Sequence-Based Pangenomic Core Detection

(Work in progress...)

Tizian Schulz, Roland Wittler, Jens Stoye

Faculty of Technology, Bielefeld University

February 11, 2021

Pangenomic Diversity

Nowadays, many individual genomes per species are available.

Pangenomic Diversity

Nowadays, many individual genomes per species are available.

Genomic comparison reveals large differences between them

Pangenomic Diversity

Nowadays, many individual genomes per species are available.

Genomic comparison reveals large differences between them

For bacterial species: Even different gene content

Pangenomic Diversity

Nowadays, many individual genomes per species are available.

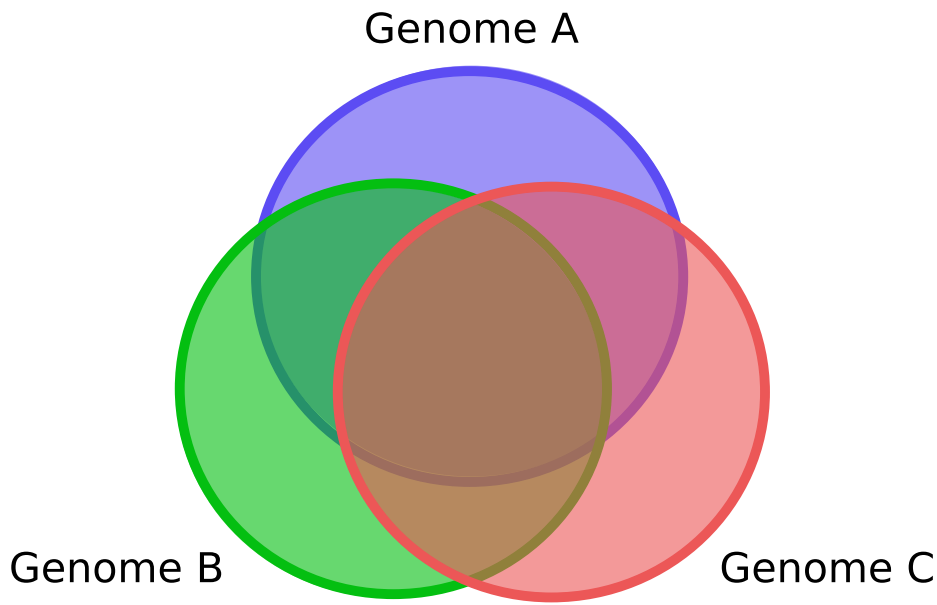
Genomic comparison reveals large differences between them

For bacterial species: Even different gene content

What defines a (bacterial) species on a genomic level?

Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

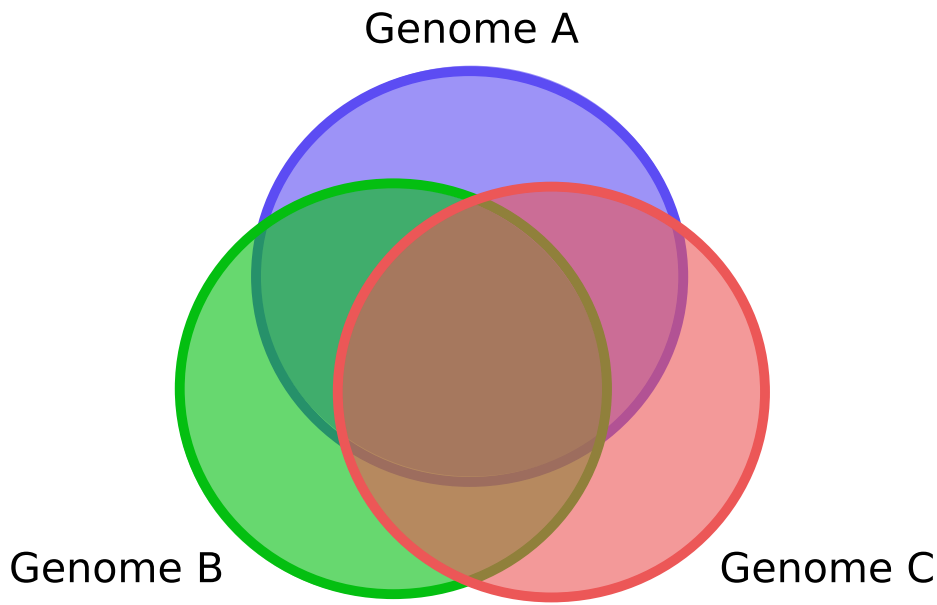


Adapted from Guillaume Holley

Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$

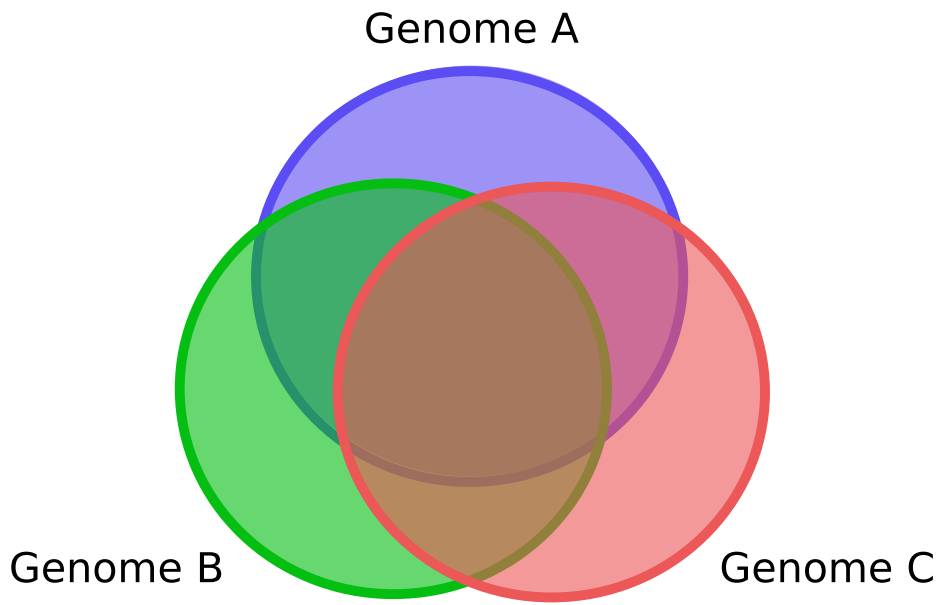


Adapted from Guillaume Holley

Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$



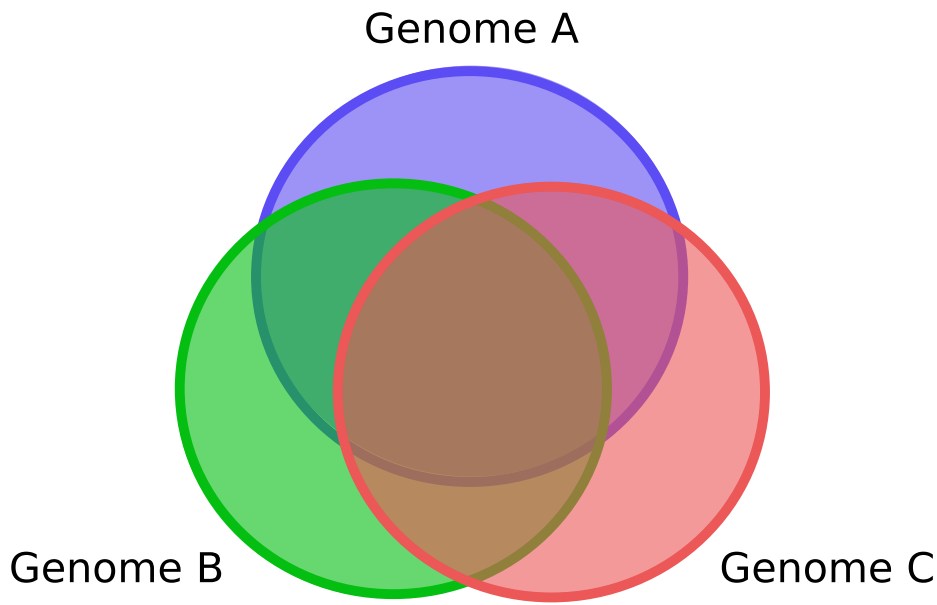
Adapted from Guillaume Holley

Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for



Adapted from Guillaume Holley

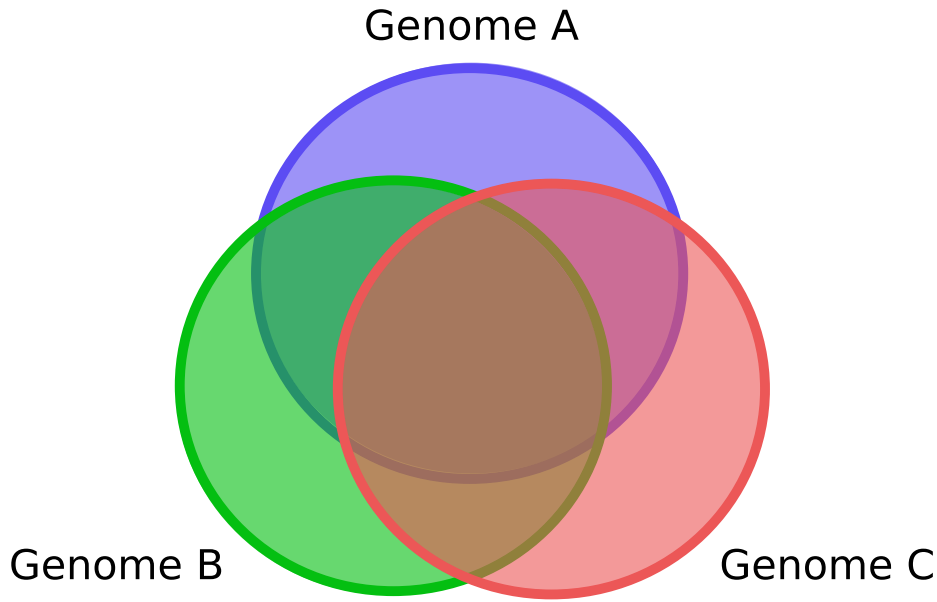
Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for

- studying genetic diversity



Adapted from Guillaume Holley

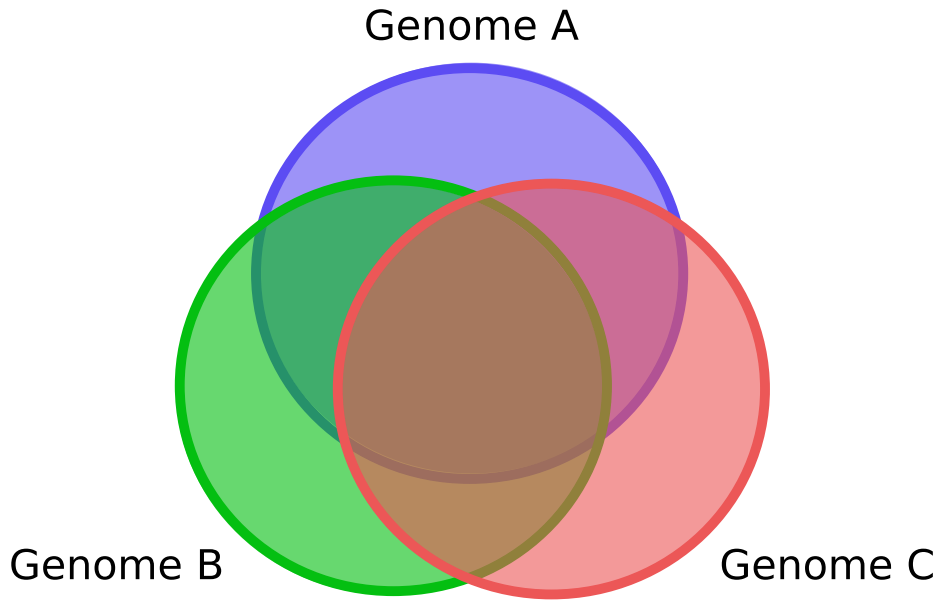
Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for

