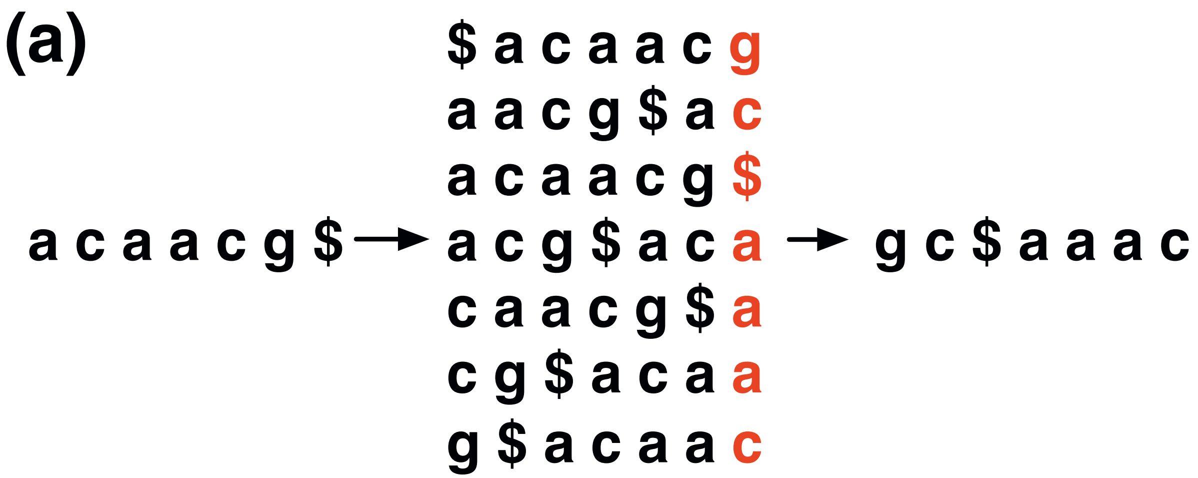
Burrows Wheeler Transform (BWT)



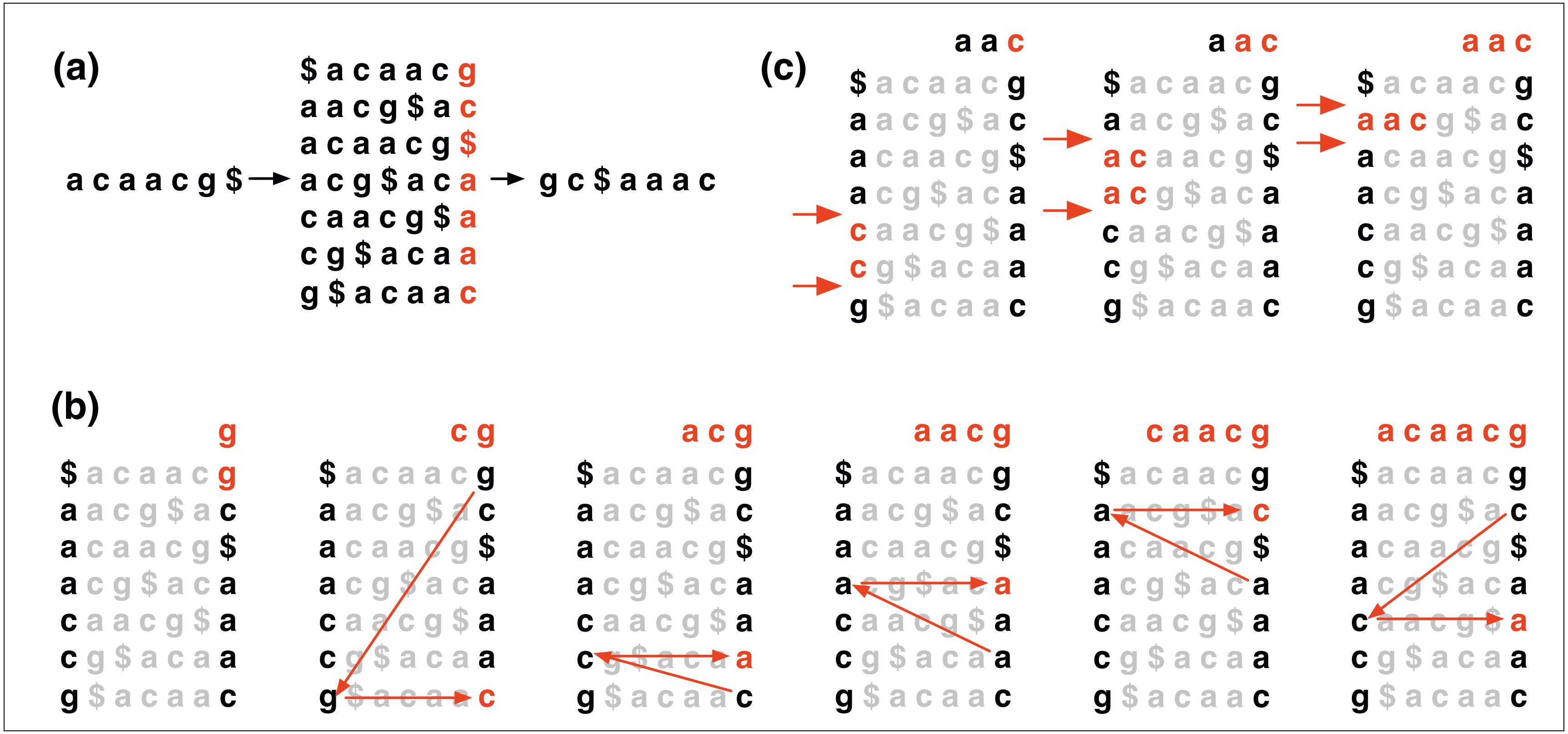
(a)

```
acaacg$ → $acaacg → gc$aaac
          aacg$ac
          acaacg$
          acg$aca
          caacg$a
          cg$acaa
          g$acaac
```

```
$acaacg
aacg$ac
acaacg$
acg$aca
caacg$a
cg$acaa
g$acaac
```

