ETH zürich

Swiss National
Science Foundation

An introduction to pangenomics

Alexander Leonard[1]

[1]ETH Zürich

2025/01/23

Caveat emptor

Pangenomics is a *rapidly evolving* and *poorly defined* field, this is just a taster

Caveat emptor

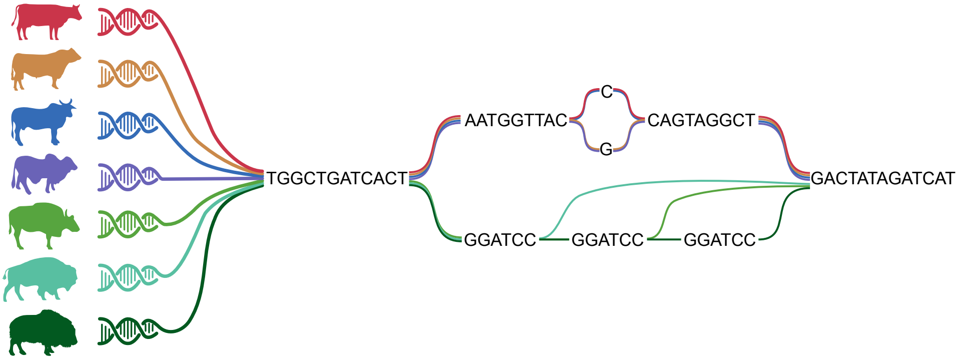
Pangenomics is a *rapidly evolving* and *poorly defined* field, this is just a taster

This also focuses on “**sequence/variation graph**” pangenomics, but there are many other types out there!

Caveat emptor

Pangenomics is a *rapidly evolving* and *poorly defined* field, this is just a taster

This also focuses on "**sequence/variation graph**" pangenomics, but there are many other types out there!



Overview

1. Introduction to pangenomics
   - Terminology
   - Graph building
   - Pangenome visualisation
2. Working with pangenomes
   - Pangenome communities
   - Pangenome validation
   - Downstream pangenomics
3. Pangenomics of the future
   - Personalised pangenomes
   - Targeted pangenomes
   - Pangenomes = biology?

What is a genome?

Encode *one* layer of information for an individual organism

Sequence of $\sim$ 1,000,000,000 nucleotides [ACTG] split into chromosomes

What is a reference genome?

Definition of a reference genome:

A reference sequence is an accepted representation that is used by researchers as a standard for comparison to DNA sequences generated in their studies.

What is a reference genome?

Definition of a reference genome:

> *A reference sequence is an accepted representation that is used by researchers as a standard for comparison to DNA sequences generated in their studies.*

We use the **same** reference genome for these **different** cows?

Routine genome assembly

Long read sequencing has *almost* solved genome assembly

Solving a puzzle is easier with larger pieces

Jarvis, E.D., Formenti, G., Rhie, A. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022). https://doi.org/10.1038/s41586-022-05325-5

Routine genome assembly

Long read sequencing has *almost* solved genome assembly

Solving a puzzle is easier with larger pieces

Jarvis, E.D., Formenti, G., Rhie, A. et al. Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022). https://doi.org/10.1038/s41586-022-05325-5

Much faster **and** much cheaper **and** much easier today

What is a **pan**genome?

Almost no consensus of what a pangenome *is*

What is a **pan**genome?

Almost no consensus of what a pangenome *is*

- a reference genome with a vcf
- a set of genome assemblies
- a list of haplotypes

What is a **pan**genome?

Almost no consensus of what a pangenome *is*

- a reference genome with a vcf
- a set of genome assemblies
- a list of haplotypes
- **a graph structure representing variation across multiple assemblies**

What is a **pan**genome?

How can we integrate information from many assemblies into one structure?

What is a **pan**genome?

How can we integrate information from many assemblies into one structure?

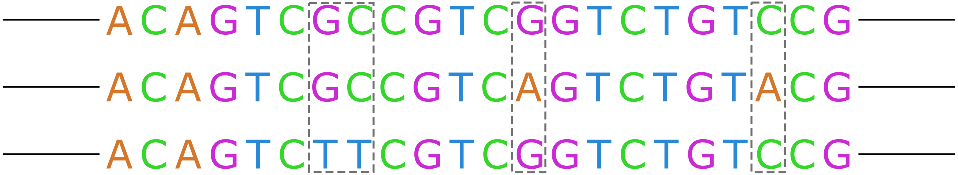
ACAGTCGCCGTCGGTCTGTCCG

ACAGTCGCCGTCAGTCTGTACG

ACAGTCTTCGTCGGTCTGTCCG

What is a **pan**genome?

How can we integrate information from many assemblies into one structure?



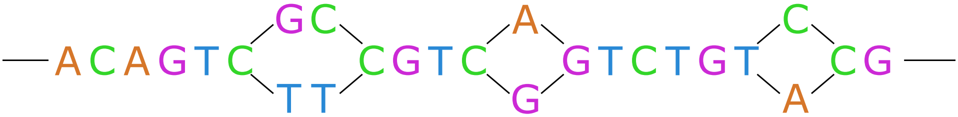
ACAGTCGCCGTCGGTCTGTCCG

ACAGTCGCCGTCAGTCTGTACG

ACAGTCTTCGTCGGTCTGTCCG

What is a **pan**genome?

How can we integrate information from many assemblies into one structure?

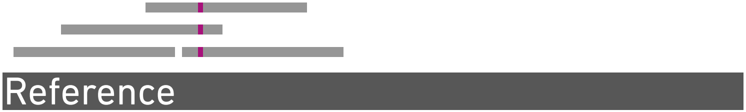


What is reference bias?

Why do we even *want* pangenomes to replace reference genomes?

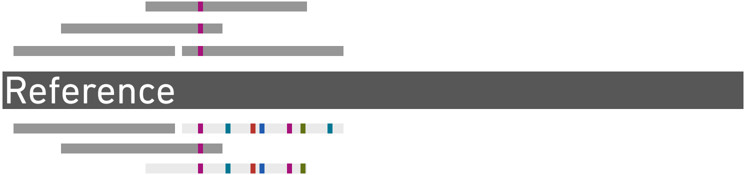
What is reference bias?

Why do we even *want* pangenomes to replace reference genomes?



What is reference bias?

