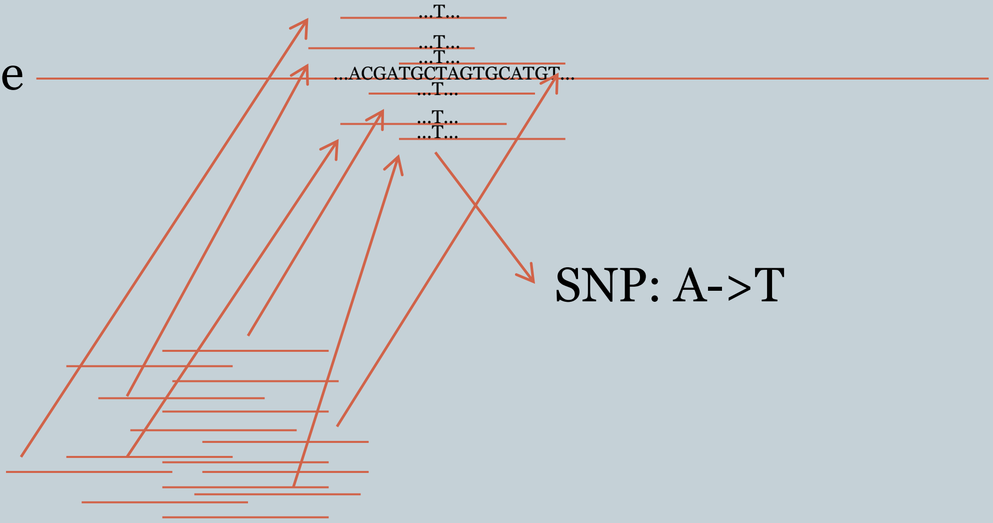
Indexing Finite Language Representation of Population Genotypes

VELI MÄKINEN
SUCCINCT DATA STRUCTURES GROUP

JOINT WORK WITH
JOUNI SIRÉN, NIKO VÄLIMÄKI, AND SERIKZHAN KAZI

Variation calling



short DNA reads aligned to same region

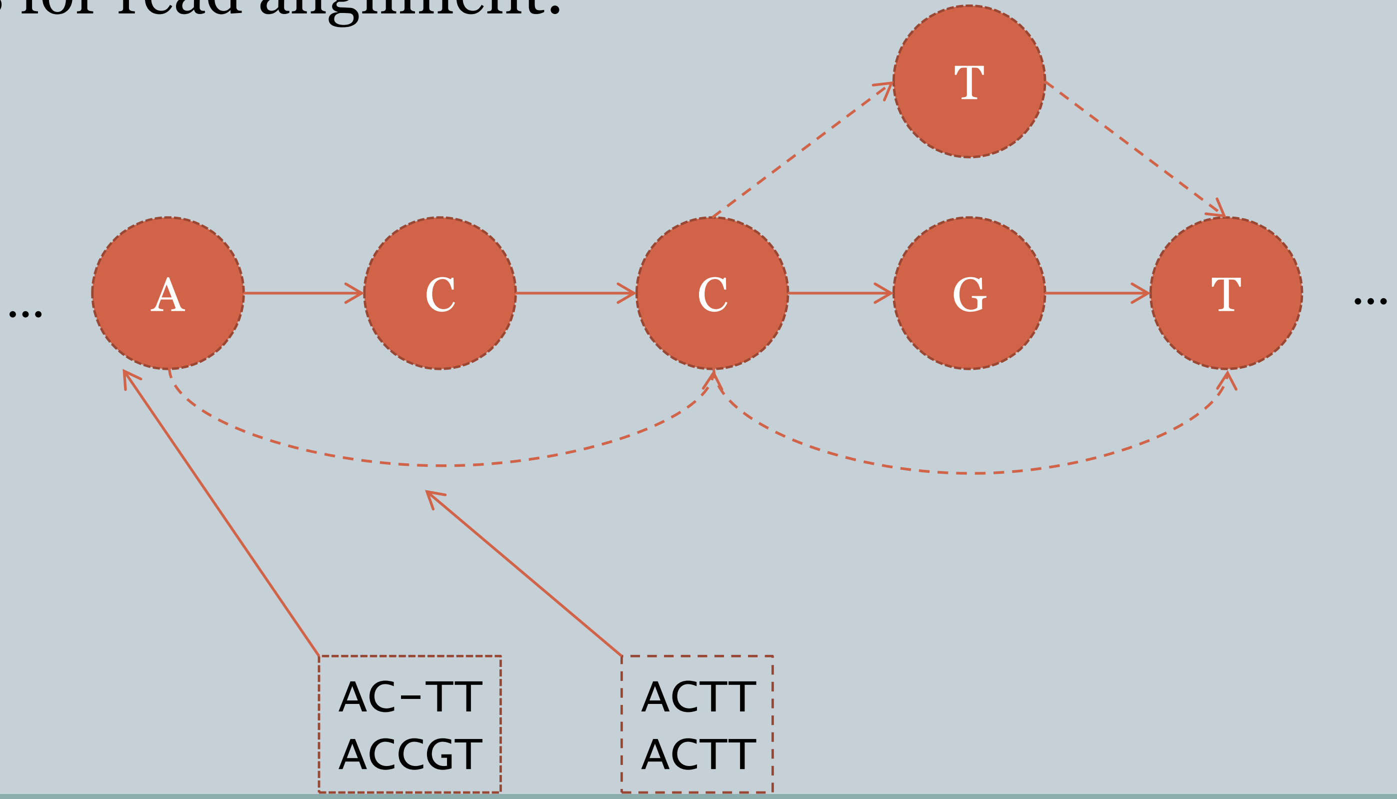
...T...
...T...
...T...
reference genome ...ACGATGCTAGTGCATGT...
...T...
...T...
...T...

