PanGeA: Pan-Genome Annotation
Indexing Annotated Human Genome Collections

Andre Brehme, Sven Brümmer, Jonas Charfreitag, Jonas Ellert,
Jannik Junghänel, Dominik Köppl, Christopher Osthues, Sven Rahmann, Dennis Rohde,
Julian Sauer, Jonas Schmidt, Lars Schäpers, Uriel Elias Wiebelitz, Jens Zentgraf

Abstract

Motivation

- Novel high-throughput sequencing methods
  - ➡ Huge sets of genomes with annotations
  - ➡ Opportunities for personalized genome-based medicine
- 1 genome ≤ 1GB (1 base pair ≙ 2 bits)
  - ➡ Pangenome of Dortmund (≈500.000 inhabitants) ≙ ≈500TB
  - ➡ Datasets too huge to maintain in commodity computers efficiently
- Genomes of two individuals are up to 99% *identical* [FCS06]
  - ➡ Opportunities for compression

Goal

Build an application, that

- maintains a pangenome
- answers queries like (approximate) pattern matching efficiently on the maintained pangenome

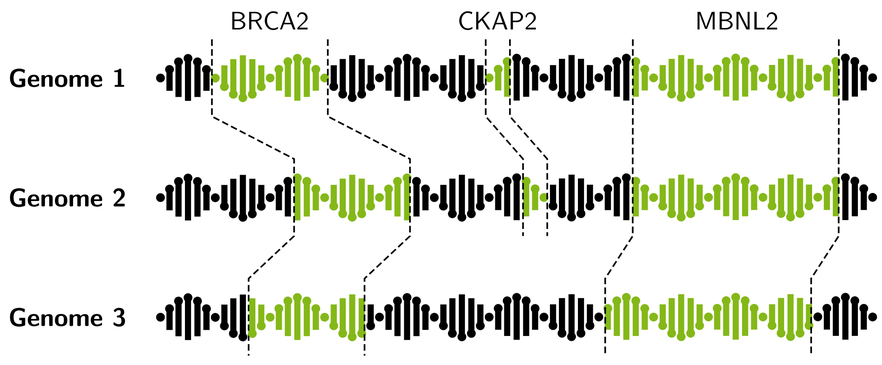
Implemented Data Structures

(a) Colored De Bruijn Graph [BBB+17] as a prefilter
(b) Journaled String Tree [RWR14]
(c) CHICO [Val16]
(d) Relative Lempel-Ziv [KPZ11]

Future Work

- Implement other approaches like grammar-based self-indexes [CN12]
- Use annotations for annotation based queries or more efficient query answering

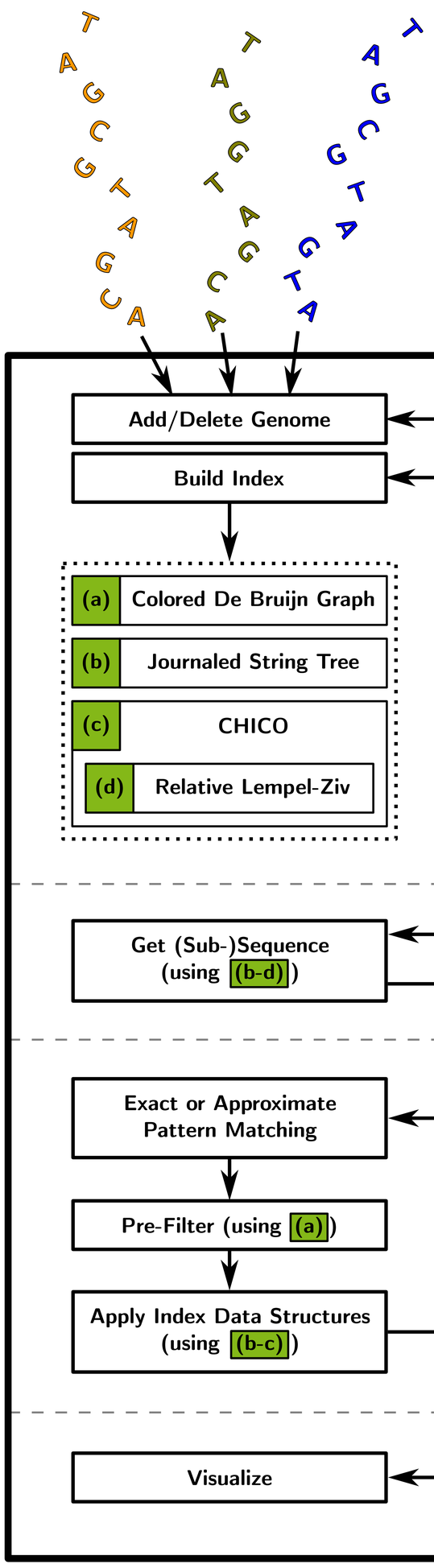
Annotations



Annotated area (for specific gene) is very similar among all genomes

- Highly compressible areas can be found with help of annotations
- Useful to fix positions for sequence alignment

```
Position:    1 2 3 4 5 6 7 8 9 10
Genome s1:   T A G C G T A G C A
Genome s2:   T A G G T A G C A
Genome s3:   T A G C G T A G T A
```