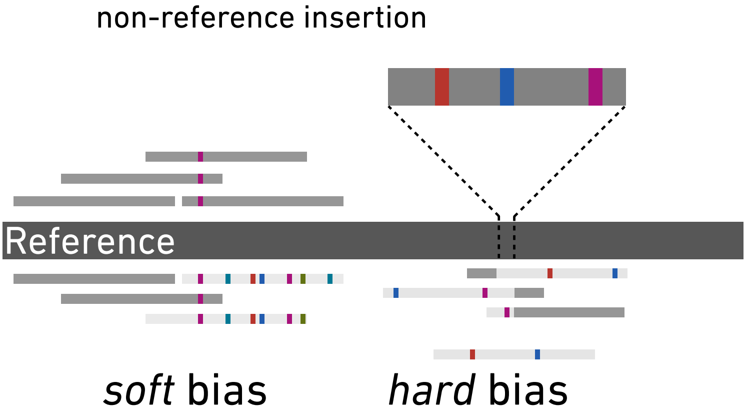
Why do we even *want* pangenomes to replace reference genomes?



soft bias

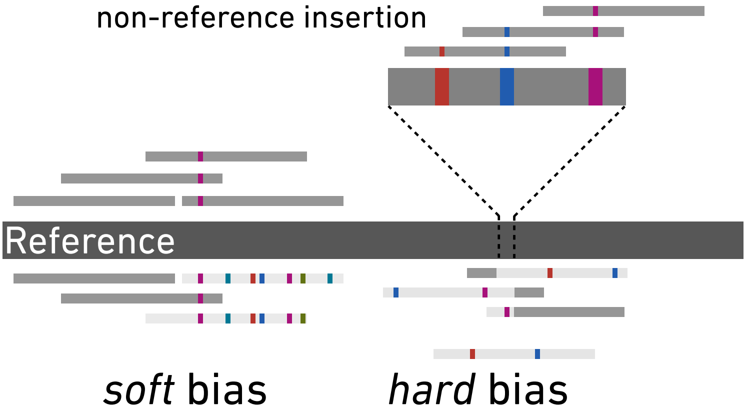
What is reference bias?

Why do we even *want* pangenomes to replace reference genomes?



What is reference bias?

Why do we even *want* pangenomes to replace reference genomes?

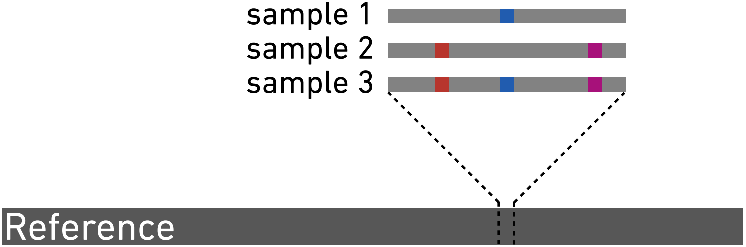


How do we represent complex variaton?

SNPs can **at worst** be quadallelic but a small SV (50 bp) can have $4^{50} \approx 1.3 \times 10^{30}$ alleles

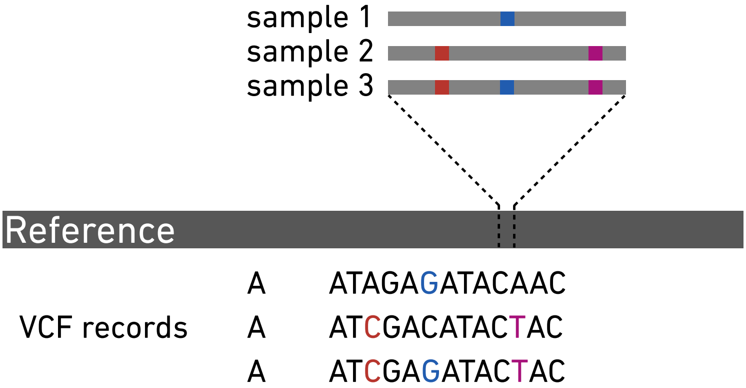
How do we represent complex variaton?

SNPs can **at worst** be quadallelic but a small SV (50 bp) can have $4^{50} \approx 1.3 \times 10^{30}$ alleles



How do we represent complex variaton?

SNPs can **at worst** be quadallelic but a small SV (50 bp) can have $4^{50} \approx 1.3 \times 10^{30}$ alleles



sample 1

sample 2

sample 3

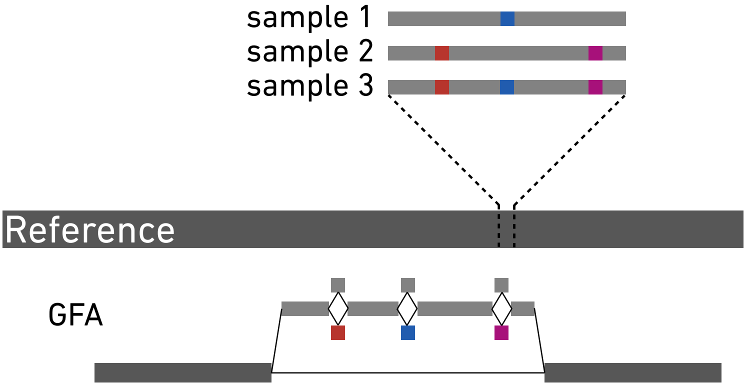
Reference

VCF records

```
A    ATAGAGATACAAC
A    ATCGACATACTAC
A    ATCGAGATACTAC
```

How do we represent complex variaton?

SNPs can **at worst** be quadallelic but a small SV (50 bp) can have $4^{50} \approx 1.3 \times 10^{30}$ alleles



Genome file formats

Most sequencing data (or anything representing genomes) are in *fasta/q*

Sequence alignments are generally in *SAM/BAM*

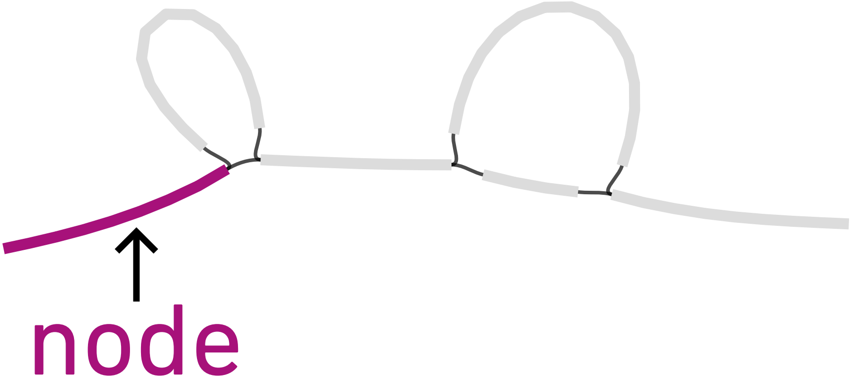
Other “annotation” files like *BED*, *GFF*, etc

Pangenome terminology

How do we describe a graph-based sequence/variation pangenome?

