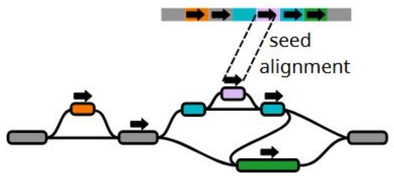
Improved Chaining in non-DAG Regions for Long Read Giraffe

UC SANTA CRUZ
Computational Genomics Lab & Platform

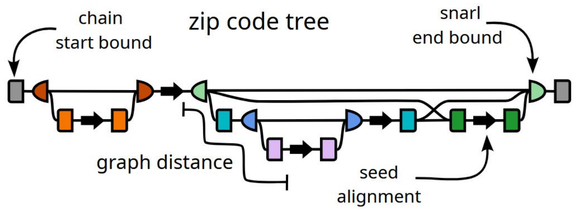
Faith Okamoto[1], Adam M. Novak[1], Xian Chang[1,2], Jouni Sirén[1], Benedict Paten[1]
[1] UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA
[2] MIAT, INRAE, Toulouse, France

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Chaining in Giraffe



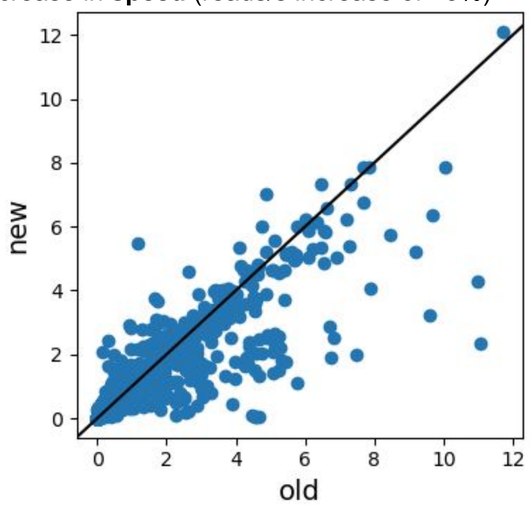
For **chaining** we must calculate distances between pairs of seeds, efficiently, in a **pangenome graph**.



We construct a **zip code tree** from the read's seeds, then traverse it to enumerate **graph distances** for transitions in between pairs of seeds.

Alignment comparisons

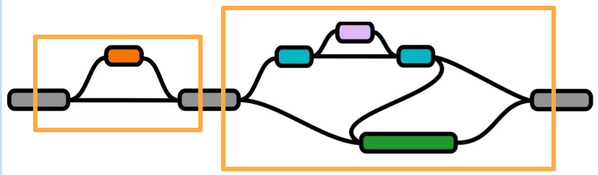
Increase in **speed** (reads/s increase of ~5%)



Comparable **accuracy**

1 million simulated reads	Old correct alignments	New correct alignments
HiFi	924760	924751
R10	925376	925386

Snarls

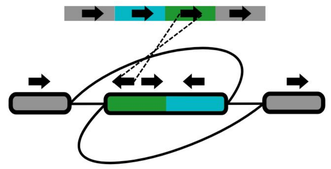


Many snarls (boxed) are **directed acyclic graphs** (DAGs) and can be put in a topological order.

What about non-DAG **cyclic snarls**?

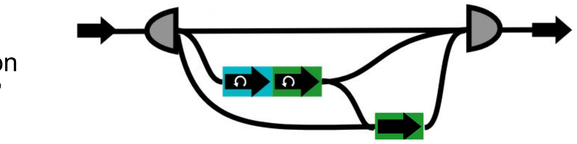
Old heuristics

The zip code tree is traversed in a **single direction**.
Therefore, cyclic snarls must be **linearized**.



Find seeds in graph

Guess orientientation and order of "runs" based on read

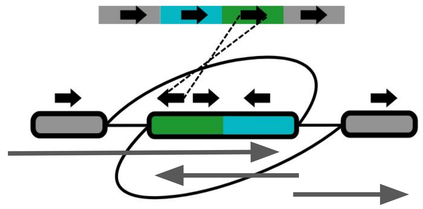


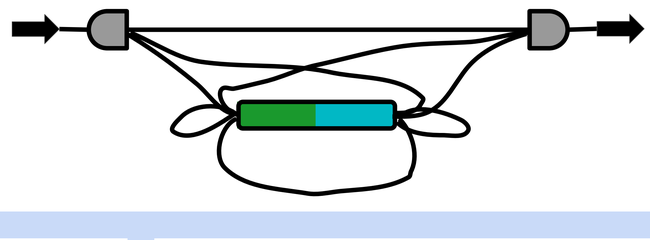
Make zip code tree with artificial children

New exact algorithm

Remove need to do linearization by **relaxing traversal rules**

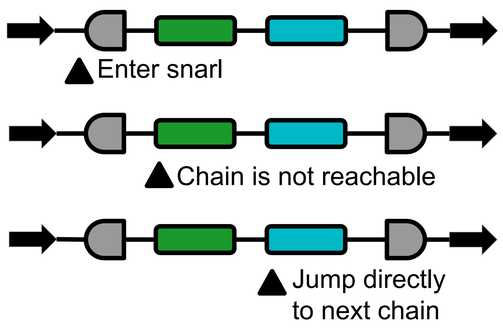
- Child chains may have **arbitrary orientation** and **order**
- Each child chain is traversed twice: **once in each direction**



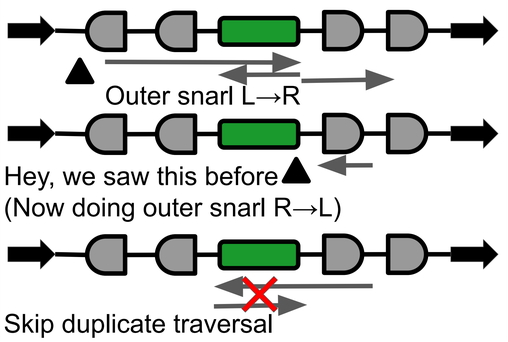


Optimization Details

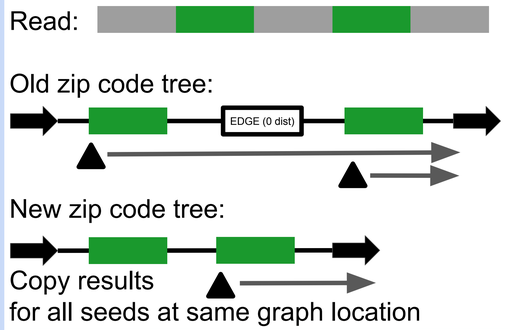
Jumping



Memorization



Duplicate minimizers

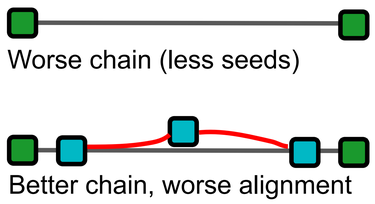


Current work/challenges

- **HiFi read mappings** worsen slightly
  - identity decreases by 10^{-6} out of 1
  - correctness decreases by 10^{-5}%
- Both HiFi and R10 have ~0.2% more **variant calling errors**

Often, errors are from finding better-looking chains which force a worse alignment.

May need to tweak mapping parameters or strategy.



Acknowledgments

Thanks to vg team for support! Also, many figures are taken from Chang *et al.* 2025 (bioXiv), of which I am a co-author.