Resequencing against a human whole genome variation graph

Erik Garrison
Sanger Wellcome Trust Institute
Quantitative Genomics 2015
May 29, 2015

UNIVERSITY OF CAMBRIDGE

Overview

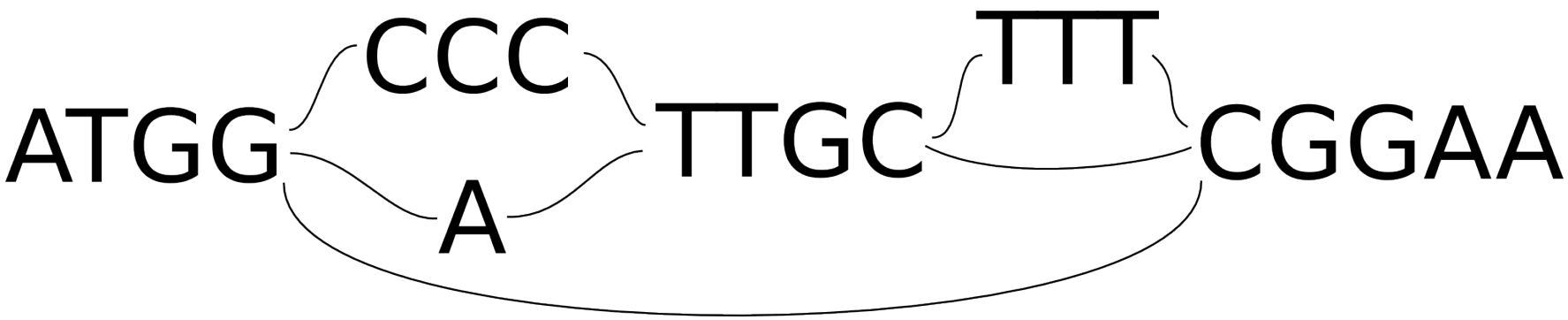
1. Variation + sequence = variation graphs
2. Local realignment to a variation graph
3. Constructing a whole genome variation graph
4. Aligning to a whole genome variation graph

(1) Variation + sequence

= variation graph

Variation graphs

A variation graph represents many genomes in the same context.



Nodes contain sequence and directed edges represent potential links between successive sequences.

A multiple sequence alignment is a variation graph

traditional MSA

(a)

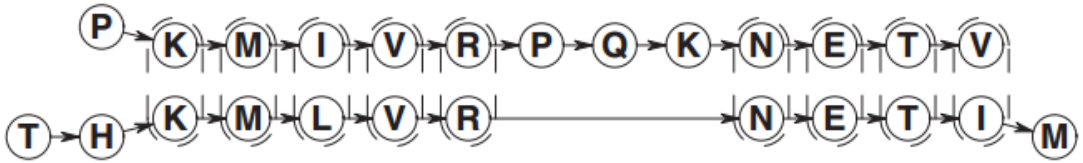
```
.  .  P  K  M  I  V  R  P  Q  K  N  E  T  V  .
T  H  .  K  M  L  V  R  .  .  .  N  E  T  I  M
```

consensus sequence

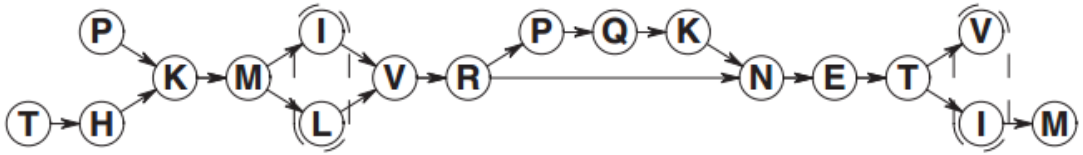
(b)

positionally-matching regions aligned

(c)

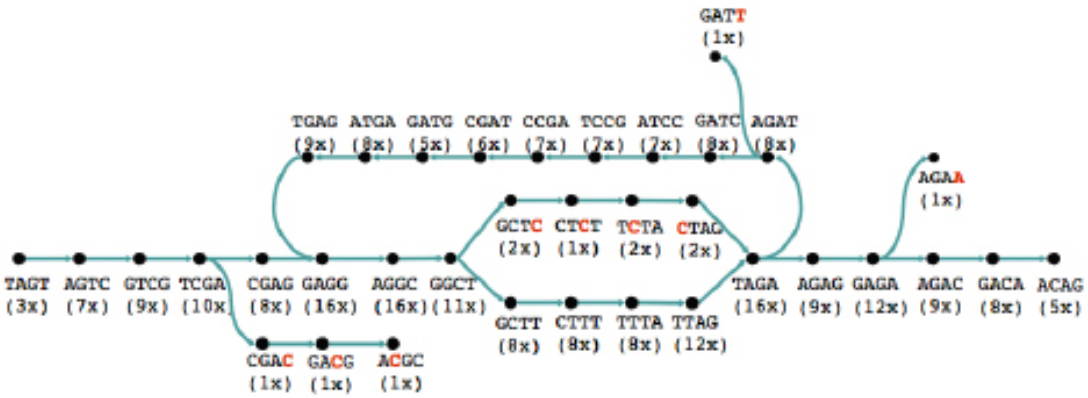


(d)



Christopher Lee, Catherine Grasso, Mark F. Sharlow. **Multiple sequence alignment using partial order graphs.** *Bioinformatics*, 2002.

