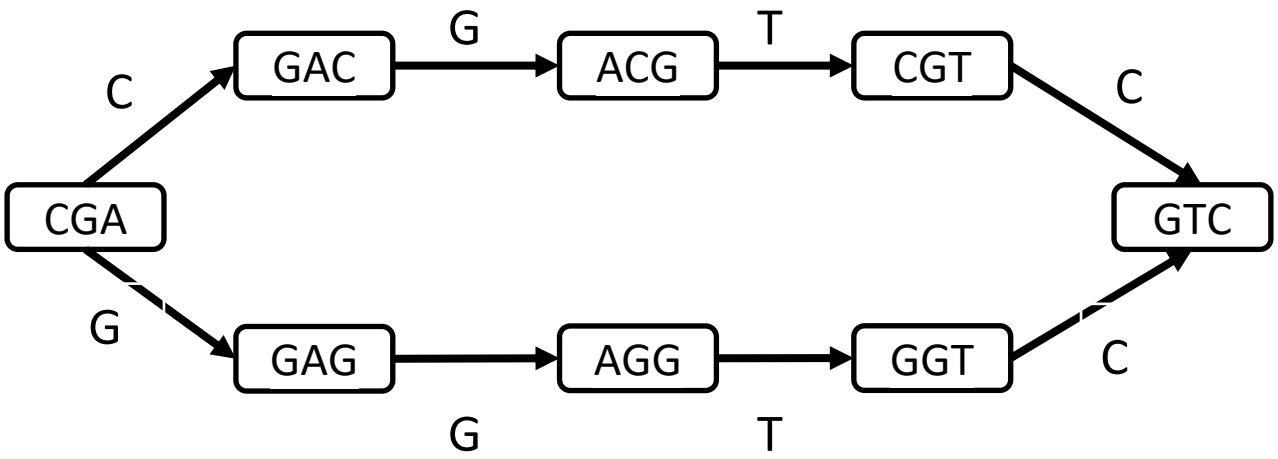
Well-Known Variant Callers

	Reference-based	Reference-free	
		Overlap-Layout Consensus	De Bruijn Graph
Advantage	Read-coherent	Read-coherent	Non-read-coherent
Disadvantage	Unculturable species	Inefficient	Efficient