- studying genetic diversity
- medical research:



Adapted from Guillaume Holley

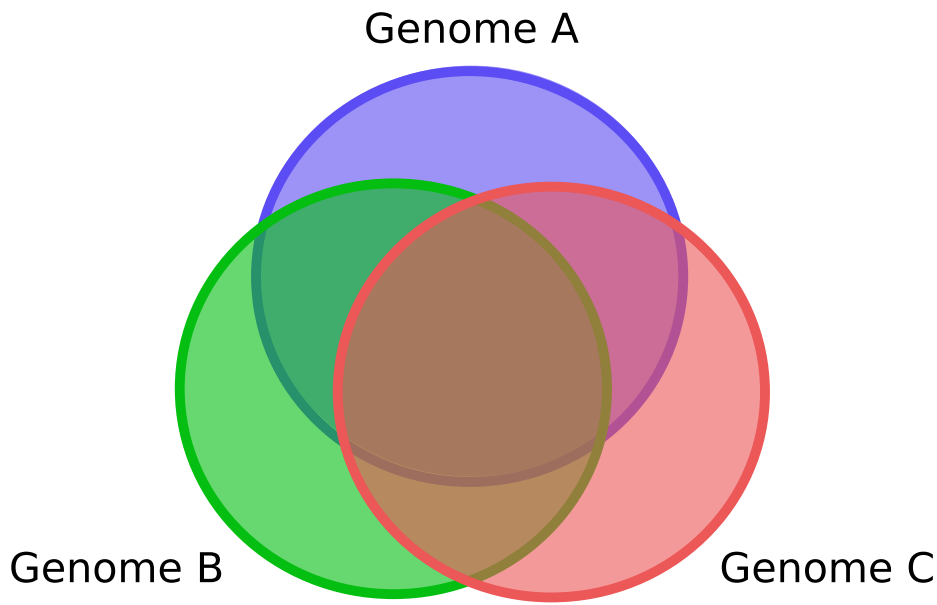
Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for

- studying genetic diversity
- medical research:
  - drug development



Adapted from Guillaume Holley

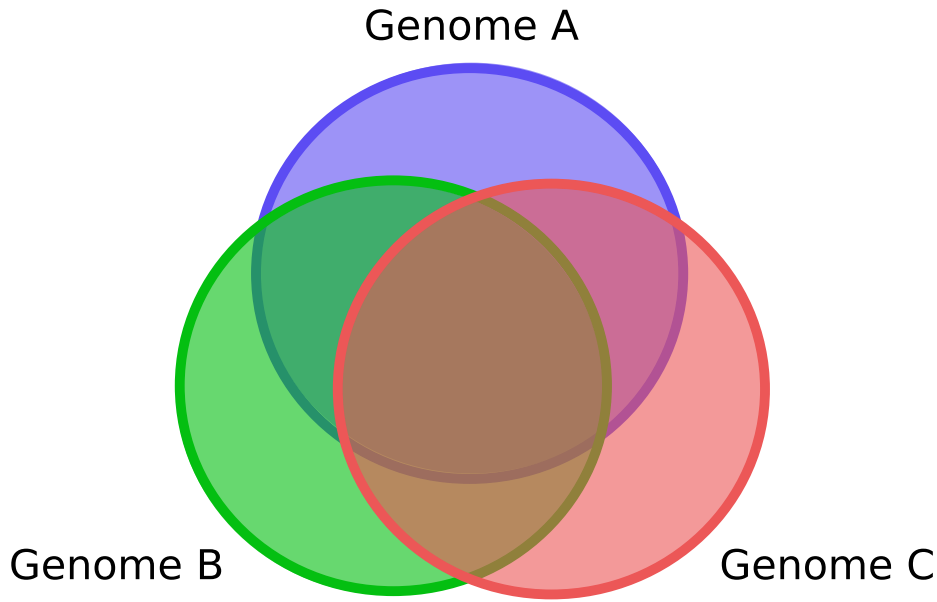
Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for

- studying genetic diversity
- medical research:
  - drug development
  - vaccine design



Adapted from Guillaume Holley

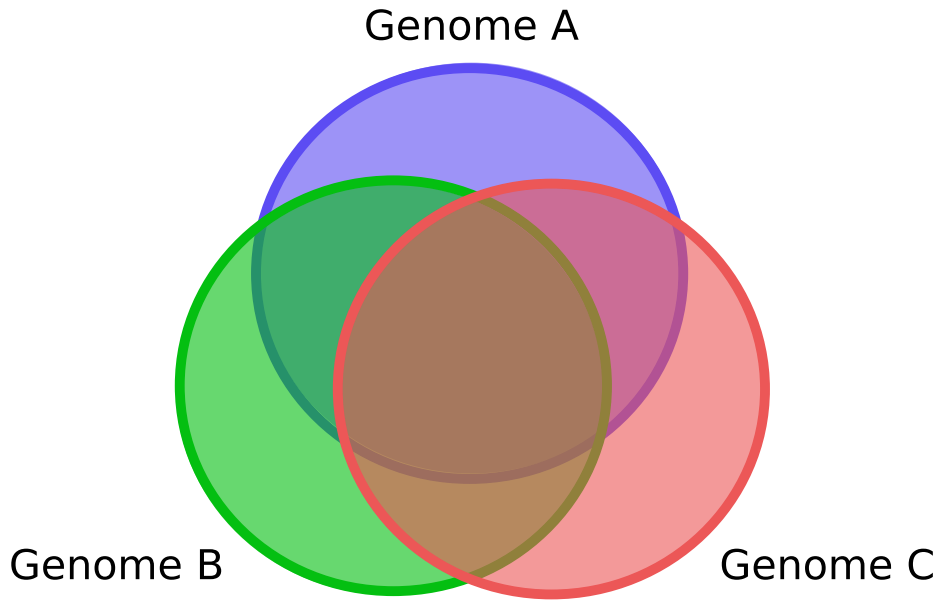
Classical Pangenomics

Let a *genome* be a set of strings over the alphabet $\Sigma_{\text{DNA}} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