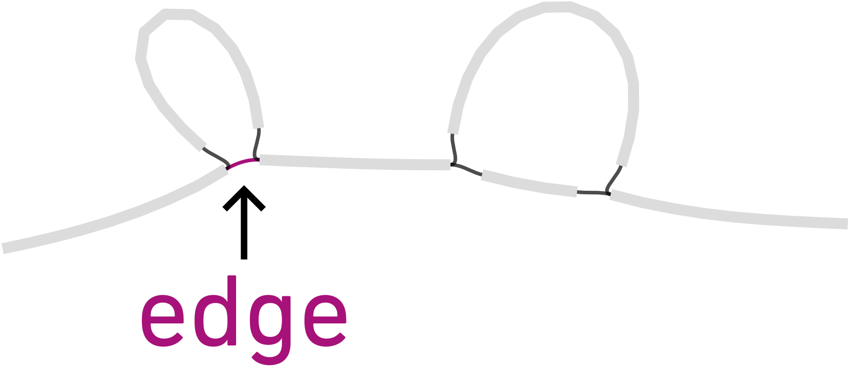
Pangenome terminology

How do we describe a graph-based sequence/variation pangenome?



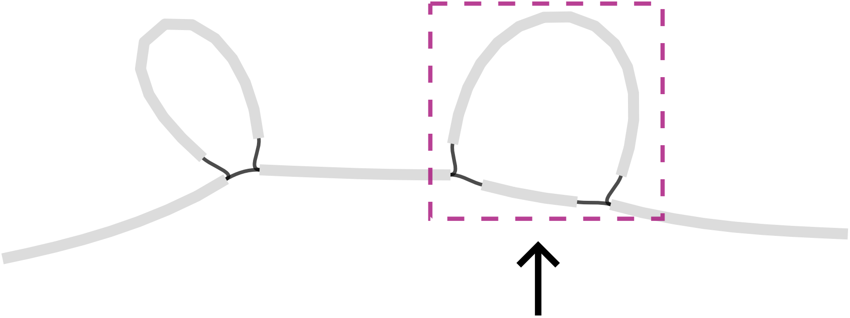
Pangenome terminology

How do we describe a graph-based sequence/variation pangenome?



Pangenome terminology

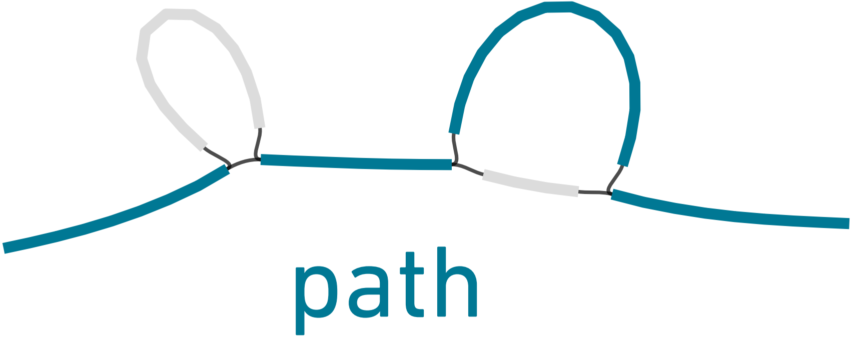
How do we describe a graph-based sequence/variation pangenome?



"bubble"

Pangenome terminology

How do we describe a graph-based sequence/variation pangenome?



Pangenome file formats

What are the pangenomic file equivalents?

Pangenome file formats

What are the pangenomic file equivalents?

GFA: **G**raphical **F**ragment *A**ssembly

Pangenome file formats

What are the pangenomic file equivalents?

GFA: **G**raphical **F**ragment *A**ssembly

Three main components:

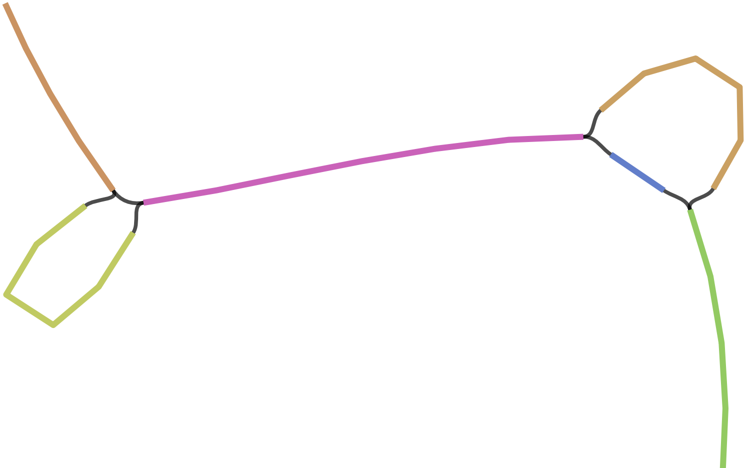
- S-lines: the sequence of the nodes
- L-lines: how the graph is connected with edges
- P-lines: how a “sample” traverses the graph (*optional*)

Pangenome file formats

```
H       VN:Z:1.0
S       1       AATTTACC
S       2       GGTAT
S       3       T
S       4       CCCGATA
S       5       GGACTA
S       6       TTAC
L       1       +       2       +       0M
L       1       +       3       +       0M
L       2       +       4       +       0M
L       3       +       4       +       0M
L       4       +       5       +       0M
L       5       +       6       +       0M
L       4       +       6       +       0M
P       Alice    1+,2+,4+,5+,6+ *
P       Bob      1+,3+,4+,6+ *
```

Pangenome file formats

That looks like



Pangenome file formats

Most downstream tools have their own “efficient” representations of *.gfa* files

- `.og`
- `.vg`
- `.xg`
- `.gbz`

Pangenome file formats

Most downstream tools have their own “efficient” representations of *.gfa* files

- `.og`
- `.vg`
- `.xg`
- `.gbz`

These graphs contain a lot of information.

GFA is human-readable, but binary formats are more compute efficient

Pangenome file formats

GAF: **G**raph **A**lignment **F**ormat

A graph “superset” of PAF (**P**airwise **m**Apping **F**ormat).

Pangenome file formats

GAF: **G**raph **A**lignment **F**ormat

A graph “superset” of PAF (**P**airwise **m**Apping **F**ormat).

Similar to *.sam* files, recording details on:

- which read
- where does it align
- how good was that alignment

Pangenome file formats

GAF: **G**raph **A**lignment **F**ormat
A graph “superset” of PAF (**P**airwise **m**Apping **F**ormat).

Similar to *.sam* files, recording details on:

- which read
- where does it align
- how good was that alignment

Likewise, this is human-readable, and so some tools prefer the binary version `.gam`.