Variant Calling using DBG

Sample 1 = CGACGTC

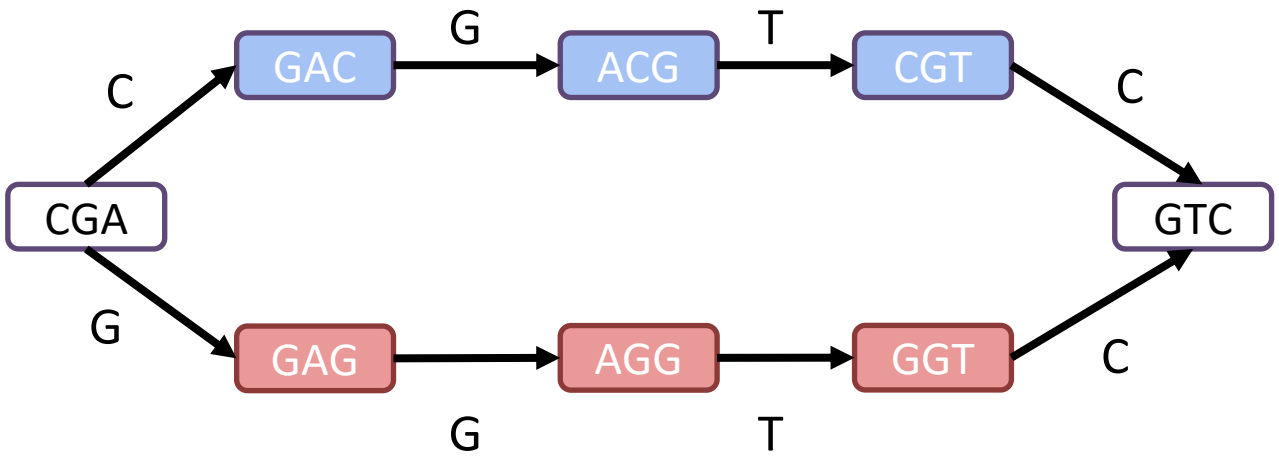
Sample 2 = CGAGGTC



Variant Calling using Colored DBG

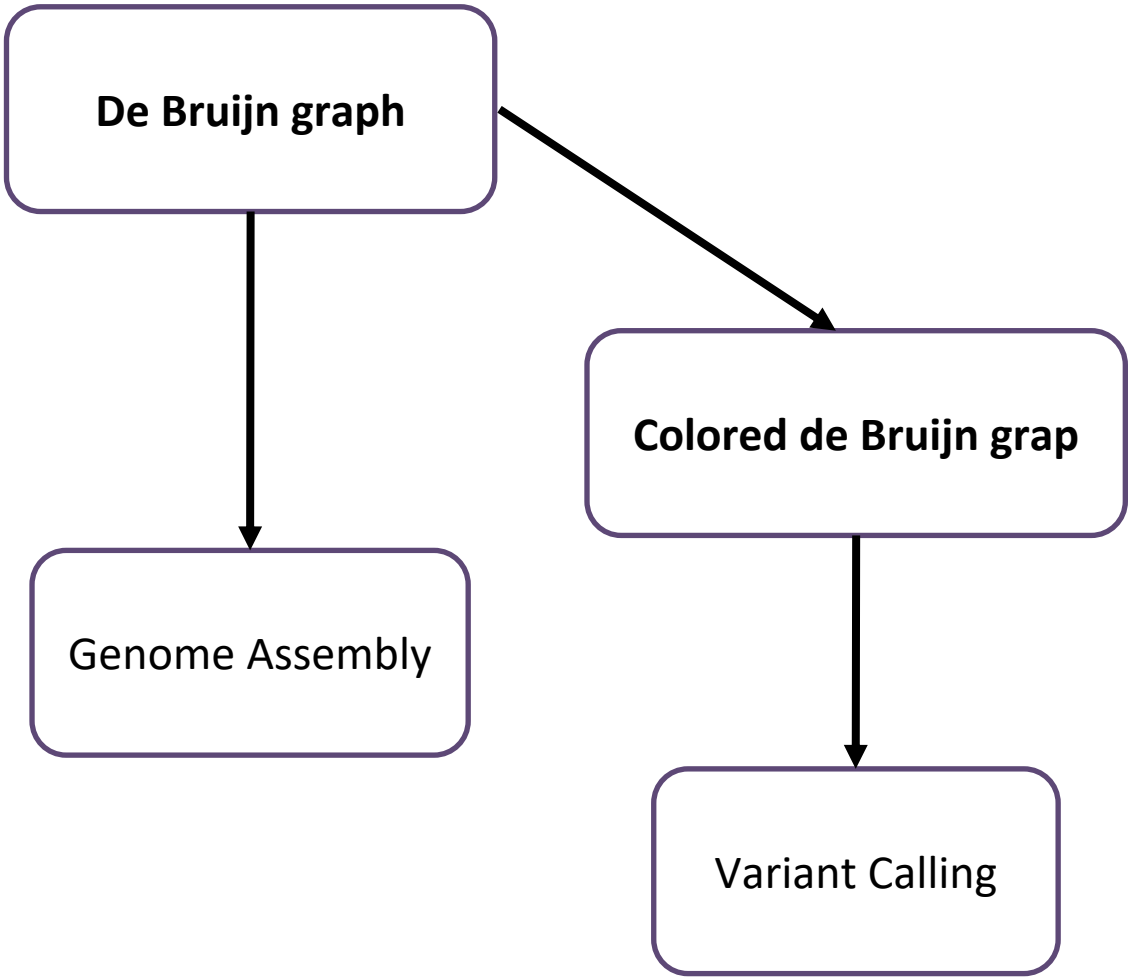
Sample 1 = CGACGTC

Sample 2 = CGAGGTC



	S1	S2
ACGT	1	0
AGGT	0	1
CGAC	1	0
CGAG	0	1
CGTC	1	0
GACG	1	0
GAGG	0	1
GGTC	0	1

Conclusion


De Bruijn graph
Colored de Bruijn grap
Genome Assembly
Variant Calling

Conclusion

- Following topics are being stutied in lots of projects:
  - Colored de Bruijn graph
  - Read colored de Bruijn graph
  - Positional de Bruijn graph
  - Variable-ordered de Bruijn graph
  - Succinct representation of de Bruijn graph

Questions?

Genome Assembly Challenges

- > 50% of human genome are repeats