- *Gene based pangenome*: All *genes* of genomes $g_1, \ldots, g_n$
- *Core genome*: All common genes of $g_1, \ldots, g_n$

Considering the pangenomic core is important for

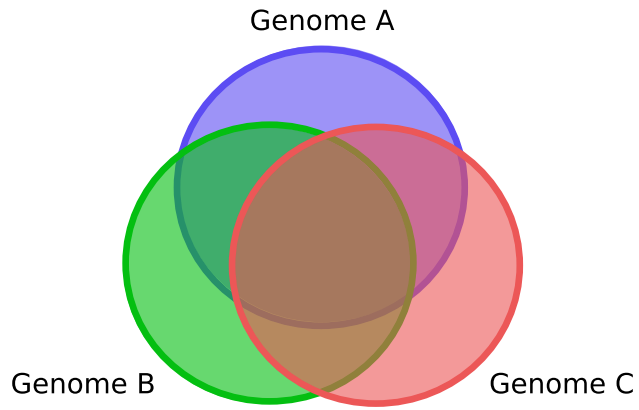
- studying genetic diversity
- medical research:
  - drug development
  - vaccine design
- crop plant breeding



Adapted from Guillaume Holley

Limitations of Gene Based Approaches

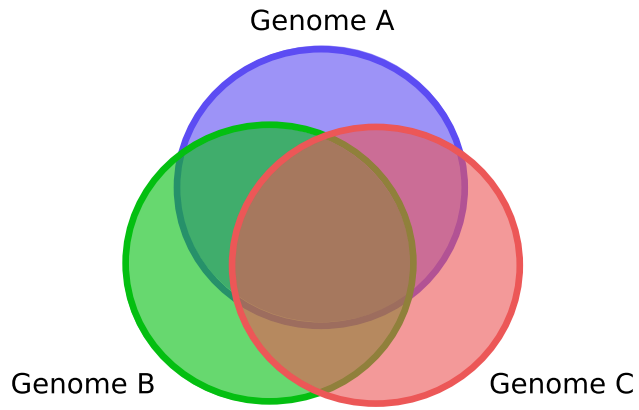
Drawbacks of gene based pangenomic approaches:



Limitations of Gene Based Approaches

Drawbacks of gene based pangenomic approaches:

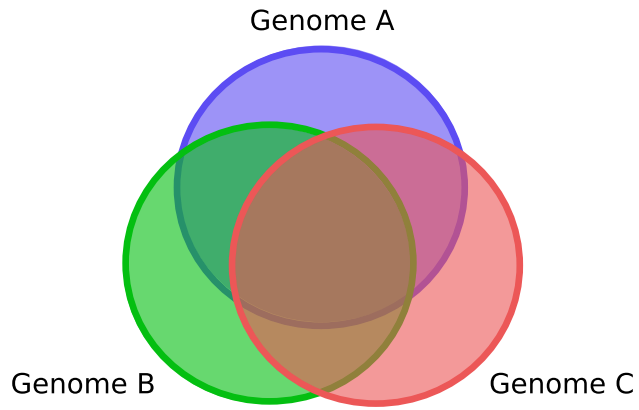
- expensive preprocessing needed (assembly, gene annotation)



Limitations of Gene Based Approaches

Drawbacks of gene based pangenomic approaches:

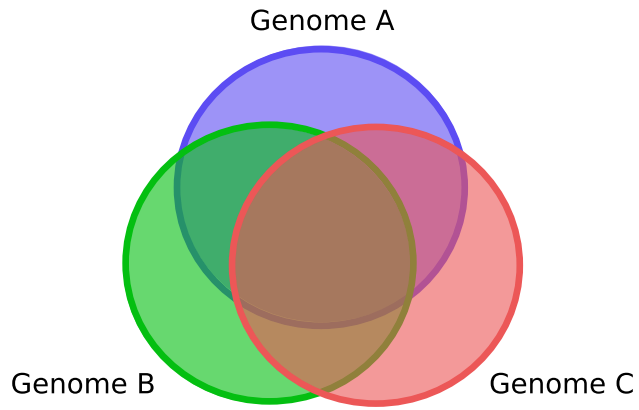
- expensive preprocessing needed (assembly, gene annotation)
- genes from different genomes need to be mapped



Limitations of Gene Based Approaches

Drawbacks of gene based pangenomic approaches:

- expensive preprocessing needed (assembly, gene annotation)
- genes from different genomes need to be mapped
- core features below (and above) gene level remain undiscovered



Sequence Based Pangenomics

Definition Sequence Based Pangenome

The *sequence based pangenome* is a set of genomes $g_1, \ldots, g_n$ of some taxonomic unit t.

Sequence Based Pangenomics

Definition Sequence Based Pangenome

The *sequence based pangenome* is a set of genomes $g_1, \ldots, g_n$ of some taxonomic unit t.

Realization: ?

Sequence Based Pangenomics

Definition Sequence Based Pangenome

The *sequence based pangenome* is a set of genomes $g_1, \ldots, g_n$ of some taxonomic unit t.