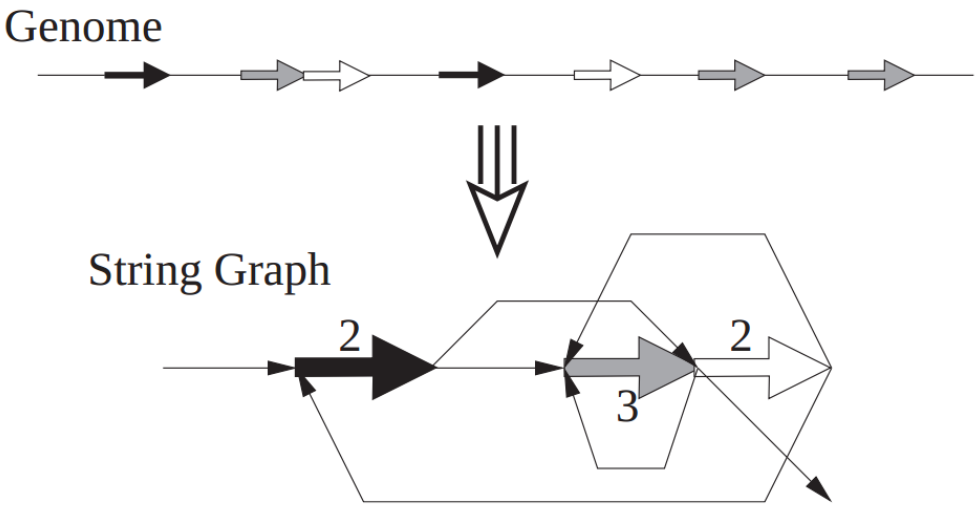
Assembly graphs are variation graphs



A de Bruijn graph can be converted into a variation graph (as previous) by setting node sequences to the first letter in the kmer and compressing non-branching runs into single nodes.

http://plus.maths.org/content/os/issue55/features/sequencing/index, credit Daniel Zerbino

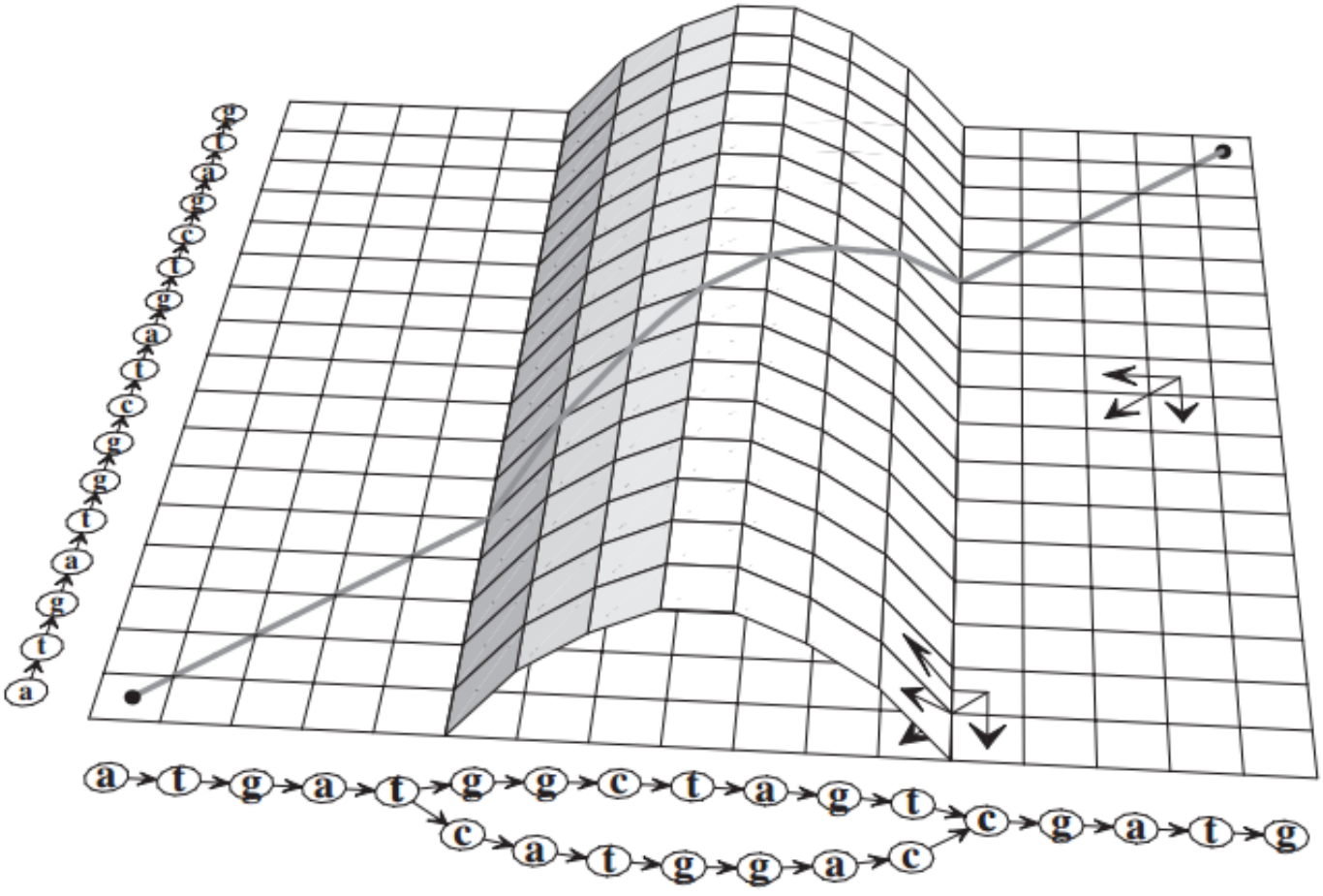
A string graph follows the same format as the variation graph.