Graph building

Building a “variation graph” starts with a set of assemblies

Graph building

Building a “variation graph” starts with a set of assemblies

We often rename chromosome names using PanSN-spec

[sample]#[haplotype]#[contig](#[fragment/subrange])

Graph building

Building a “variation graph” starts with a set of assemblies

We often rename chromosome names using PanSN-spec

[sample]#[haplotype]#[contig](#[fragment/subrange])

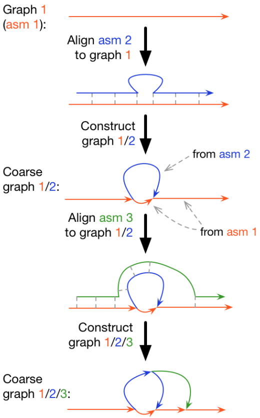
- avoids conflicts of many e.g. “>chr1” sequences
- encodes some metadata *within* the file
- enables selectively grouping/renaming values by “classification”

minigraph

Augments a linear reference “backbone” with *sufficiently* new variation

minigraph

Augments a linear reference “backbone” with *sufficiently* new variation

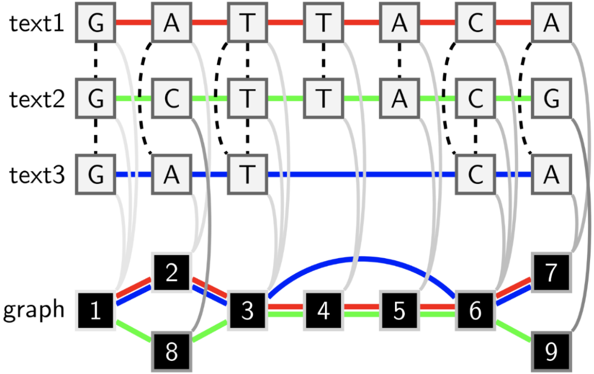


pggb

All-versus-all alignment, followed by complicated cleaning of the graph structure

pggb

All-versus-all alignment, followed by complicated cleaning of the graph structure



Different approaches

	≥ 50 bp	< 50 bp	Reference-based	Lossless	N+1	Compute
`minigraph`	Yes	No	Yes	No	Easy	Laptop
`cactus`	Yes	Yes	No-ish	Yes	Easy-ish	Cluster
`pggb`	Yes	Yes	No	Yes	Rebuild	Big cluster

Different approaches

	≥ 50 bp	< 50 bp	Reference-based	Lossless	N+1	Compute
`minigraph`	Yes	No	Yes	No	Easy	Laptop
`cactus`	Yes	Yes	No-ish	Yes	Easy-ish	Cluster
`pggb`	Yes	Yes	No	Yes	Rebuild	Big cluster

We can perfectly reconstruct any assembly from a *lossless* graph

Different approaches

	≥ 50 bp	< 50 bp	Reference-based	Lossless	N+1	Compute
`minigraph`	Yes	No	Yes	No	Easy	Laptop
`cactus`	Yes	Yes	No-ish	Yes	Easy-ish	Cluster
`pggb`	Yes	Yes	No	Yes	Rebuild	Big cluster

We can perfectly reconstruct any assembly from a *lossless* graph

Pick the approach that best matches **your** research question

Other pangenome tools

Variation is a powerful tool, but easy to get overwhelmed by

Other pangenome tools

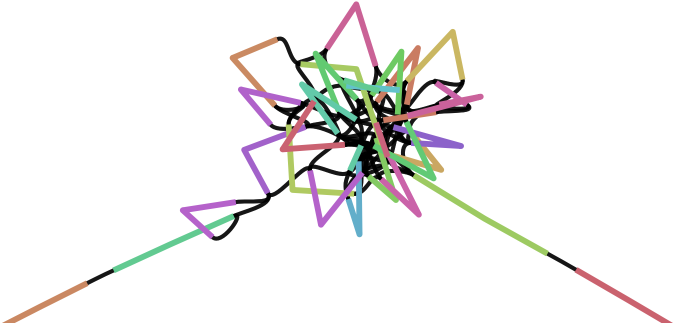
Variation is a powerful tool, but easy to get overwhelmed by

- `pangene`
- `pgr-tk`
- many dBG tools (`bifrost` etc.)

Other pangenome tools

Variation is a powerful tool, but easy to get overwhelmed by

- `pangene`
- `pgr-tk`
- many dBG tools (`bifrost` etc.)



Pangenome visualisation

IGV (**I**ntegrative **G**enomics **V**iewer, https://igv.org/doc/desktop/) is a useful tool for visualising different formats of genomic data:

- read alignments
- bed files
- gene annotations

Pangenome visualisation

IGV (**I**ntegrative **G**enomics **V**iewer, https://igv.org/doc/desktop/) is a useful tool for visualising different formats of genomic data:

- read alignments
- bed files
- gene annotations

Is there a pangenomic equivalent?

Visualising **pan**genomic data

Everything is more complicated in the pangenomic world

Visualising **pan**genomic data

Everything is more complicated in the pangenomic world

What are we trying to visualise?

- Synteny between many assemblies?
- Genic regions in a pangenome?
- Alignments to a pangenome?

Interactive visualisation

How do we visualise the *.gfa* output of pangenome construction?

Interactive visualisation

How do we visualise the *.gfa* output of pangenome construction?

One of the most common tools is `BandageNG` (https://github.com/asl/BandageNG).

Interactive visualisation

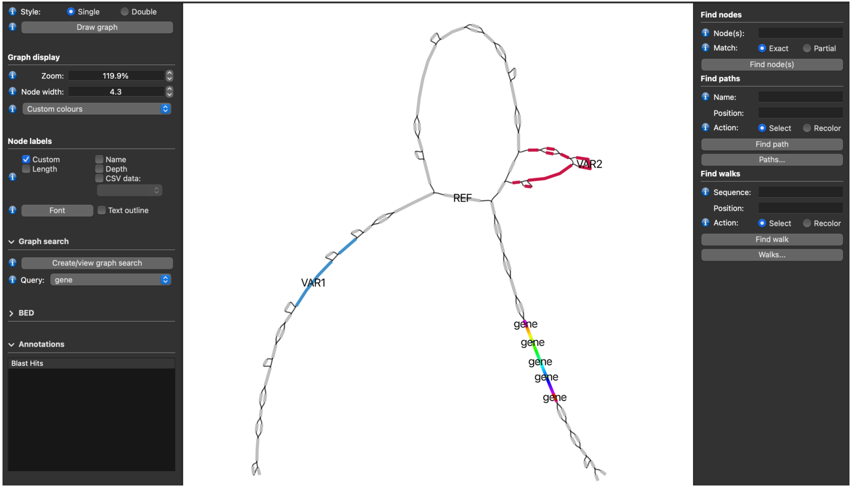
How do we visualise the *.gfa* output of pangenome construction?