Realization: $\rightarrow$ e.g. colored de Bruijn graph (C-DBG)

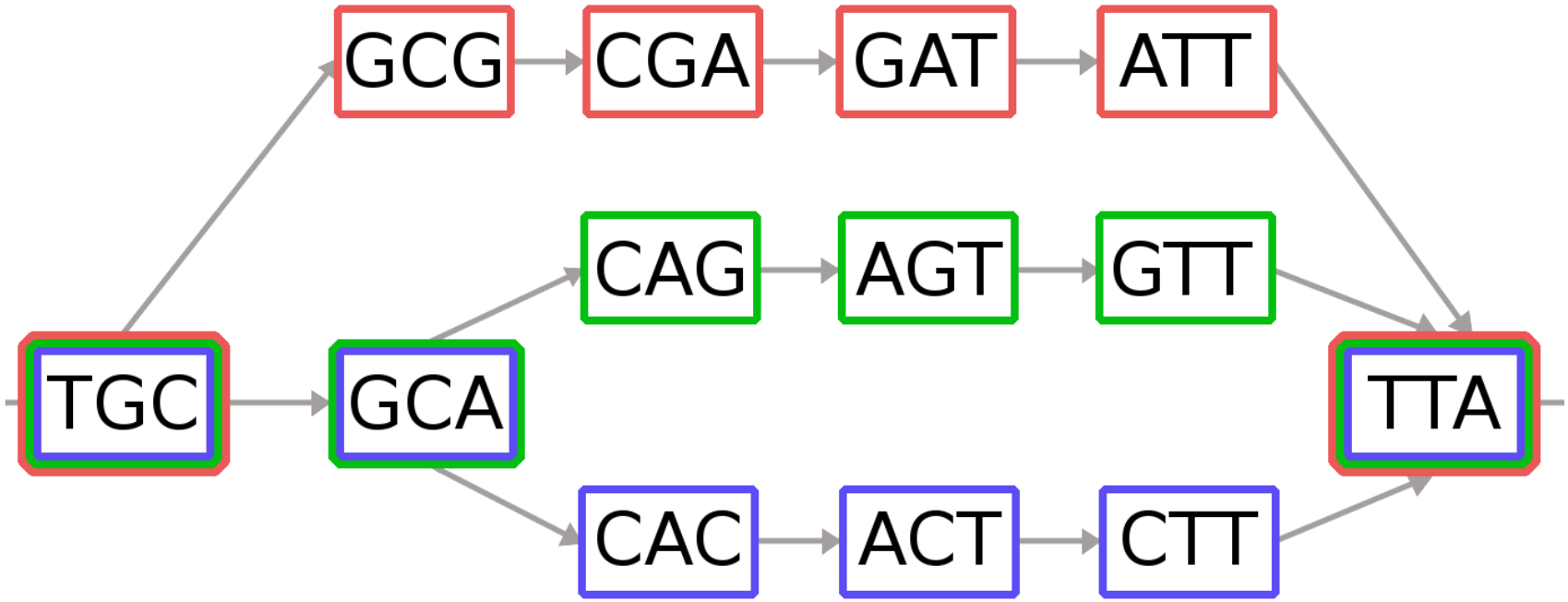


Image courtesy of Guillaume Holley

The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

Goal: Generalization of gene based core

The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

Goal: Generalization of gene based core

Idea: Take all nodes (k-mers) shared between all input genomes

The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

Goal: Generalization of gene based core

Idea: Take all nodes (k-mers) shared between all input genomes

Problem:

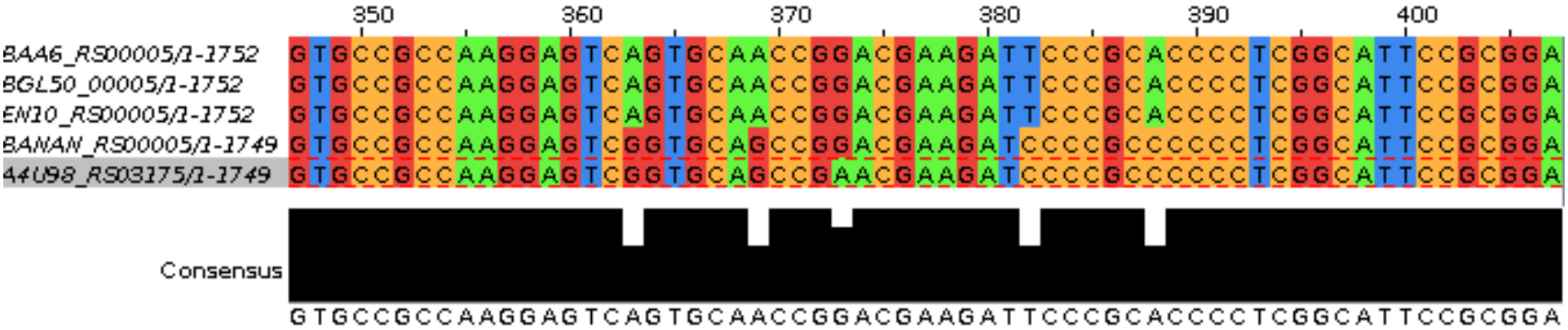
The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

Goal: Generalization of gene based core

Idea: Take all nodes (k-mers) shared between all input genomes

Problem:



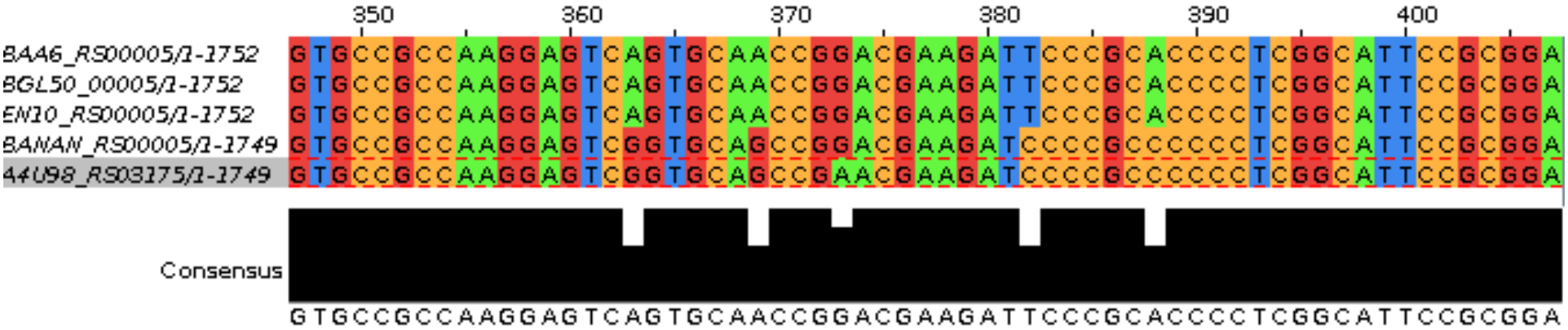
The Sequence Based Core

How to define the core of a sequence based, graphical pangenome?