Eugene Myers. **The fragment assembly string graph**. *Bioinformatics*, 2005.

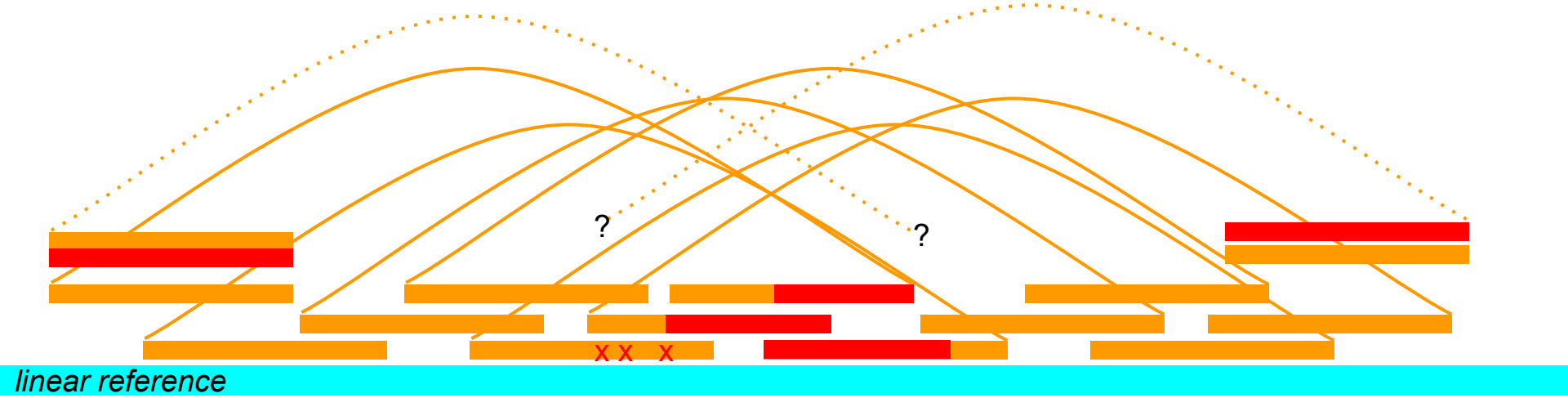
(2) Local realignment to a variation graph