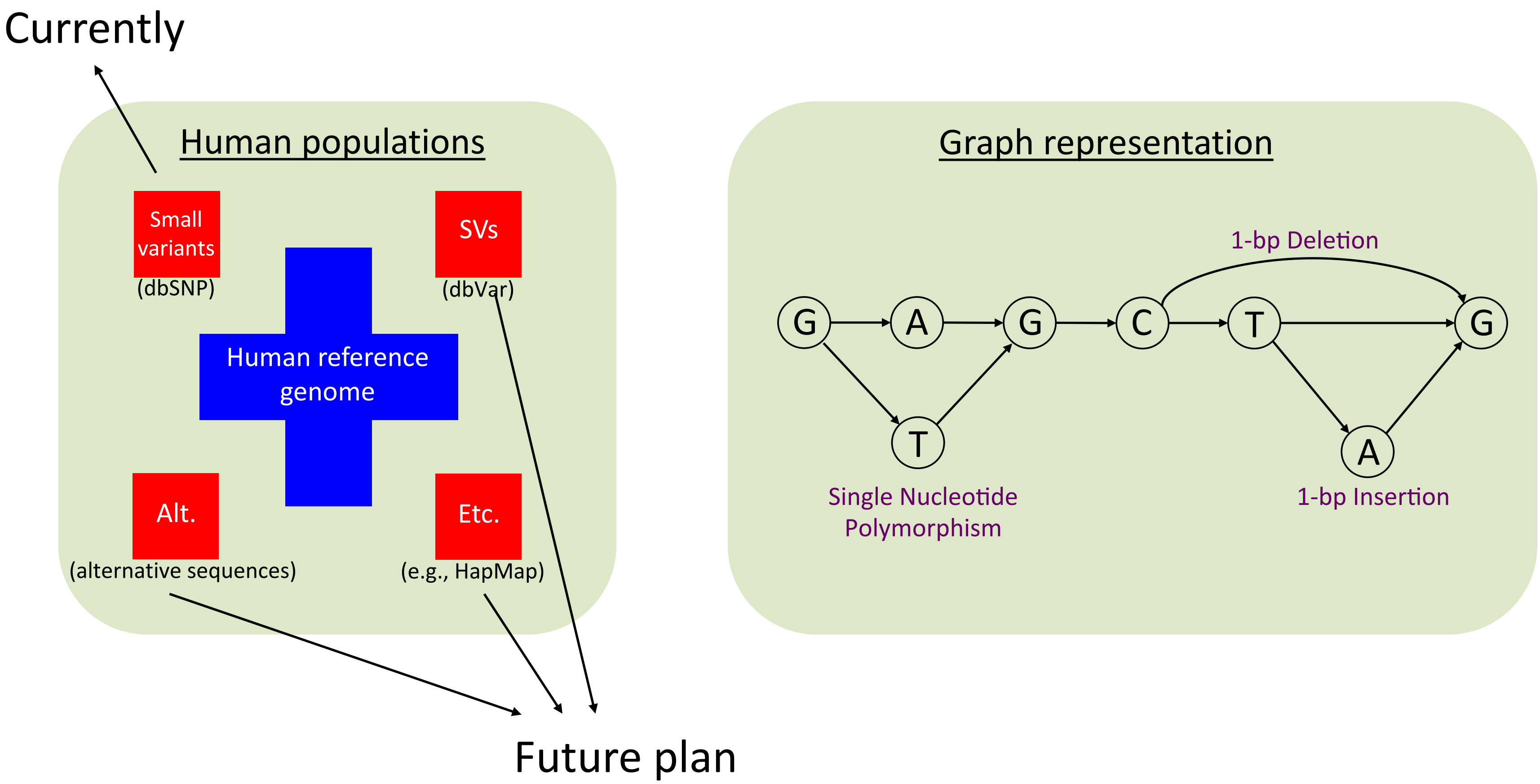
Bowtie paper

Burrows Wheeler Transform (BWT)



Bowtie paper

HISAT2 – Encompassing Diversity



Currently

Human populations

Small variants (dbSNP)