One of the most common tools is `BandageNG` (https://github.com/asl/BandageNG).

We'll explore this in the practical, but it has several advantages:

- easy to install
- quick to load small-to-moderate sized graphs
- extensive analytic functionality

Interactive visualisation



Style: ● Single ○ Double

Zoom: 119.9%

Node width: 4.3

Custom colours

☑ Custom □ Name
□ Length □ Depth
□ CSV data:

□ Text outline

Query: gene

Node(s):

Match: ● Exact ○ Partial

Name:

Position:

Action: ● Select ○ Recolor

Sequence:

Position:

Action: ● Select ○ Recolor

Static visualisation

Large graphs (many nodes and/or edges) are complex to render.

Static visualisation

Large graphs (many nodes and/or edges) are complex to render.

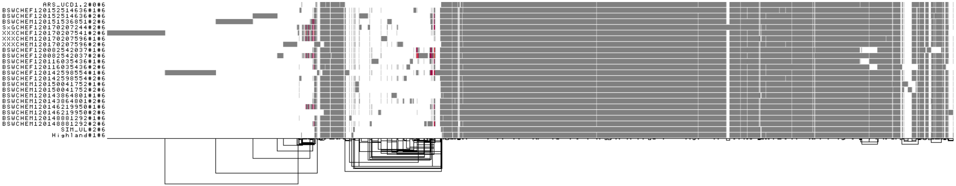
Let the computer do the **hard** work and render a static representation!

Static visualisation

Break pangenome down into multiple linear blocks

Static visualisation

- Break pangenome down into multiple linear blocks

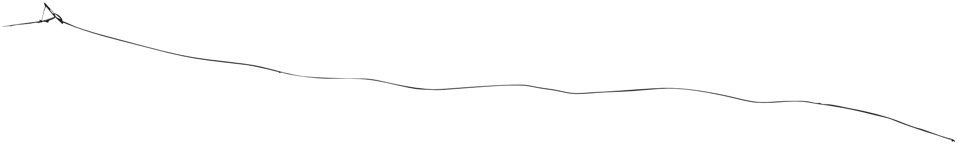


Static visualisation

"Optimally" lay out nodes/edges in 2D with a *Hogwild!* algorithm.

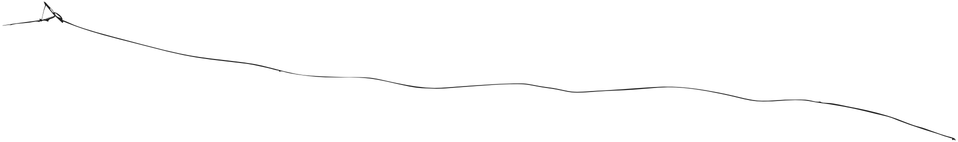
Static visualisation

"Optimally" lay out nodes/edges in 2D with a *Hogwild!* algorithm.



Static visualisation

"Optimally" lay out nodes/edges in 2D with a *Hogwild!* algorithm.



This step took ~**30%** of the entire HPRC pipeline runtime!

Pangenome communities

Building pangenomes per chromosome is much easier than genome-wide

Pangenome communities

Building pangenomes per chromosome is much easier than genome-wide

What if

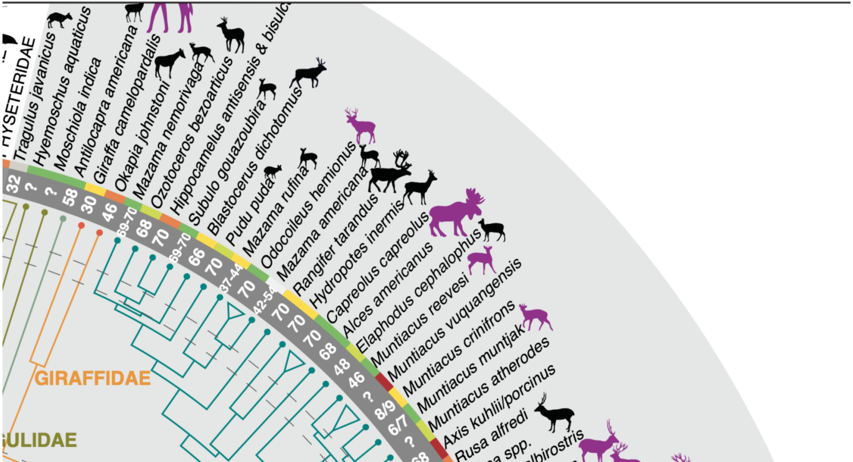
- we care about interchromosomal events
- we don’t know how to define “per chromosome”
- we don’t have assigned chromosomes

Nonuniform karyotypes

Even “similar” species can undergo complex chromosomal evolution

Nonuniform karyotypes

Even “similar” species can undergo complex chromosomal evolution



Community detection

pggb implemented community detection

- map whole genomes all-versus-all
- build a *weighted* network from all submappings
- use graph theory community-detection algorithms

Community detection

pggb implemented community detection

- map whole genomes all-versus-all
- build a *weighted* network from all submappings
- use graph theory community-detection algorithms

Translocations or complex rearrangements are also *identified*

Community detection

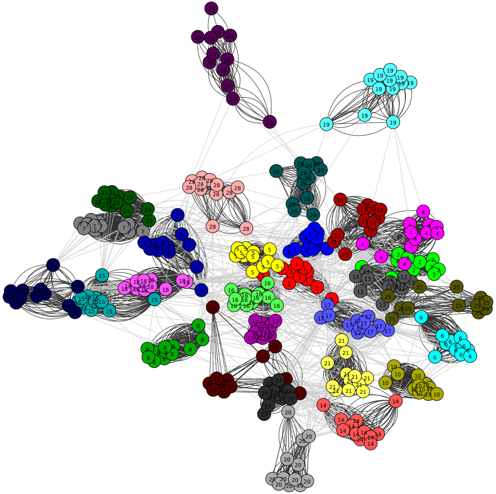
pggb implemented community detection

- map whole genomes all-versus-all
- build a *weighted* network from all submappings
- use graph theory community-detection algorithms

Translocations or complex rearrangements are also *identified*

Distinguishing signal from noise is hard for small/infrequent mappings

Community detection



Pangenome validation

How do we know if the pangenome we built is any good?

Pangenome validation

How do we know if the pangenome we built is any good?

What does *good* even mean to us?