Local alignment against a graph

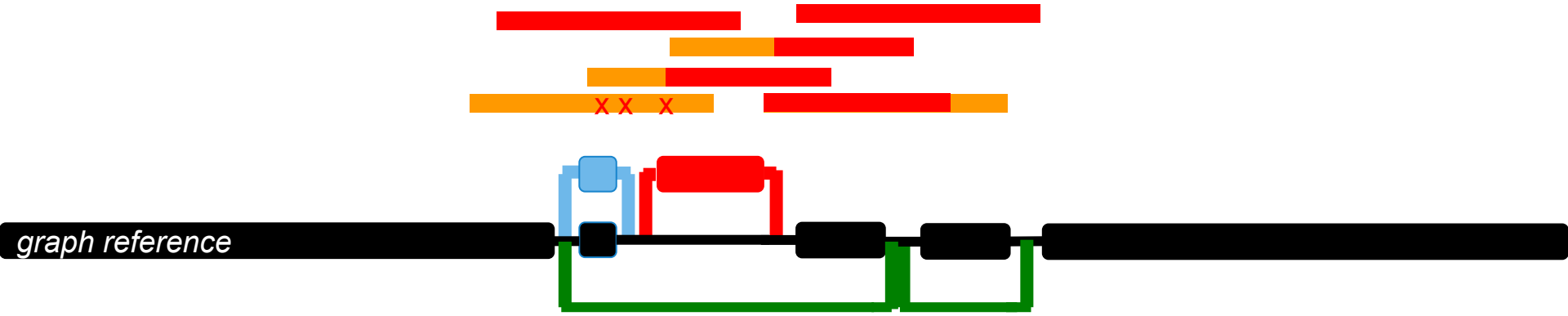


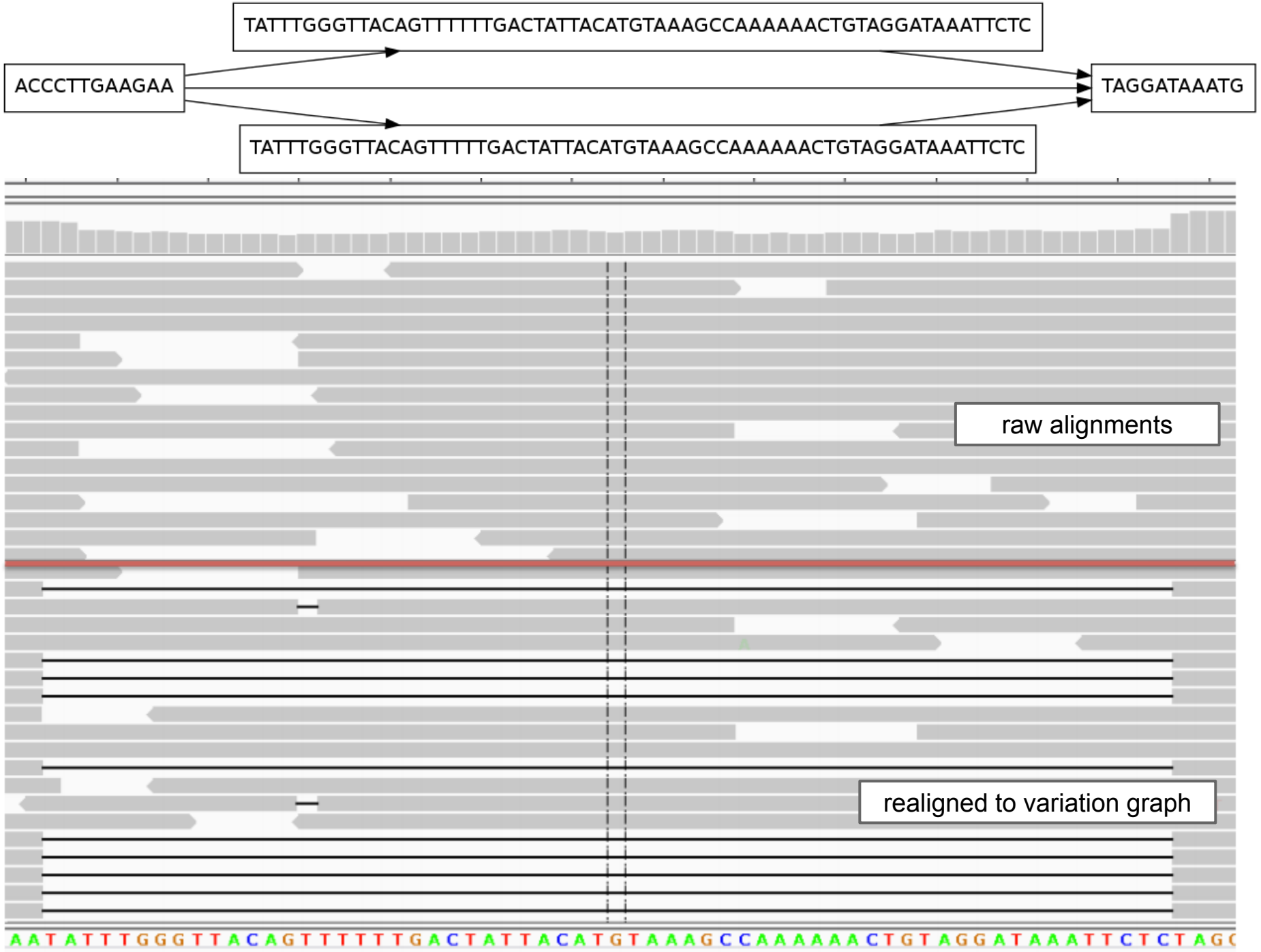
Christopher Lee, Catherine Grasso, Mark F. Sharlow. **Multiple sequence alignment using partial order graphs.** *Bioinformatics*, 2002.