SVs (dbVar)

Human reference genome

Alt. (alternative sequences)

Etc. (e.g., HapMap)

Future plan

Graph representation

1-bp Deletion

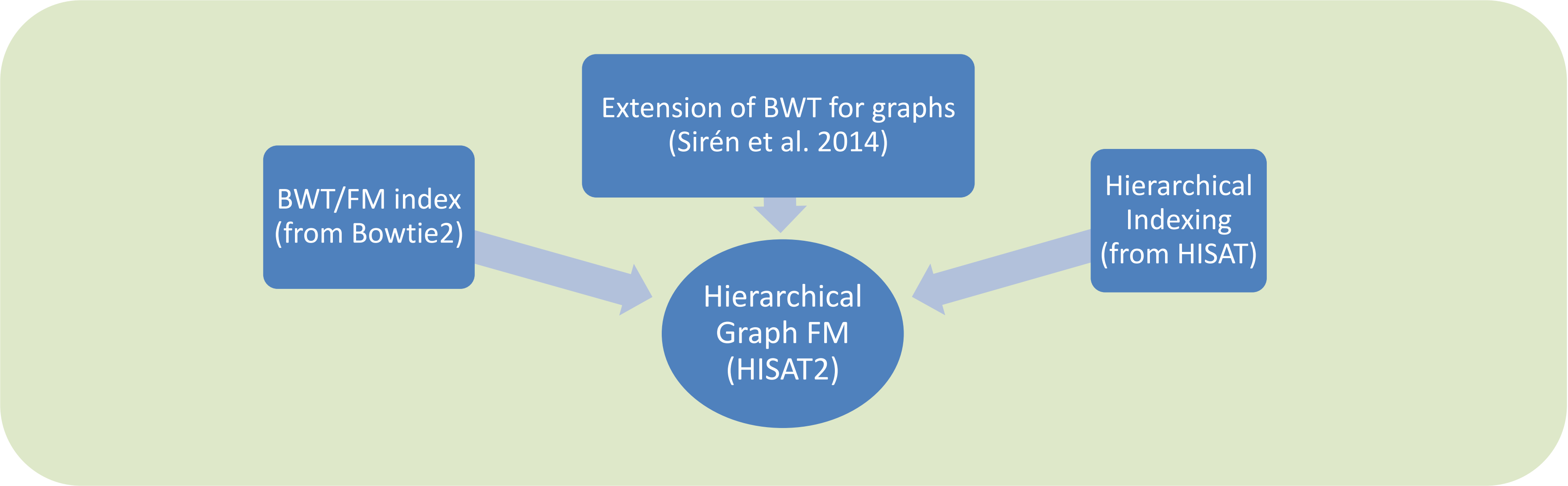
Single Nucleotide Polymorphism

1-bp Insertion

First version of HISAT2 (Sept. 2015)

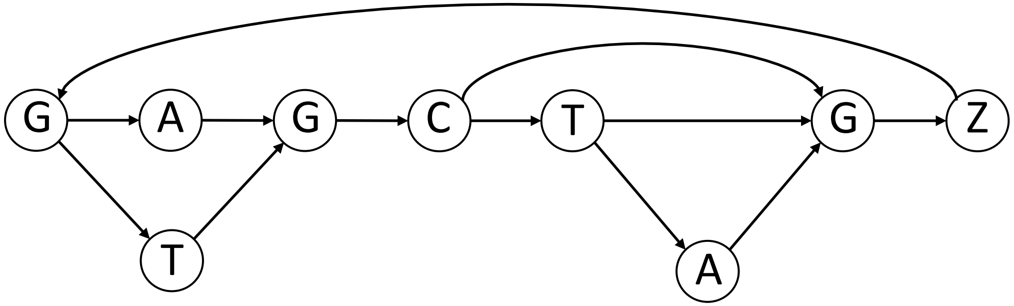
- Memory lean (6.7 GB for human genome and 12.3M small variants including 1.3M insertions and deletions)
- Rapid alignment comparable to HISAT

HISAT2 – Key Technologies


Extension of BWT for graphs (Sirén et al. 2014)
BWT/FM index (from Bowtie2)
Hierarchical Indexing (from HISAT)
Hierarchical Graph FM (HISAT2)

- BWT (1994) BWT for linear string
  - Burrows M, Wheeler DJ: A Block Sorting Lossless Data Compression Algorithm. Technical Report 124. Palo Alto, CA: Digital Equipment Corporation; 1994.
- FM (2000) BWT + metadata for fast lookup
  - Ferragina, P. & Manzini, G. Proc. 41st Annual Symposium on Foundations of Computer Science 390–398 (IEEE Comput. Soc.; 2000).
- Bowtie2 (2012) Very efficient implementation of BWT/FM index
  - Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
- GCSA (2014) BWT for graph
  - Sirén J, Välimäki N, Mäkinen V (2014) Indexing graphs for path queries with applications in genome research. IEEE/ACM Transactions on Computational Biology and Bioinformatics 11: 375–388. doi: 10.1109/tcbb.2013.2297101
- HISAT (2015) Hierarchical Indexing
  - Kim D, Langmead B and Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 2015
- HISAT2 - GFM and HGFM *[in progress]* Graph BWT + metadata for fast lookup
  - Kim D., Paggi J., Salzberg S.L.