SNP: A->T

millions of short DNA reads

Enhanced variation calling

- Why always only one reference is used?
- We propose to use reference + known variations as the basis for read alignment:

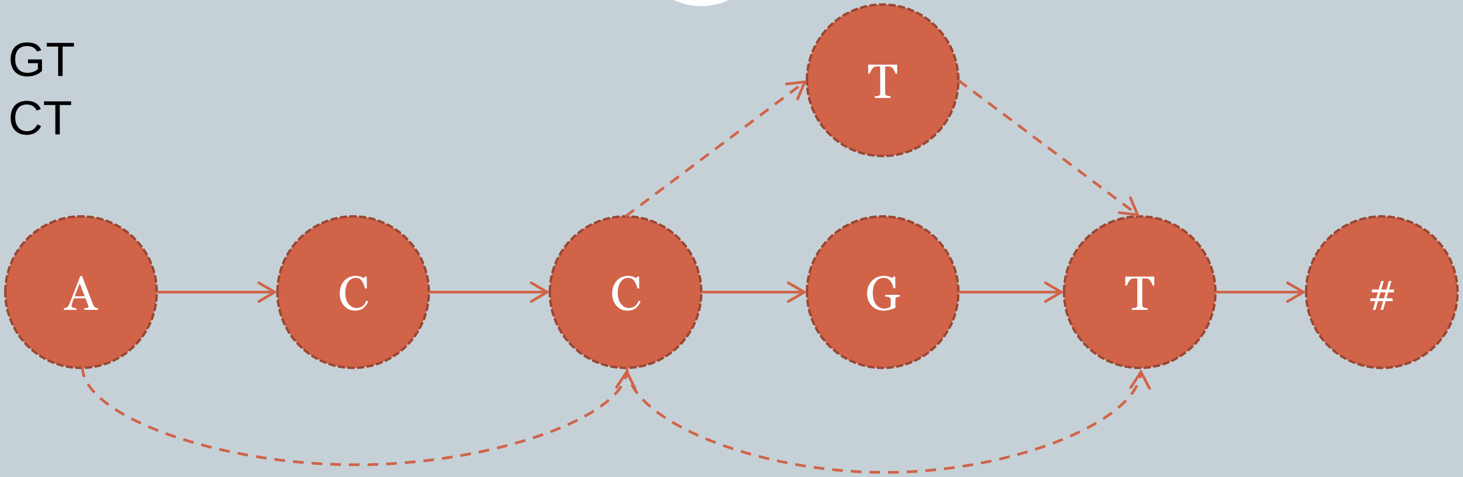


Enhanced variation calling

- Sirén, Välimäki, Mäkinen. Indexing Finite Language Representation of Population Genotypes. *WABI 2011*.
  - Generalization of Burrows-Wheeler transform for finite automata
    - Based on our work in RECOMB 2009 for multiple genomes.
  - Supports alignment of reads alike the other read aligners
    - Given a pattern P of length m, one can count the paths starting with P in O(m) time
    - Locate using the standard sampling mechanism.
    - Extends to approximate search with the general backtracking & branch-and-bound mechanism.
  - Similar space usage as for other read aligners
    - Less than 70 MB for multiple alignment of 4 assemblies of human chromosome 18 each about 76 Mbp long

Some insights

Search GT
Search CT



		GBWT	Distinguishing prefixes	Edges
		T	#	1
		#	A	2
	0 G	A	CC	1
1 G	2 C	AC	CG	1
4 C			CT	2
		C	G	1
		CGT	T#	1
		C	TT	1

All paths in lex order

#
ACCGT#
ACCTT#
ACGT#
ACTT#
CCGT#
CCTT#
CGT#
CT#
CTT#
GT#
T#
TT#

Summary

- Make finite automaton from reference + SNP data or from multiple alignment.
- Make it reverse deterministic (skipped details).
- Sort distinguishing prefixes (prefix doubling, radix sort, others?)
- Output GBWT.
- Read alignment almost identical to normal BWT read aligners.

What now?

- Index construction for human genome + SNPs requires really much RAM (terabytes)
- Summer 2011->now : Distributed construction algorithm almost ready
  - Choose p pivot prefixes, and let p machines sort their parts independently.
  - Each machine needs to access the whole automaton:
    - Compressed graph representation required.
- Aiming to release first version of the index with HG+simple common SNPs still this year.

Thanks for listening!

QUESTIONS?
COMMENTS?
NEW IDEAS?