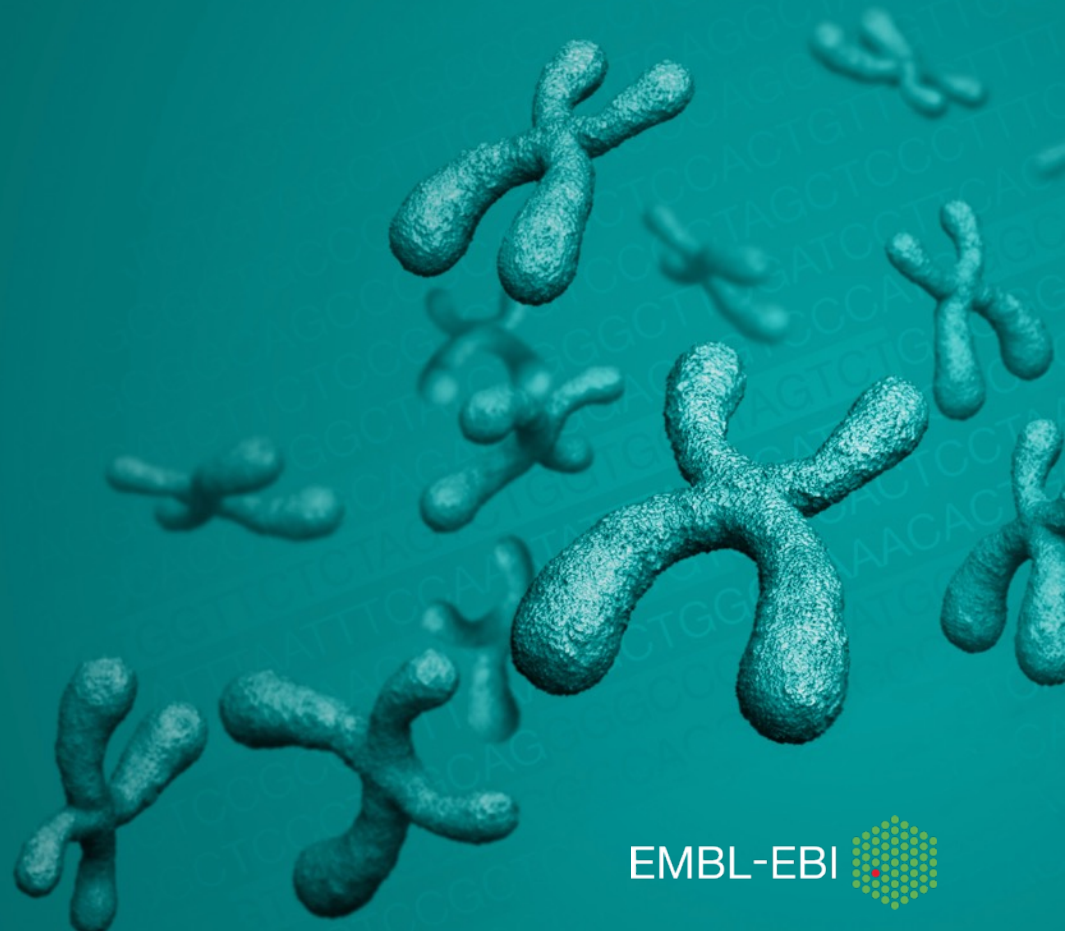
Pangenome graph structures reveal more genetic variation between divergent genomes

Rachel Colquhoun, Michael Hall, Derrick Crook, Zamin Iqbal

rmcolq@ebi.ac.uk



EMBL-EBI

Talk outline

1. The problem we want to solve
2. Bacterial inheritance 101
3. A new pangenome reference approach
4. Initial results for *E. coli*
5. How this infrastructure can be extended to mixtures

Motivation

- *Enterobacteriaceae* are commonly found in normal microflora in the human gastrointestinal tract

Motivation

- *Enterobacteriaceae* are commonly found in normal microflora in the human gastrointestinal tract

Enterobacter sp.

E. coli

Klebsiella sp.