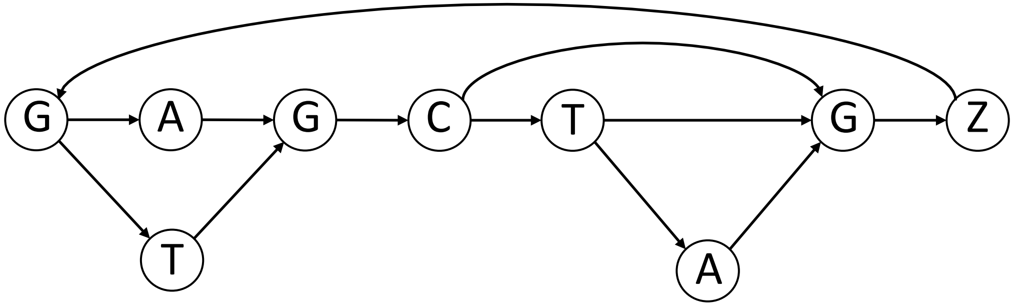
Basic Terms (Path, String, Order)



- A path defines a string
  - For example, a path G->A->G->C defines a string, GAGC

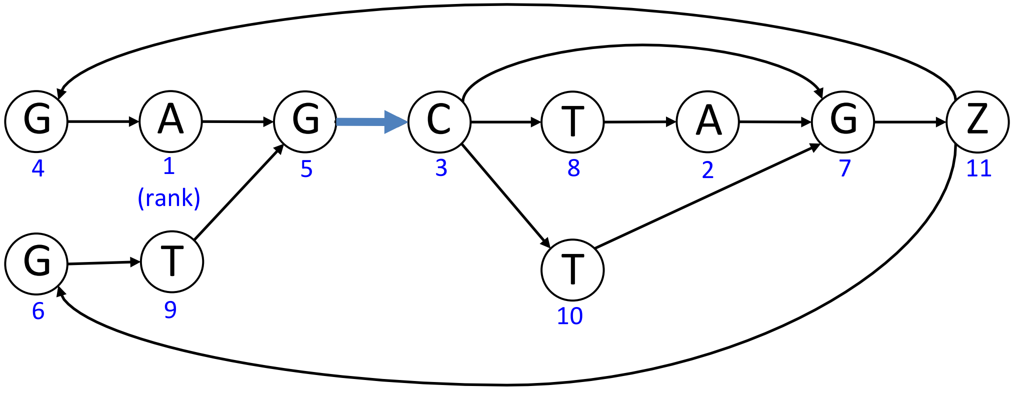
- Strings can be ordered lexicographically
  - For example, AGC comes before GTG, which comes before TGZ

Prefix-sorted Graph



Original graph

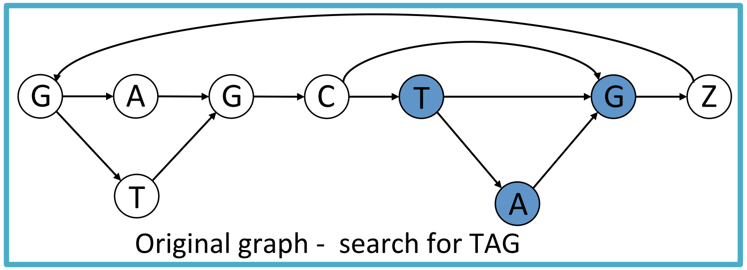
Prefix-doubling and pruning (Sirén et al. 2014)