Pangenome analyses

There are several tools useful for checking pangenome construction and content

- `gfatools`
- `odgi`
- `panacus`
- `gretl`

Pangenome graph statistics

After building a graph, the simplist statistics to check are:

- total sequence length
- maxmimum and average node size
- node depth distribution

Pangenome graph statistics

Graphs can be described by the number of nodes and edges they contain.

Different graphs (e.g., `pggb` versus `minigraph`) may have similar length, but very different node/edge counts.

Pangenome graph statistics

Graphs can be described by the number of nodes and edges they contain.

Different graphs (e.g., `pggb` versus `minigraph`) may have similar length, but very different node/edge counts.

Consider the average node size (pangenome length / number of nodes) or average edge degree (number of nodes / number of edges)

Pangenome graph statistics

Graphs can be described by the number of nodes and edges they contain.

Different graphs (e.g., `pggb` versus `minigraph`) may have similar length, but very different node/edge counts.

Consider the average node size (pangenome length / number of nodes) or average edge degree (number of nodes / number of edges)

Should be *reasonable* values (how many bases do you expect before a SNP?)

Pangenome graph statistics

From `gfatools stat` on a large, base-level bovine pangenome of chromosome 1 (159 Mb)

```
Number of segments: 10140559
Number of links: 14371940
Number of arcs: 28743880
Total segment length: 200985993
Average segment length: 19.820
Max degree: 106924
Average degree: 1.417
```

Pangenome graph statistics

From `gfatools stat` on a large, base-level bovine pangenome of chromosome 1 (159 Mb)

```
Number of segments: 10140559
Number of links: 14371940
Number of arcs: 28743880
Total segment length: 200985993
Average segment length: 19.820
Max degree: 106924
Average degree: 1.417
```

The total pangenome size should *approximately* be equal to the reference plus all variation.

Pangenome openness

How does the growth of a pangenome change with more samples?

Pangenome openness

How does the growth of a pangenome change with more samples?

We can use *Heap's law* from text analysis: $N \propto n^{-\alpha}$

Pangenome openness

How does the growth of a pangenome change with more samples?

We can use *Heap's law* from text analysis: $N \propto n^{-\alpha}$

If $\alpha > 1$, the pangenome is **closed**, otherwise if $\alpha \leq 1$, the pangenome is **open**.

Pangenome openness

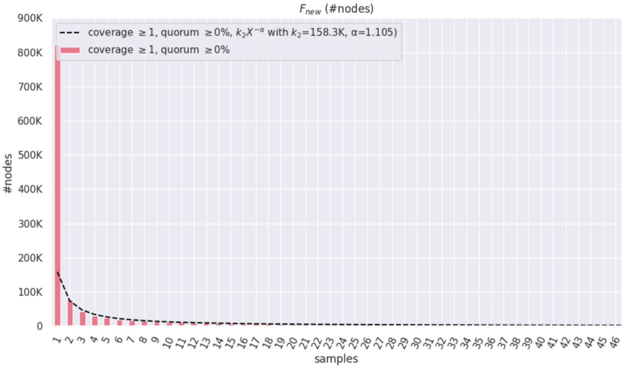
With enough samples, we can estimate α

Care is needed about how much variation is *expected* to be shared . . .

Pangenome openness

With enough samples, we can estimate α

Care is needed about how much variation is *expected* to be shared . . .



Pangenome layers

Pangenome openness effectively addresses the total unique sequence.

What about different levels of intersection?

Pangenome layers

Pangenome openness effectively addresses the total unique sequence.
What about different levels of intersection?

We can characterise pangenome *nodes* as:

- **core**: present in all/most samples
- **shell**: present in at least two samples
- **cloud**: present in only one sample
- **flexible/dispensable**: varies, but something like shell/cloud

Pangenome layers

With enough samples, we expect minimal sequence to be “core”

Pangenome layers

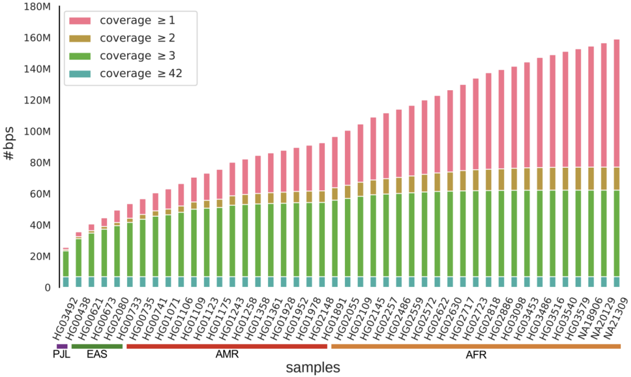
With enough samples, we expect minimal sequence to be “core”

Misassemblies can also further reduce the “core”

Pangenome layers

With enough samples, we expect minimal sequence to be “core”

Misassemblies can also further reduce the “core”



Downstream pangenomics

Once we have a “good” pangenome, what can we actually do with it?

Calling pangenome variants

We can also call variants *within* the pangenome with `vg deconstruct`

Calling pangenome variants

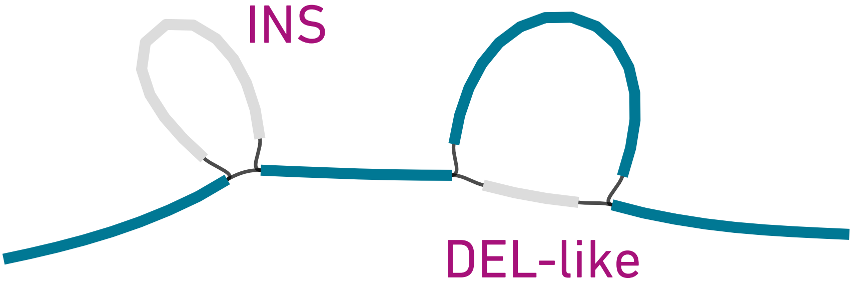
We can also call variants *within* the pangenome with `vg deconstruct`

“Project” back into linear space (losing *some* pangenomic benefits)

Calling pangenome variants

We can also call variants *within* the pangenome with `vg deconstruct`

“Project” back into linear space (losing *some* pangenomic benefits)



Aligning to pangenomes

Linear-reference alignment is “simple”

Aligning to pangenomes

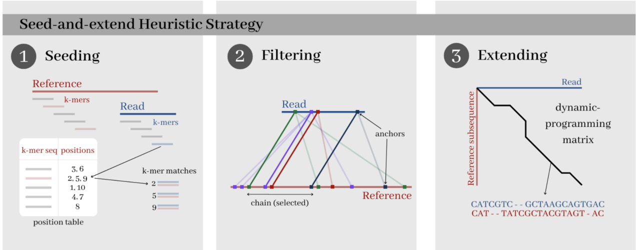
Linear-reference alignment is “simple”