Salmonella enterica

Proteus mirabilis

Motivation

- *Enterobacteriaceae* are commonly found in normal microflora in the human gastrointestinal tract
- Horizontal gene transfer often occurs among *Enterobacteriaceae* and between pathogenic and commensal strains

Schjørring & Krogfelt, Int J Microbiol (2011)
Stecher et al., PNAS (2012)

Motivation

- *Enterobacteriaceae* are commonly found in normal microflora in the human gastrointestinal tract
- Horizontal gene transfer often occurs among *Enterobacteriaceae* and between pathogenic and commensal strains

Schjørring & Krogfelt, Int J Microbiol (2011)
Stecher et al., PNAS (2012)

- To fully understand the dynamic interactions with and within such species in the microbiome we need to be able to compare diverse genomes

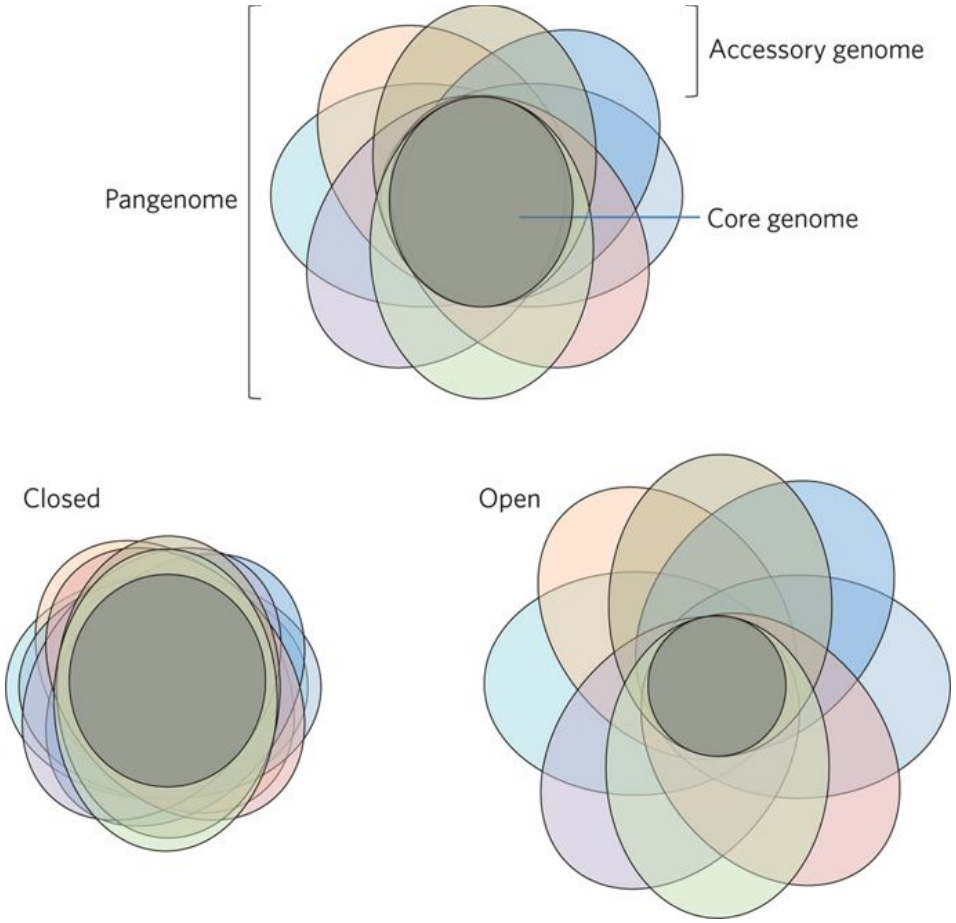
Motivation

- *Enterobacteriaceae* are commonly found in normal microflora in the human gastrointestinal tract
- Horizontal gene transfer often occurs among *Enterobacteriaceae* and between pathogenic and commensal strains

Schjørring & Krogfelt, Int J Microbiol (2011)
Stecher et al., PNAS (2012)

- To fully understand the dynamic interactions with and within such species in the microbiome we need to be able to compare diverse genomes (at the single nucleotide level)