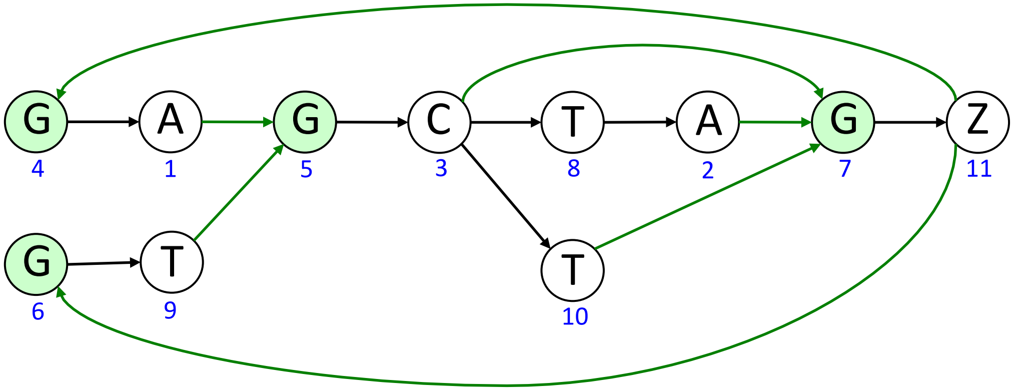
Prefix-sorted graph

Last-First (LF) mapping

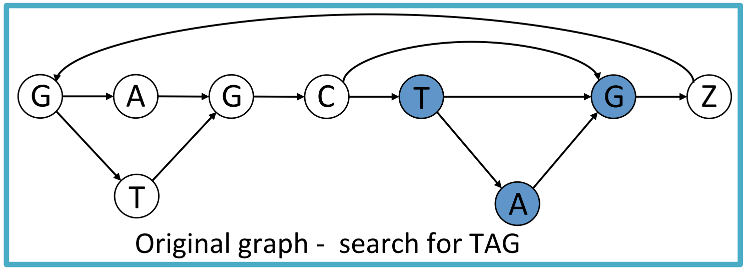
Outgoing edge(s)			Incoming edge(s)	
Node rank	First		Last	Node rank
1	A		G	1
2	A		T	2
3	C		G	3
	C		Z	4
	C		A	5
4	G		T	
5	G		Z	6
6	G		A	7
7	G		C	
8	T		T	
9	T		C	8
10	T		G	9
11	Z		C	10
	Z		G	11



Original graph - search for TAG

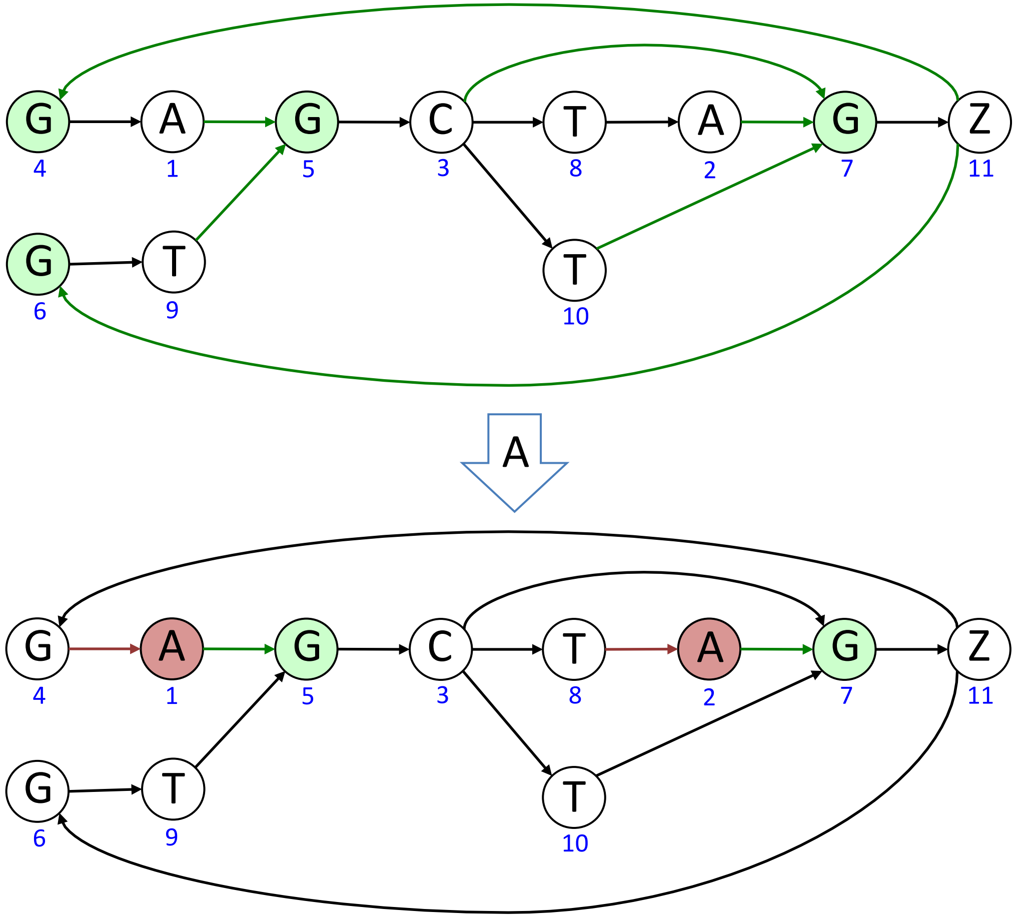


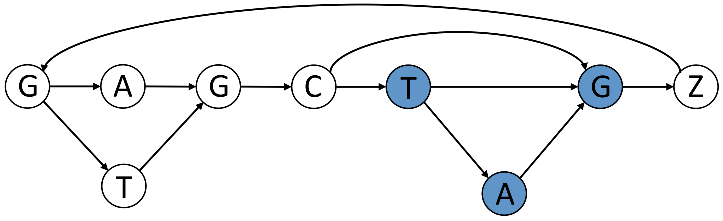
Outgoing edge(s)			Incoming edge(s)	
Node rank	First		Last	Node rank
1	A		G	1
2	A		T	2
3	C C C		G	3
4	G		Z	4
5	G		A T	5
6	G		Z	6
7	G		A C T	7
8	T		C	8
9	T		G	9
10	T		C	10
11	Z Z		G	11