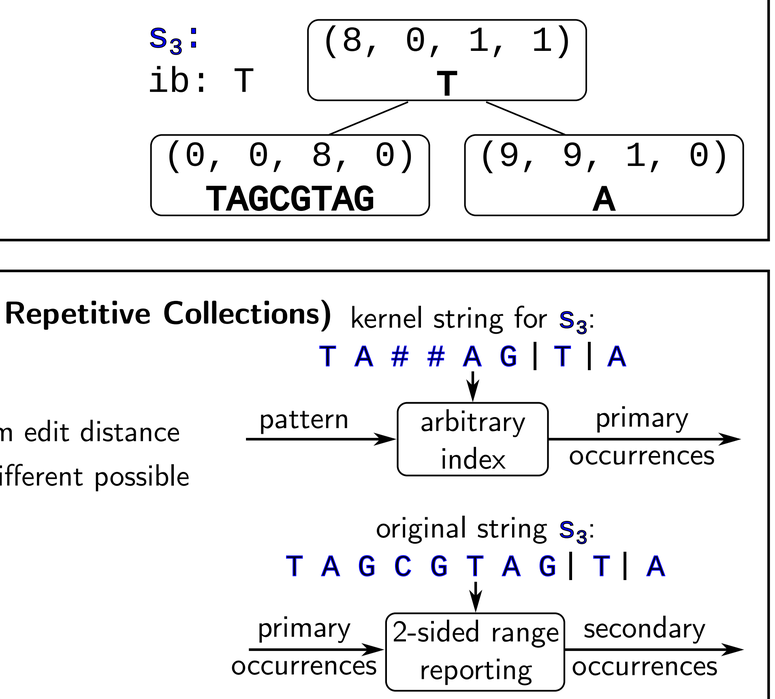
PanGeA Index

Index Features Usage

1. **Add/Delete genomes**
2. **Build implemented data structures (a) - (d)**
   Existing structures are extended dynamically
3. **Retrieve genomes** or subsequences of genomes by ID and position or name of annotated area
4. **Pattern Matching**
   Return a list of all occurrences of the given pattern in all genomes
5. **Visualize** the genome collection

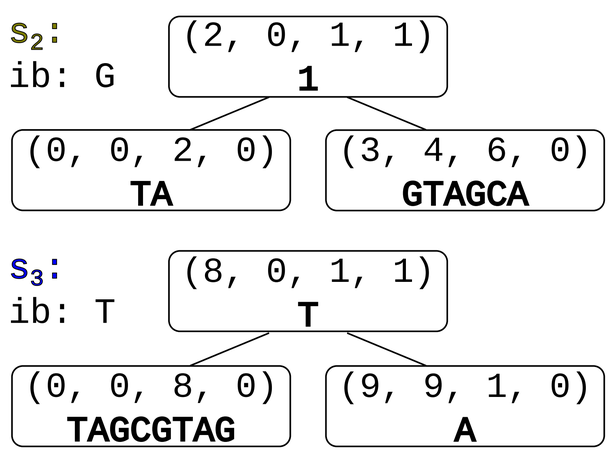
(a) Colored De Bruijn Graph:

- Edges represent *k+1*-mers: AGC → GCG
  - ➡ AGCG occurs in s_1 and s_3
- May return false positives due to circles
  - ➡ Use DBG as a pre-filter
- Supports approximate pattern matching
- Succinct representation allows memory efficient implementation [BOSS12]



(b) Journaled String Trees

- Represent each genome as an order of copy operations that transform the reference genome into the other genome
- Save an insertion buffer for each genome and a tuple (*vp*, *pp*, *l*, *t*) for each copy operation
- Save tuples in balanced binary search tree

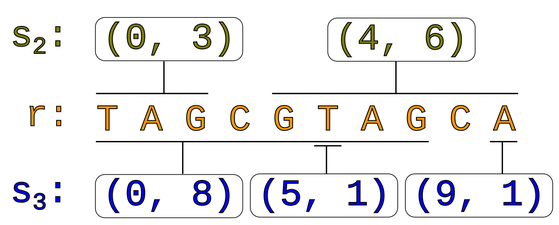


(c) CHICO (Compressed Hybrid Index for Repetitive Collections)

- Exact and approximate pattern matching
- Predefined maximum pattern length and maximum edit distance
- Utilizes RLZ-parsing to distinguish between two different possible match types
  - Primary occurrence:
    The pattern spans multiple RLZ-phrases
  - Secondary occurrence:
    The pattern is contained in a single RLZ-phrase

(d) Relative Lempel-Ziv

- User chooses reference genome `r` (here `r` = s_1)
- Each genome is interpreted as a composition of reference genome substrings
- Each substring is stored as a (start position, length)-tuple
- Retrieval of each genome in linear time to the genome length



[BBB+17] Keith Belk, Christina Boucher, Alexander Bowe, Travis Gagie, Paul Morley, Martin D Muggli, Noelle R Noyes, Simon J Puglisi, and Rober Raymond. Succinct Colored de Bruijn Graphs. *Bioinformatics*, to appear, 2017.
[BOSS12] Alexander Bowe, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Succinct de Bruijn Graphs. *Proceedings of WABI '12*, pages 225-235, 2012.
[CN12] Francisco Claude and Gonzalo Navarro. Improved Grammar-Based Compressed Indexes. *Proceedings of SPIRE '12*, pages 180-192, 2012.
[FCS06] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nature Reviews - Genetics*, 7(2):85-97, 2006.
[KPZ11] Shanika Kuruppu, Simon J. Puglisi, and Justin Zobel. Relative Lempel-Ziv Compression of Genomes for Large-scale Storage and Retrieval. *Proceedings of SPIRE '11*, pages 201-206, 2011.
[RWR14] Rene Rahn, David Weese, and Knut Reinert. Journaled string tree scalable data structure for analyzing thousands of similar genomes on your laptop. *Bioinformatics*, 30(24):3499-3505, 2014.
[Val16] Daniel Valenzuela. CHICO: A Compressed Hybrid Index for Repetitive Collections. *Proceedings of SEA '16*, pages 326-338, 2016.

technische universität dortmund
aido
Chair 11:
Algorithm Engineering
Project Group 605