Pangenome diversity

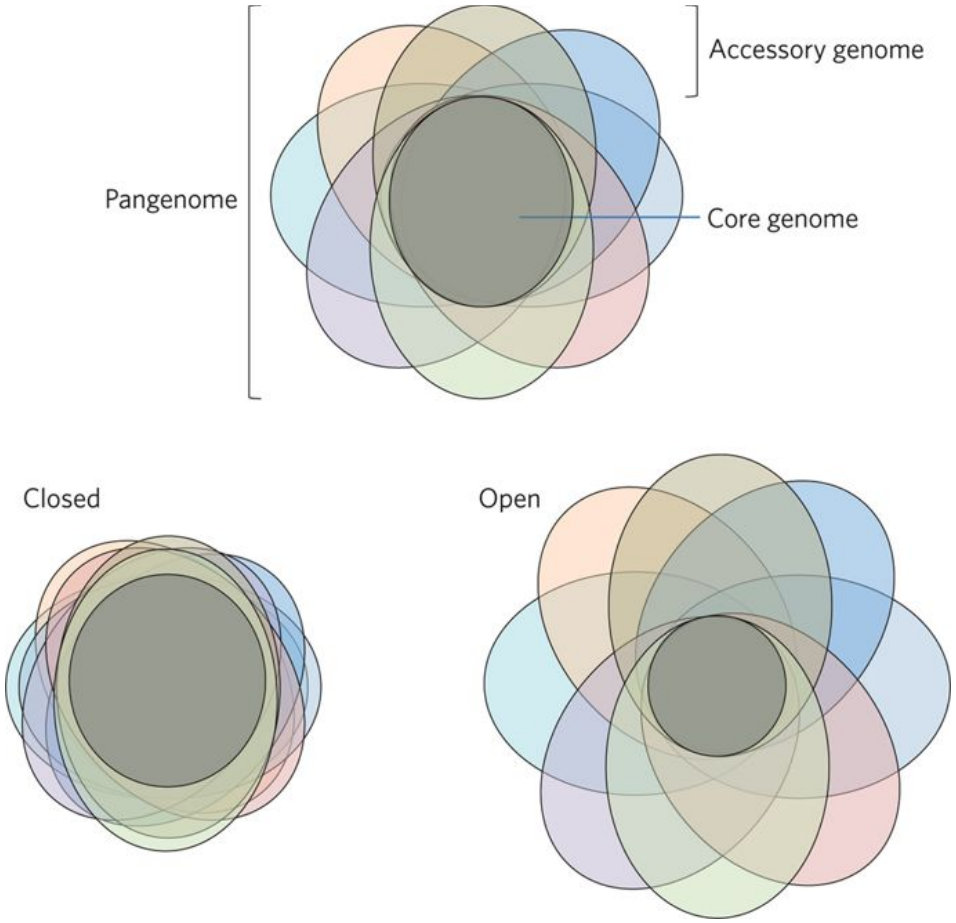


McInerney et al., Nat Micro (2017)

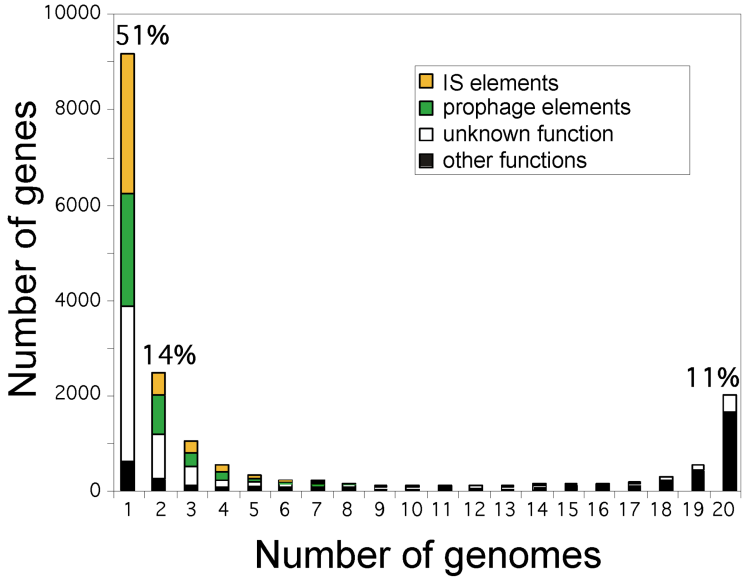
For *E. coli:*

- A single genome contains ~5000 genes
- The pangenome contains ~90,000 genes

Pangenome diversity



McInerney et al., Nat Micro (2017)