Goal: Generalization of gene based core

Idea: Take all nodes (k-mers) shared between all input genomes

Problem:

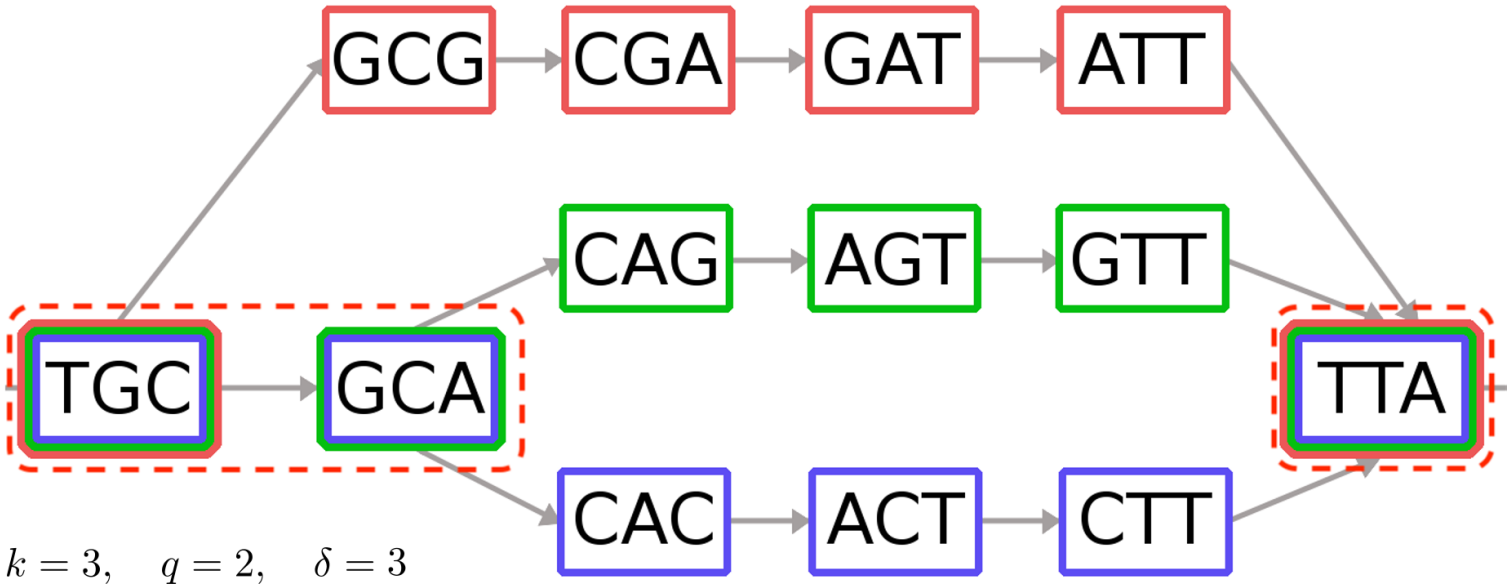


Variations occur frequently even in core sequences!

Our Core Definition

Definition Core k-mer

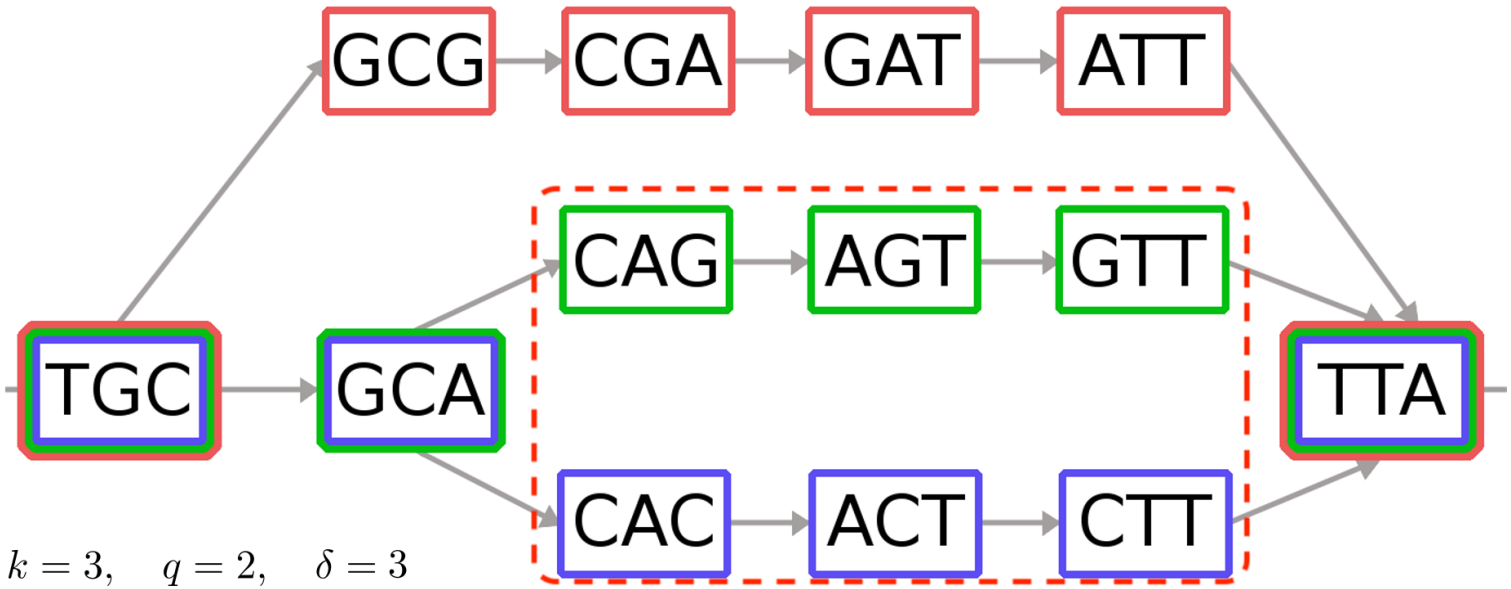
Let $G = (V, E, C)$ be a C-DBG representing a pangenome $p = \{g_1, g_2, \ldots, g_n\}$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. A k-mer $v \in V$ is called *core k-mer* if and only if $|C(v)| \geq q$