- check if next base is a match
- genomic distance matches insert size
- opposite strand is the reverse complement

Aligning to pangenomes

Linear-reference alignment is “simple”

- check if next base is a match
- genomic distance matches insert size
- opposite strand is the reverse complement



Aligning to pangenomes

`vg giraffe` was a huge step forward for read-to-graph alignment

Aligning to pangenomes

`vg giraffe` was a huge step forward for read-to-graph alignment

Many algorithms silently assume “DAGs” (**D**irected **A**cyclic **G**raph)

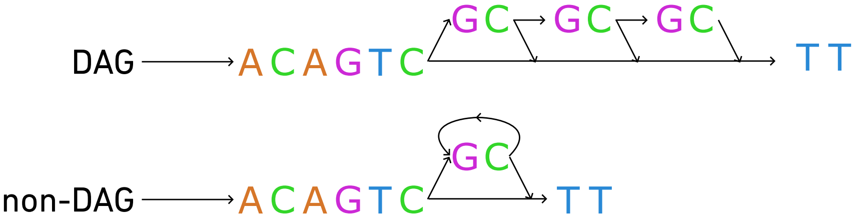
non-DAGs allow revisiting a node (maybe infinitely times)

Aligning to pangenomes

`vg giraffe` was a huge step forward for read-to-graph alignment

Many algorithms silently assume “DAGs” (**D**irected **A**cyclic **G**raph)

non-DAGs allow revisiting a node (maybe infinitely times)



Long read alignment

Long reads span more bubbles in graphs, exponentially complicating alignment

Long read alignment

Long reads span more bubbles in graphs, exponentially complicating alignment

Currently limited number of “production” tools

- `GraphAligner`
- `vg giraffe-lr` soon!

Personalised pangenomes

Pangenomes are critical to give coordinates to all sequence

Personalised pangenomes

Pangenomes are critical to give coordinates to all sequence

We want to **maintain** those coordinates across all analyses

Personalised pangenomes

Pangenomes are critical to give coordinates to all sequence

We want to **maintain** those coordinates across all analyses

Can we “filter” out graph complexity that isn’t useful for a *given* sample?

Irrelevant pangenomic variation

Given any genomic sequencing, we can easily calculate a set of *k*-mers for that sample

Irrelevant pangenomic variation

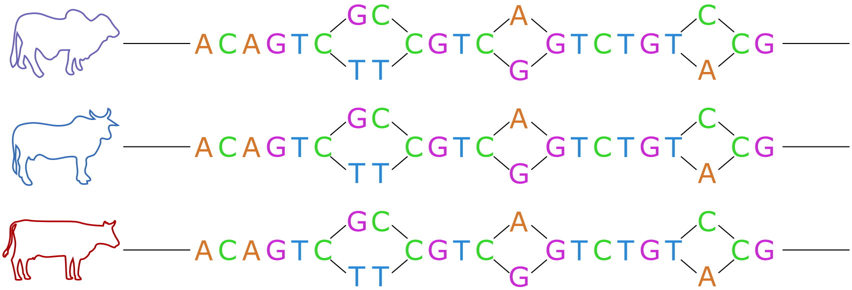
Given any genomic sequencing, we can easily calculate a set of *k*-mers for that sample

Retain nodes/edges which span those *k*-mers, rather than filtering by allele frequency

Irrelevant pangenomic variation

Given any genomic sequencing, we can easily calculate a set of *k*-mers for that sample

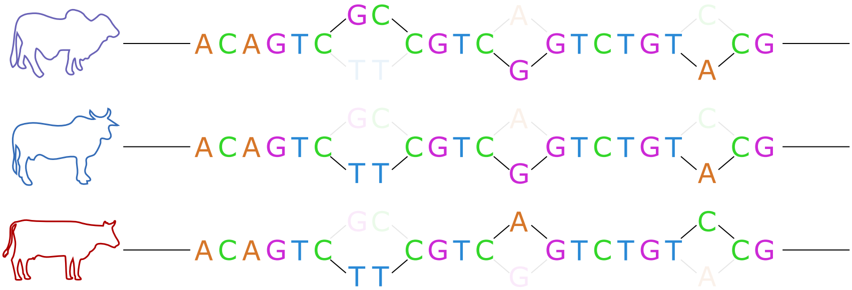
Retain nodes/edges which span those *k*-mers, rather than filtering by allele frequency



Irrelevant pangenomic variation

Given any genomic sequencing, we can easily calculate a set of *k*-mers for that sample

Retain nodes/edges which span those *k*-mers, rather than filtering by allele frequency



ACAGTC GC CGTC G GTCTGT A CG
ACAGTC TT CGTC G GTCTGT A CG
ACAGTC TT CGTC A GTCTGT C CG

Upstream blackbox

Pangenomes can be challenging and don't always match downstream input formats

Upstream blackbox

Pangenomes can be challenging and don't always match downstream input formats

A user could provide a complete reference pangenome and (short) reads

Inside a black box, we can then run

- `vg haplotype` (personalise the pangenome)
- `vg giraffe` (align to the pangenome)
- `vg surject` (convert back to linear coordiantes)
- e.g. `DeepVariant` (call variants as per usual)

Upstream blackbox

Pangenomes can be challenging and don't always match downstream input formats

A user could provide a complete reference pangenome and (short) reads

Inside a black box, we can then run

- `vg haplotype` (personalise the pangenome)
- `vg giraffe` (align to the pangenome)
- `vg surject` (convert back to linear coordiantes)
- e.g. `DeepVariant` (call variants as per usual)

Improved variant calls without *direct* exposure to the pangenome

Targeted pangenomes

A *reference* pangenome should cover the entire genome

Targeted pangenomes

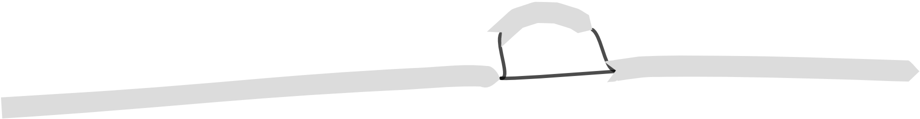
A *reference* pangenome should cover the entire genome