Original graph - search for TAG

Outgoing edge(s)			Incoming edge(s)	
Node rank	First		Last	Node rank
1	A		G	1
2	A		T	2
3	C C C		G	3
4	G		Z	4
5	G		A T	5
6	G		Z	6
7	G		A C T	7
8	T		C	8
9	T		G	9
10	T		C	10
11	Z Z		G	11

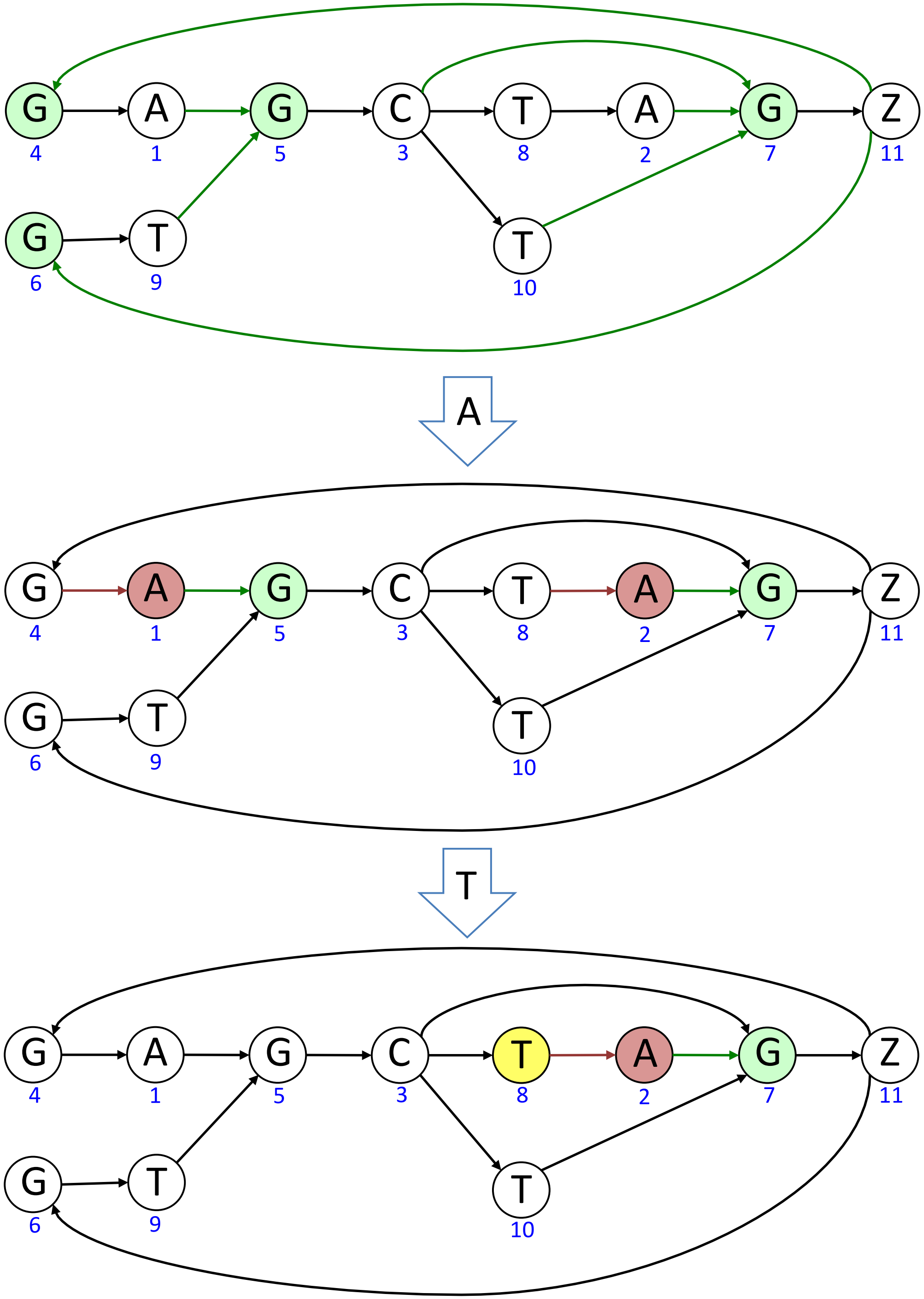




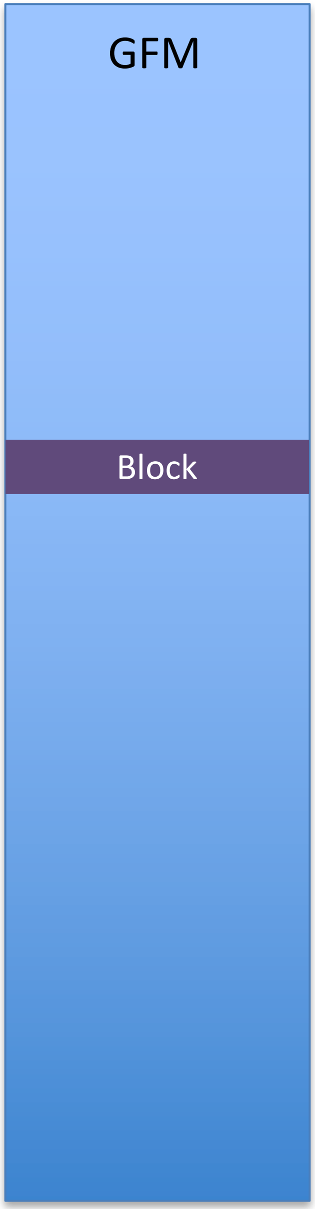
Original graph - search for TAG

Outgoing edge(s)	
Node rank	First
1	A
2	A
3	C C C
4	G
5	G
6	G
7	G
8	T
9	T
10	T
11	Z Z