Seeding graph-based alignments

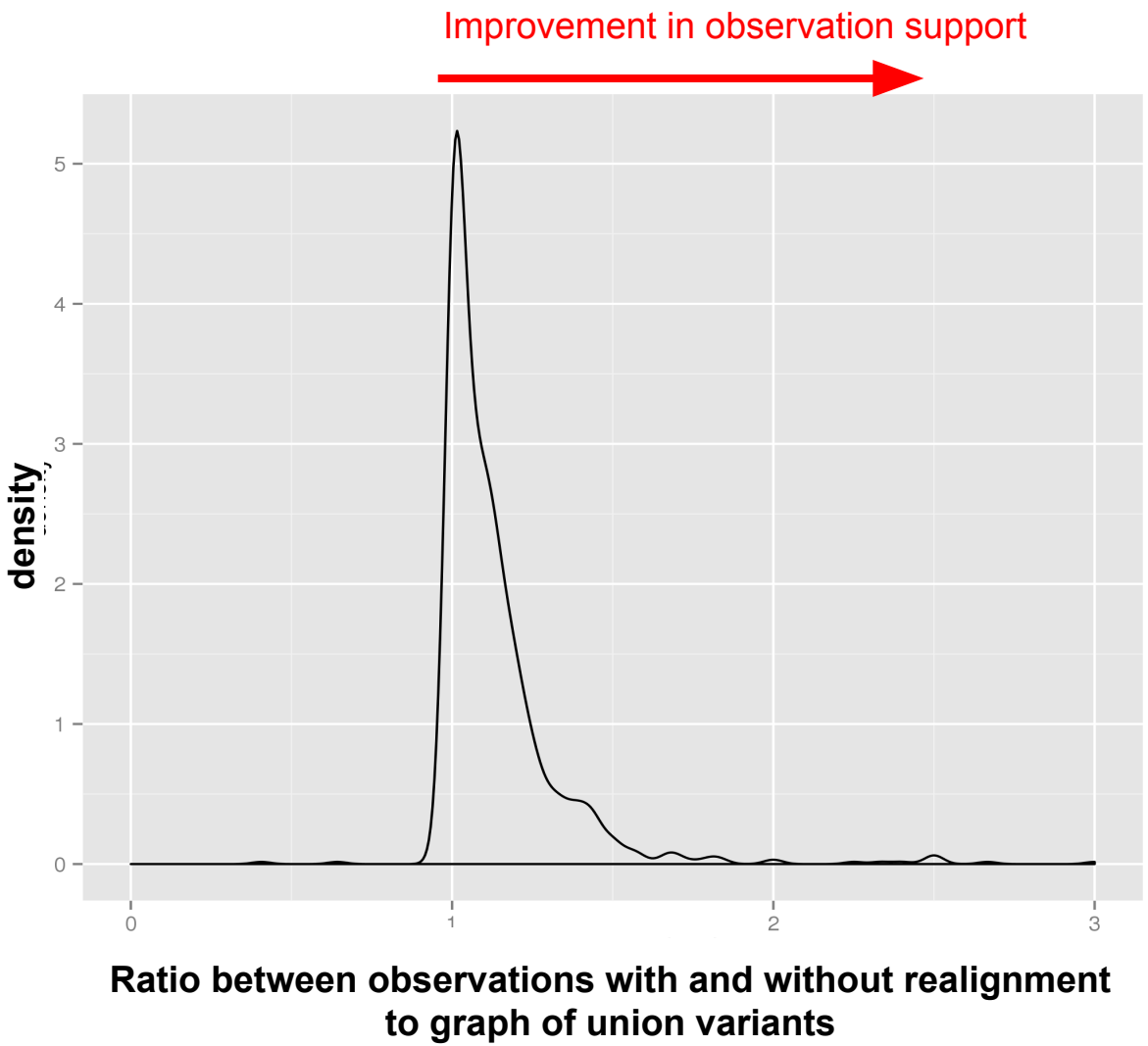


Test imperfectly-mapped reads against graph.




TATTTGGGTTACAGTTTTTTGACTATTACATGTAAAGCCAAAAAAACTGTAGGATAAATTCTC
ACCCTTGAAGAA
TAGGATAAATG
TATTTGGGTTACAGTTTTTGACTATTACATGTAAAGCCAAAAAAACTGTAGGATAAATTCTC
raw alignments
realigned to variation graph
AATATTTGGGTTACAGTTTTTTTGACTATTACATGTAAAGCCAAAAAAACTGTAGGATAAATTCTCTAG

realignment to variation graph reduces reference bias



Standard alignment is frustrated even by small variants

(tested against 1mb segment on chr20 using 1000Gp3 union allele list)

(3) Constructing a whole genome variation graph

Construction

```
POS     ID      REF     ALT
...
```

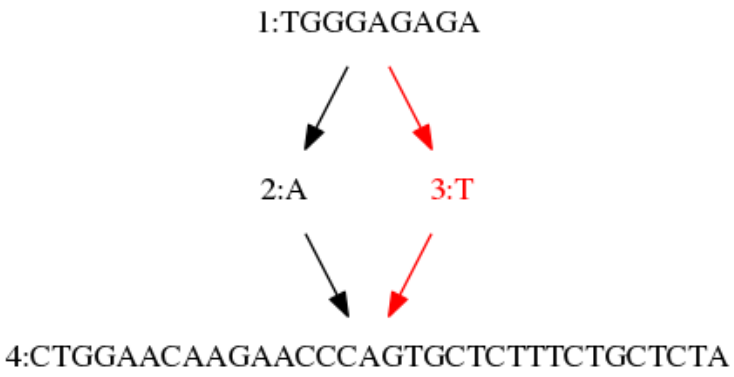
1:TGGGAGAGAACTGGAACAAGAACCCAGTGCTCTTTCTGCTCTA

For each variant

1. cut the reference path around the variant
2. add the novel (ALT) sequence to the graph

Construction

```
POS     ID      REF     ALT
10      .       A       T
...
```



For each variant

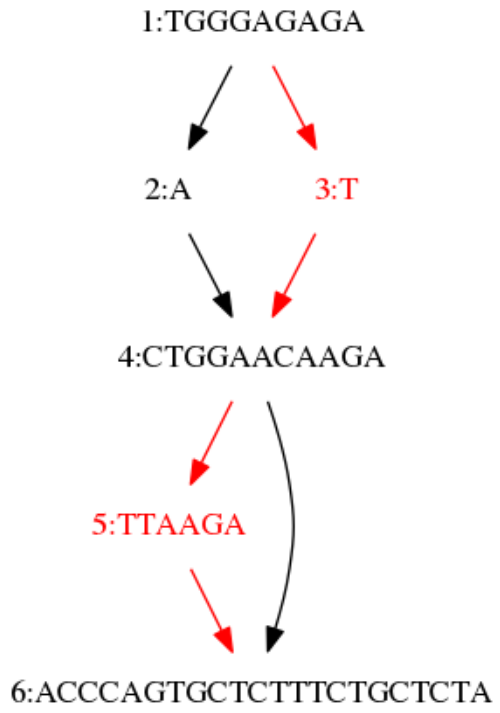
1. cut the reference path around the variant
2. add the novel (ALT) sequence to the graph

Construction

POS	ID	REF	ALT
10	.	A	T
21	.	A	ATTAAGA
...			