Most pangenome papers focus on one/several *QTL*

Targeted pangenomes

A *reference* pangenome should cover the entire genome

Most pangenome papers focus on one/several *QTL*

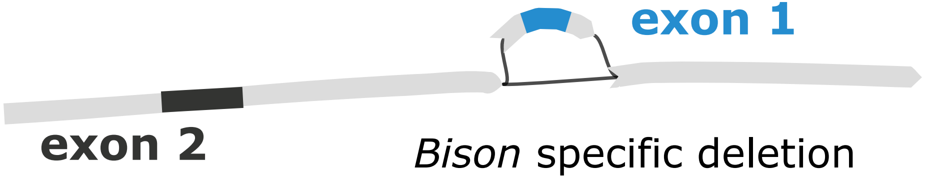


Bison specific deletion

Targeted pangenomes

A *reference* pangenome should cover the entire genome

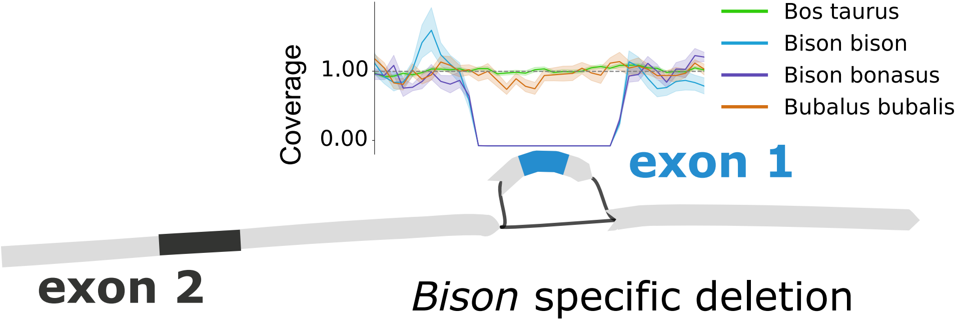
Most pangenome papers focus on one/several *QTL*



Targeted pangenomes

A *reference* pangenome should cover the entire genome

Most pangenome papers focus on one/several *QTL*



Bison specific deletion

Manual QTL pangenome

For a given reference-annotated region, we can:

Manual QTL pangenome

For a given reference-annotated region, we can:

- lift over equivalent reference coordinates into other assemblies
- extract relevant section of those assemblies
- build a pangenome from these sequences

A better approach

`impg` outlines a different approach

A better approach

`impg` outlines a different approach

- conduct the hard all-to-all mapping once
- extract *transitive* regions based on a set of coordinates
- build a pangenome from those sequences

A new whole-genome approach?

Building many small pangenomes is easier than one big pangenome

Can we go from per *chromosome* to per *window*?

A new whole-genome approach?

Building many small pangenomes is easier than one big pangenome

Can we go from per *chromosome* to per *window*?

Recombine pieces into chromosome-scale graphs with `gfalace`

A new whole-genome approach?

Building many small pangenomes is easier than one big pangenome

Can we go from per *chromosome* to per *window*?

Recombine pieces into chromosome-scale graphs with `gfalace`

Some unresolved concerns:

- boundary conditions are poorly defined
- events spanning the "split length" might be lost
- detecting subgraph isomorphisms is hard

Acrocentric recombination

Caveat: biologists probably knew before the computer people

Acrocentric recombination

Caveat: biologists probably knew before the computer people

Initial human pangenome construction lead to huge tangles in *some* chromosomes

Acrocentric recombination

Caveat: biologists probably knew before the computer people

Initial human pangenome construction lead to huge tangles in *some* chromosomes

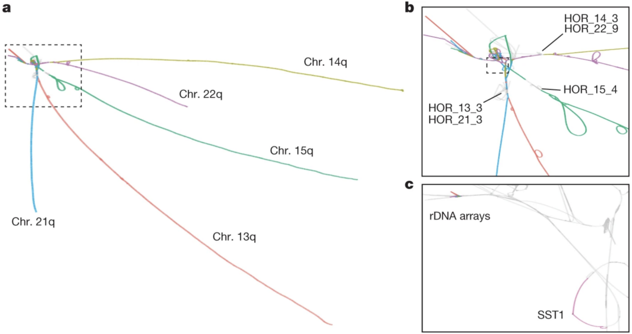
Pangenomes (at minimum) offer a new perspective on existing questions

Acrocentric recombination

Pseudo-homologous regions near centomeres drive Robertsonian translocations

Acrocentric recombination

Pseudo-homologous regions near centomeres drive Robertsonian translocations



{Guarracino et al. 2023}

Braided snarls

Strange pangenomic structures are generally worrying

Braided snarls

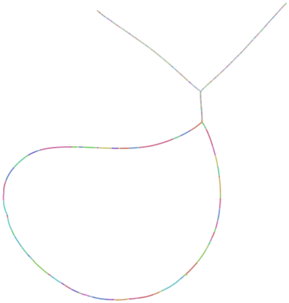
Strange pangenomic structures are generally worrying

They might actually reveal biology in way we didn’t anticipate!

Braided snarls

Strange pangenomic structures are generally worrying

They might actually reveal biology in way we didn't anticipate!



Superpangenomes

Typically "pangenomes" refer to a single species

Superpangenomes

Typically “pangenomes” refer to a single species

Superpangenomes include more diverse assemblies, e.g., *genus*-level

Superpangenomes

Typically “pangenomes” refer to a single species

Superpangenomes include more diverse assemblies, e.g., *genus*-level

Hyper/**Mega**/**Ultra**pangenomes?

Superpangenomes

What happens if we include many related species into a pangenome?

Superpangenomes

What happens if we include many related species into a pangenome?

- ultraconserved elements are still roughly single nodes
- species-specific variation are distinct paths through bubbles
- phylogeny-related information present in nested bubbles

Summary – starting with pangenomes

Pangenomes can integrate many genomes into one structure to mitigate reference bias

Summary – starting with pangenomes

Pangenomes can integrate many genomes into one structure to mitigate reference bias

Building pangenomes is still hard, but quickly getting easier

Summary – starting with pangenomes

Pangenomes can integrate many genomes into one structure to mitigate reference bias

Building pangenomes is still hard, but quickly getting easier

Pangenome openess or graph statistics help us know if our graphs are “good”

Summary – working with pangenomes

We can use the pangenome as a *reference* or as a *resource*

Summary – working with pangenomes

We can use the pangenome as a *reference* or as a *resource*

T2T assemblies and pangenomes *can* unlocking entirely new perspectives

Summary – working with pangenomes

We can use the pangenome as a *reference* or as a *resource*

T2T assemblies and pangenomes *can* unlocking entirely new perspectives

Population-scale read alignment and “direct” pangenomic analyses are *becoming* possible

Hands on pangenomics

During the activity we'll look at

- building a small `minigraph` pangenome
- visualising that pangenome in `BandageNG`
- using `gfatools` to find regions of interest

Questions?