Incoming edge(s)	
Last	Node rank
G	1
T	2
G	3
Z	4
A T	5
Z	6
A C T	7
C	8
G	9
C	10
G	11

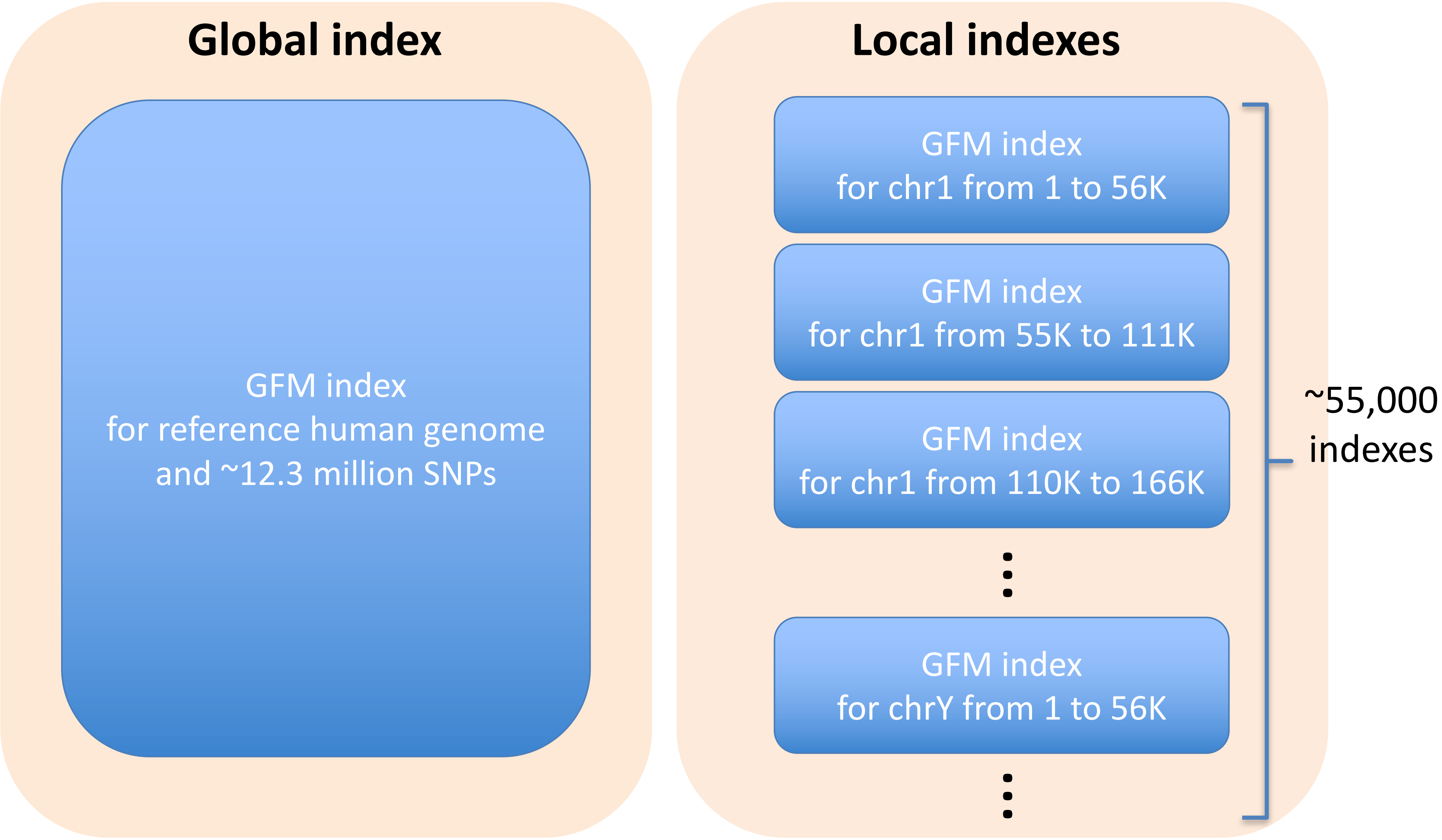


HISAT2 - Graph FM index (GFM)



- Each block
  - Stores a few hundred labels for outgoing and incoming edges
  - Stores numbers such as accumulated occurrences of A, C, G, T

HISAT2 - Hierarchical Graph FM index (HGFM)