For each variant

1. cut the reference path around the variant
2. add the novel (ALT) sequence to the graph

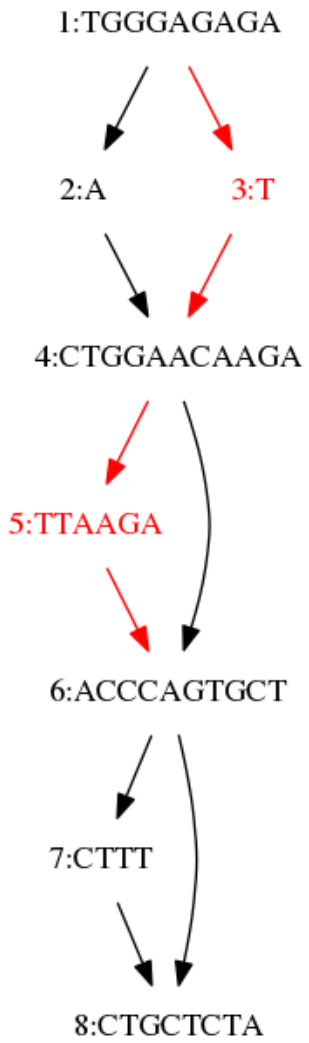


Construction

POS	ID	REF	ALT
10	.	A	T
21	.	A	ATTAAGA
31	.	TCTTT	T

For each variant

1. cut the reference path around the variant
2. add the novel (ALT) sequence to the graph



Data model (vg.proto)

Use protocol buffers to define objects

Graph, Node, Edge, Path, Alignment, Edit

```
message Graph {
    repeated Node node = 1;
    repeated Edge edge = 2;
    repeated Path path = 3;
}
message Edge {
    required int64 from = 1;
    required int64 to = 2;
    optional bytes data = 3;
    optional Annotation annotation = 4;
}
```

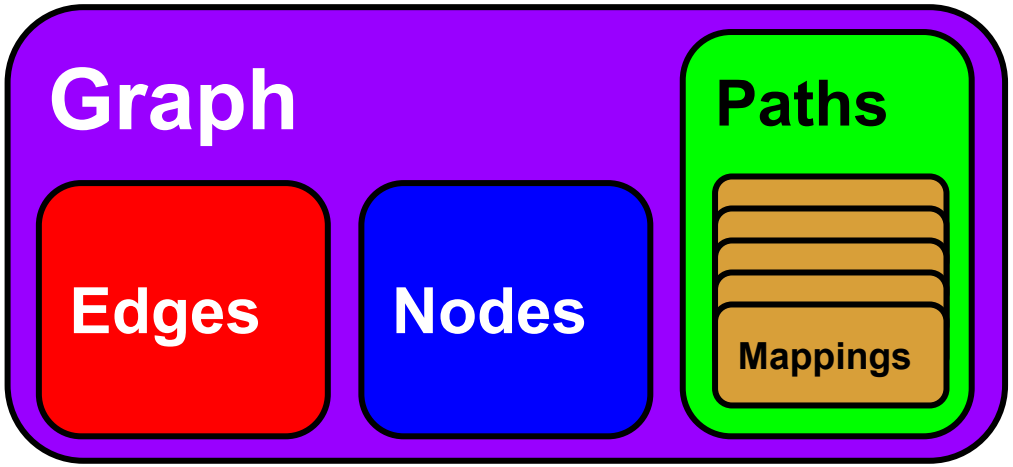
```
message Path {
    optional int32 target_position = 1;
    optional string name = 2;
    repeated Mapping mapping = 3;
}
```

```
message Node {
    optional string sequence = 1;
    optional string name = 2;
    required int64 id = 3;
    optional bytes data = 4;
    optional Annotation annotation = 5;
}
```

```
message Mapping {
    required int64 node_id =
1;
    repeated Edit edit = 2;
}
```

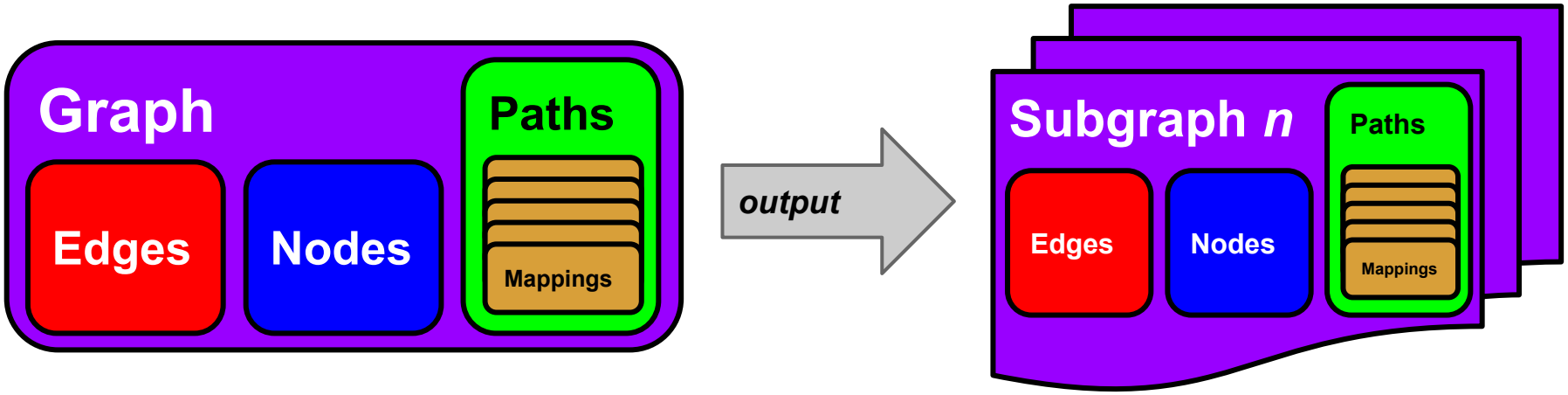
Format

Basic entity is a Graph, which is composed of nodes, edges, and paths.