Our Core Definition

Definition Bridging *k*-mer

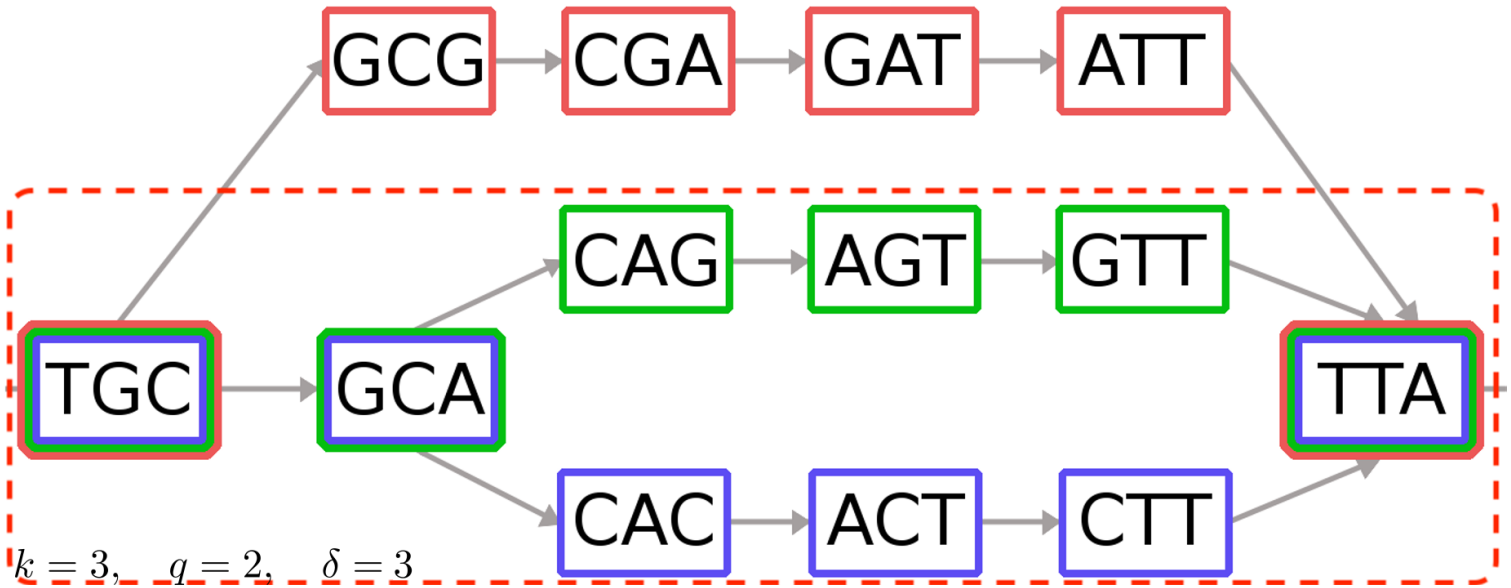
A *k*-mer $v \in V$ is called *bridging k-mer* if and only if it lies on a path π connecting two core *k*-mers and $|\pi| \leq \delta + 2$



Our Core Definition

Definition Core Genome

The *core genome* of p is defined as the set of all core and bridging k-mers of G.



Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Advantages:

Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Advantages:

- no assembly

Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Advantages:

- no assembly
- no annotations

Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Advantages:

- no assembly
- no annotations
- no gene mapping

Problem statement

Core Detection Problem

Given a pangenome $p = \{g_1, g_2, \ldots, g_n\}$ represented as a C-DGB of dimension $k \geq 1$. Let $q \in [1, n]$ and $\delta \geq 0$ be two integers. The *Core Detection Problem* is to find the core genome of p.

Advantages:

- no assembly
- no annotations
- no gene mapping

How can we find the core?

Algorithm

Algorithm

Idea:

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Idea: Use *compacted* C-DBG

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Idea: Use *compacted* C-DBG

- k-mers connected by simple path are merged

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Idea: Use *compacted* C-DBG

- k-mers connected by simple path are merged
- BFS only needed at the end of simple paths ($\leq 2 \cdot$#unitigs)

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Idea: Use *compacted* C-DBG

- k-mers connected by simple path are merged
- BFS only needed at the end of simple paths ($\leq 2 \cdot$#unitigs)

Further speed gain: Make use of information collected during past BFSs

Algorithm

Idea:

- for each core k-mer explore graph by BFS of depth δ
- if core k-mer found, mark all k-mers on path as *bridging*

$\rightarrow$ Problem: Search space grows exponentially in δ if graph is complex

Idea:

- start BFS in the middle of a path connecting core k-mers
- requires only 2 BFSs of depth $\frac{\delta}{2}$

$\rightarrow$ But: Search needed for every **non**-core k-mer

Idea: Use *compacted* C-DBG

- k-mers connected by simple path are merged
- BFS only needed at the end of simple paths ($\leq 2 \cdot$#unitigs)

Further speed gain: Make use of information collected during past BFSs

Implementation (based on Bifrost) is called *Corer*

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Parameters:

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Parameters:

- $k \in \{21, 31\}$

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Parameters:

- $k \in \{21, 31\}$
- $q = n$ (100%)

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Parameters:

- $k \in \{21, 31\}$
- $q = n$ (100%)
- $\delta \in \{0, 40, 100, 300\}$

Evaluation – setting

Evaluation on pangenome data sets of different sizes:

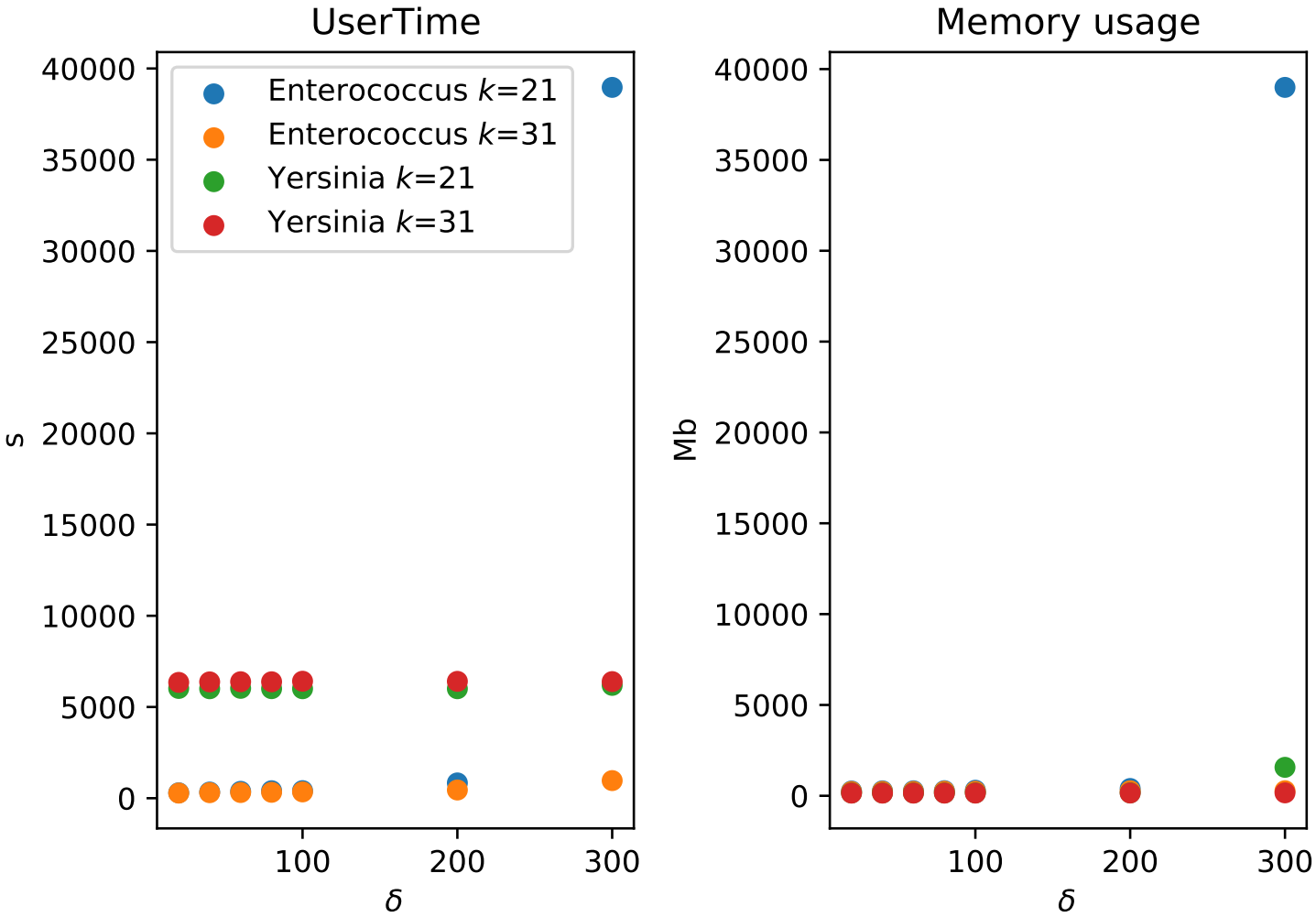
- Yersinia pestis ($n = 48$)
- Enterococcus faecium ($n = 153$)

Parameters:

- $k \in \{21, 31\}$
- $q = n$ (100%)
- $\delta \in \{0, 40, 100, 300\}$

Experiments performed single-threaded on compute cluster

Runtime and memory usage



Comparison to gene based approach

Yersinia Pestis n=48

Comparison to gene based approach

Enterococcus Faecium n=153

Comparison to sequence based approaches

Comparison planned to

Comparison to sequence based approaches

Comparison planned to

- Panseq (Laing *et al.* 2010)

Comparison to sequence based approaches

Comparison planned to

- Panseq (Laing *et al.* 2010)
- Sibelia (Minkin *et al.* 2013)

Comparison to sequence based approaches

Comparison planned to

- Panseq (Laing *et al.* 2010)
- Sibelia (Minkin *et al.* 2013)
- SplitMEM (Marcus *et al.* 2014)

Comparison to sequence based approaches

Comparison planned to

- Panseq (Laing *et al.* 2010)
- Sibelia (Minkin *et al.* 2013)
- SplitMEM (Marcus *et al.* 2014)
- iMGE (Wang *et al.* 2014)

Comparison to sequence based approaches

Comparison planned to

- Panseq (Laing *et al.* 2010)
- Sibelia (Minkin *et al.* 2013)
- SplitMEM (Marcus *et al.* 2014)
- iMGE (Wang *et al.* 2014)
- ...

Outlook

Outlook

Directions of future work:

Outlook

Directions of future work:

- generalization for accessory genome detection

Outlook

Directions of future work:

- generalization for accessory genome detection
- hierarchical core genome representation

Outlook

Directions of future work:

- generalization for accessory genome detection
- hierarchical core genome representation
- generation of a “reference core genome”

End

Thank you for your attention!