Global index

GFM index
for reference human genome
and ~12.3 million SNPs

Local indexes

GFM index
for chr1 from 1 to 56K

GFM index
for chr1 from 55K to 111K

GFM index
for chr1 from 110K to 166K

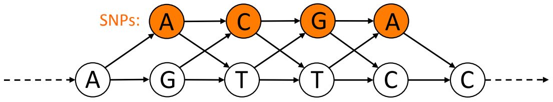
⋮

GFM index
for chrY from 1 to 56K

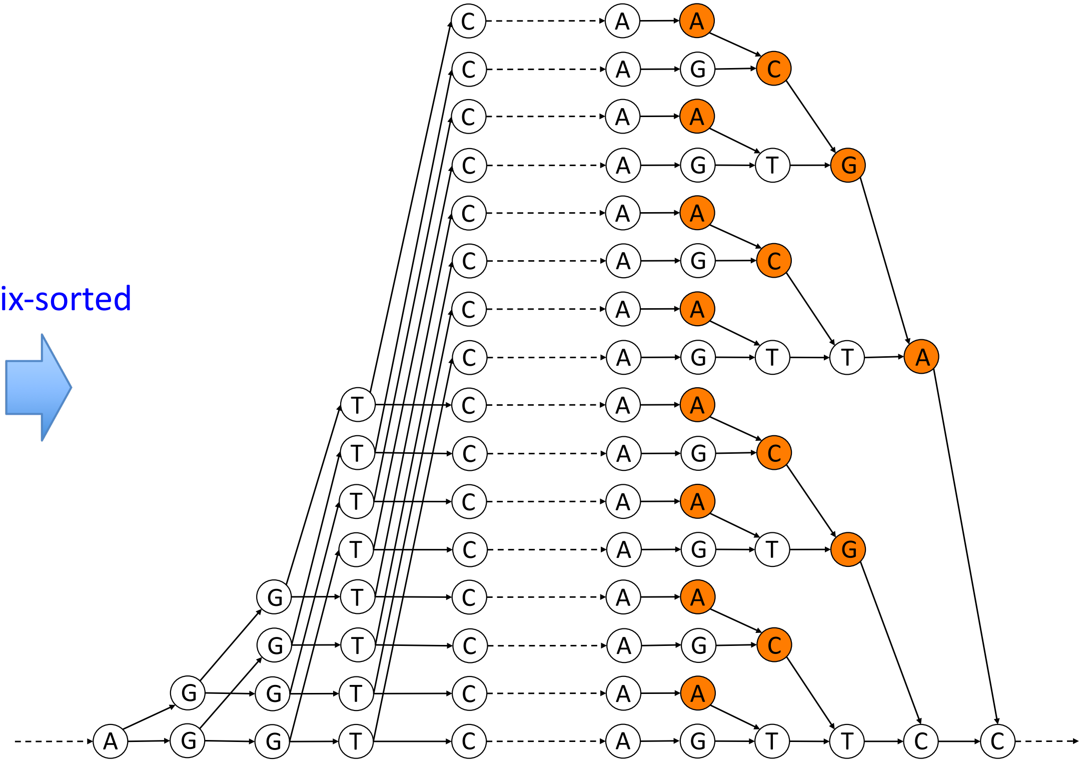
⋮

~55,000
indexes

Adjacent SNPs

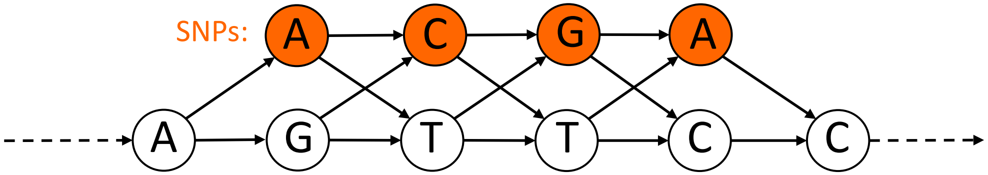


Prefix-sorted

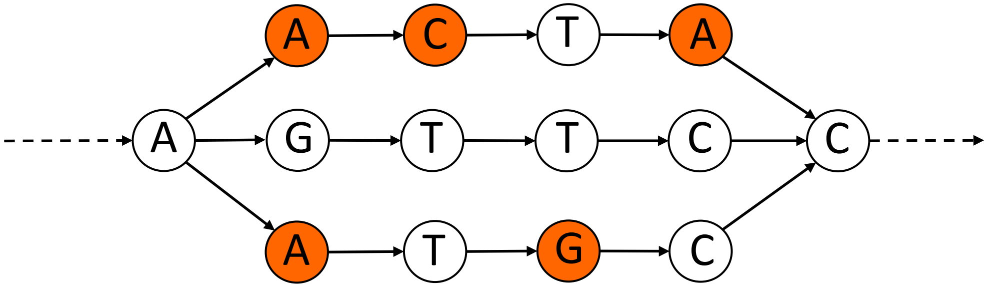


~ 240
(60 per SNP)

One Solution



Total combinations



- HapMap
- Haplotype Reference Consortium (HRC)
- 1,000 Genomes Project

Combinations found in human populations