Serialization

To serialize the graph, we generate a stream of sub-graphs that can be reassembled into the whole.



in memory **serialized stream**

Indexing

Instead of indexing the serialized graph representation (e.g. BAM, VCF), write the graph into a disk-backed key/value store:

{"key":"+g+26+n", "value":{"id": 26, "sequence": "TATTTGAAGT"}}
{"key":"+g+26+p+1+103", "value":{"node_id": 26, "edit": []}}
{"key":"+g+26+t+24", "value":{"from": 24, "to": 26}}
{"key":"+g+26+t+25", "value":{"from": 25, "to": 26}}
{"key":"+g+27+f+29", "value":{"from": 27, "to": 29}}
…
{"key":"+k+TCATACTACTG+69", "value":-2}
{"key":"+k+TCATACTACTG+70", "value":-4}
{"key":"+k+TCATACTACTG+72", "value":-5}
{"key":"+k+TCATATGTCCA+167", "value":29}
{"key":"+k+TCATATGTCCA+169", "value":-1}
{"key":"+k+TCATATGTCCA+171", "value":-2}
...

This allows our graphs and kmer indexes to have > main memory size. (Uses rocksdb.)

(4) Aligning to a whole genome variation graph

Constructing a whole human genome variation graph

Constructed 1000G phase3 + GRCh37 variation graph using `vg construct`.

5h20m on 32-core system @Sanger

3.07G on disk

3.181 Gbp of sequence in graph

286 million nodes (11 bp/node)

376 million edges (8.5 bp/edge)

Indexing 1000G+GRCh37

for each node:

for each *k*-path extending no more than *n* edges or *k* bp away from the node:

→ index the k-mers of the *k*-path

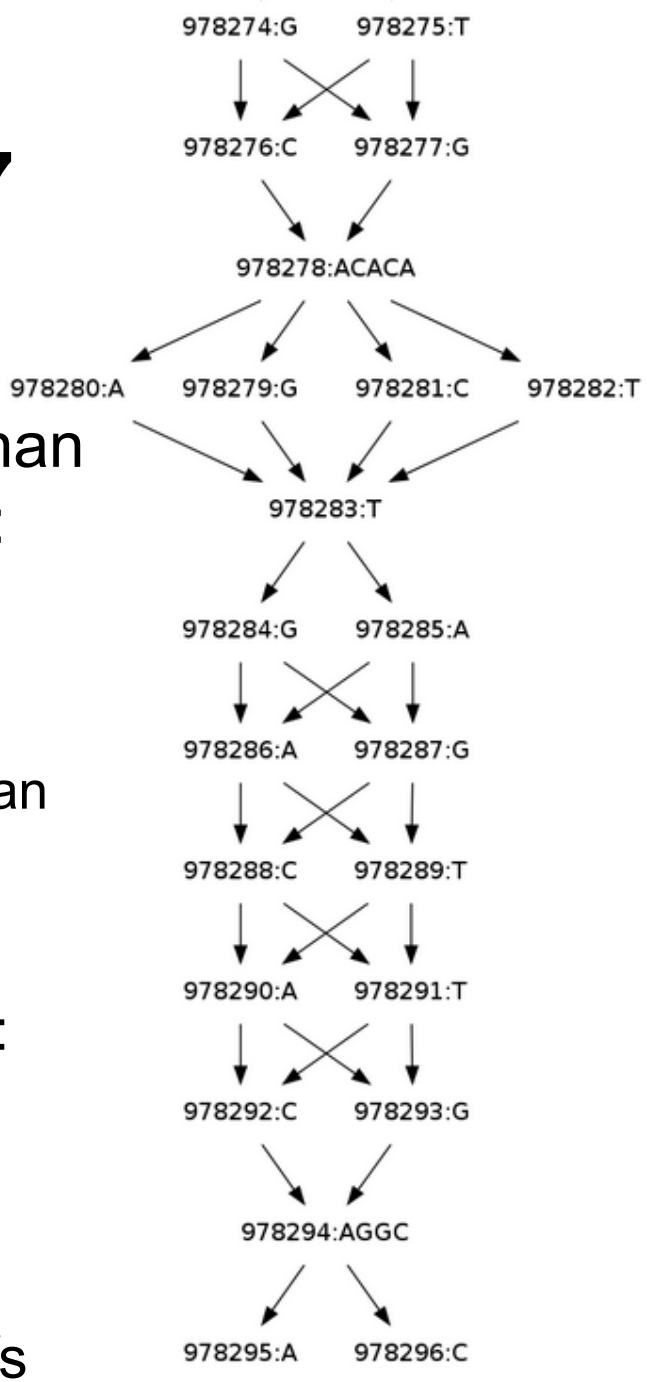
(Edge bounding limits the k-mer space, which can be very important in real data. See →)

for 27-mers crossing no more than 7 edges:

36h on a 32-core host @Sanger

Index is **80G** on disk, can be cached in memory

Can align **~1000 read/CPU/s** when on tmpfs



Aligning to 1000G+GRCh37

For each k-mer in the read

    If the k-mer is informative

        → sort hits by node id

For each cluster of hits : max(id)-min(id) < *M*

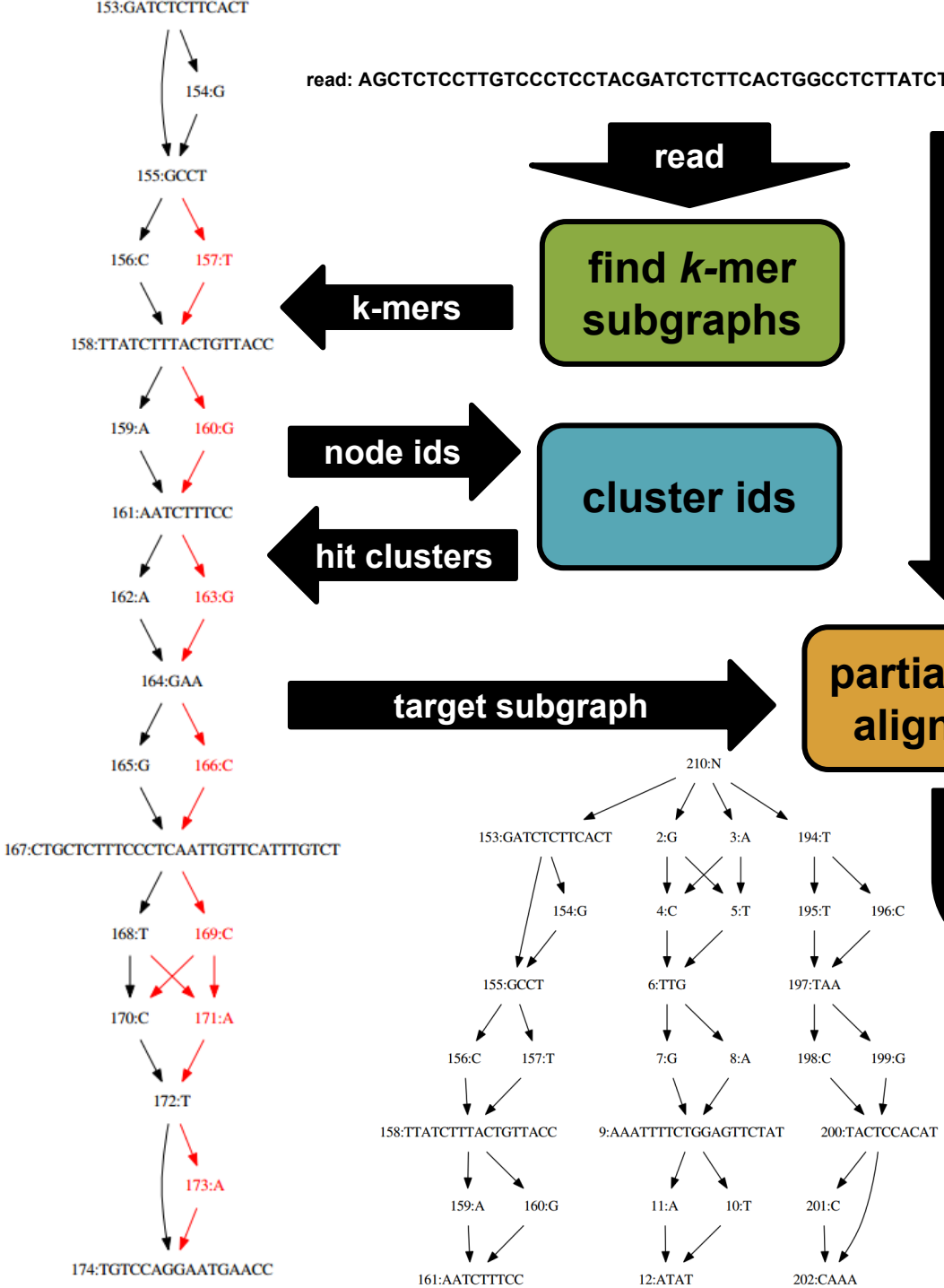
    Get local region of graph from index

    Align to the local graph using POA → alignment

If alignment failed, decrease k-mer length or stride

8000 read/s on 32-core machine @Sanger

~40 hours for a deep 150x2 Illumina X10 run



k-mer based alignment of short reads to a variation graph

Alignment

Surjection to SAM/BAM/CRAM

To immediately start working with existing tools, we can project our alignments back into the reference space.

Method simply subsets graph to only those paths in a particular reference, then forces alignment locally. see: vg surject

Thanks!

Gabor Marth
Deniz Kural
Wan-Ping Lee
Alistair Ward
Richard Durbin
Josh Randall
Benedict Patton
Zamin Iqbal
Jerome Kelleher
Jared Simpson
David Haussler
Daniel Zerbino
Jouni Siren