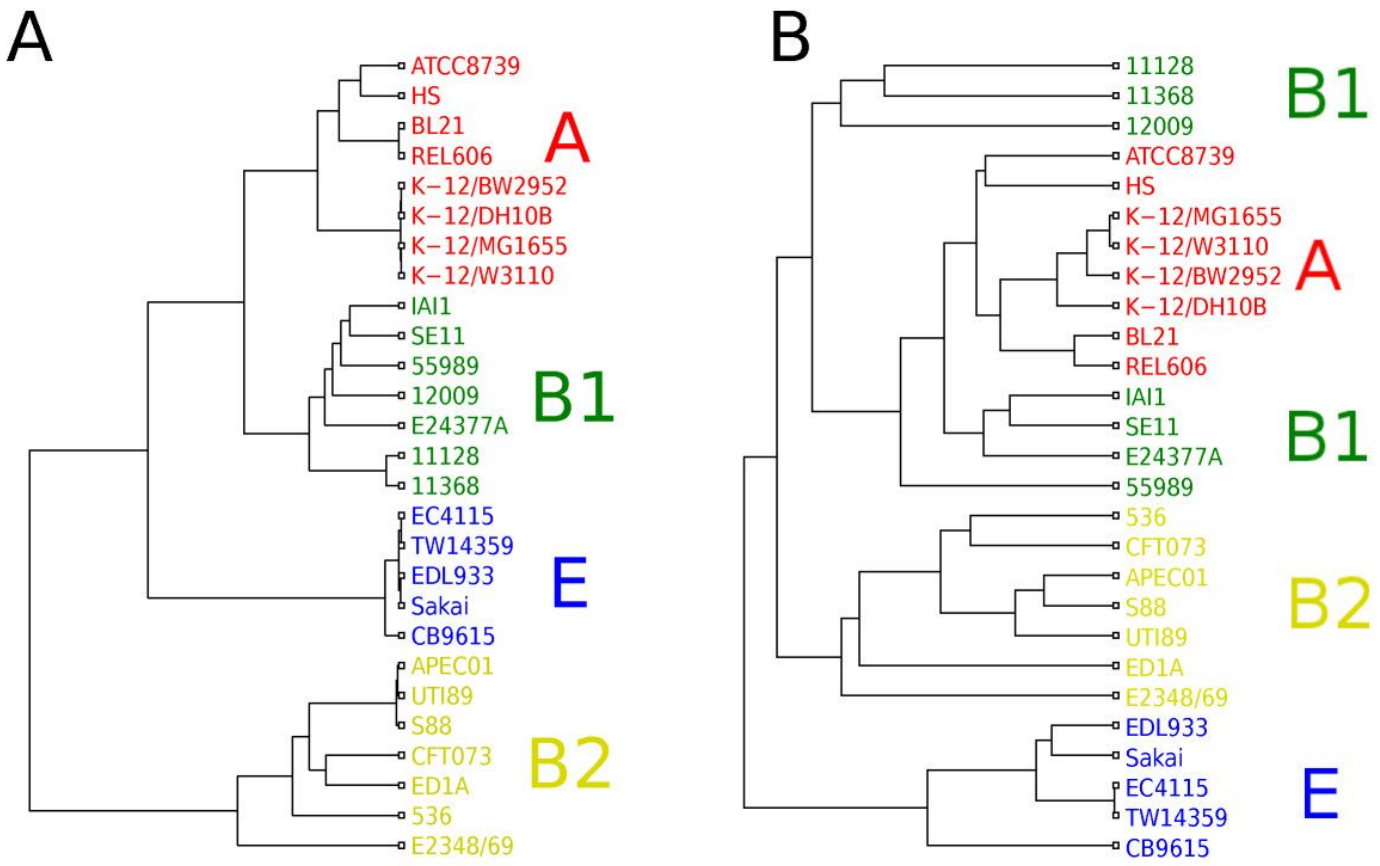


Touchon et al., PloS Genetics (2009)

For *E. coli:*

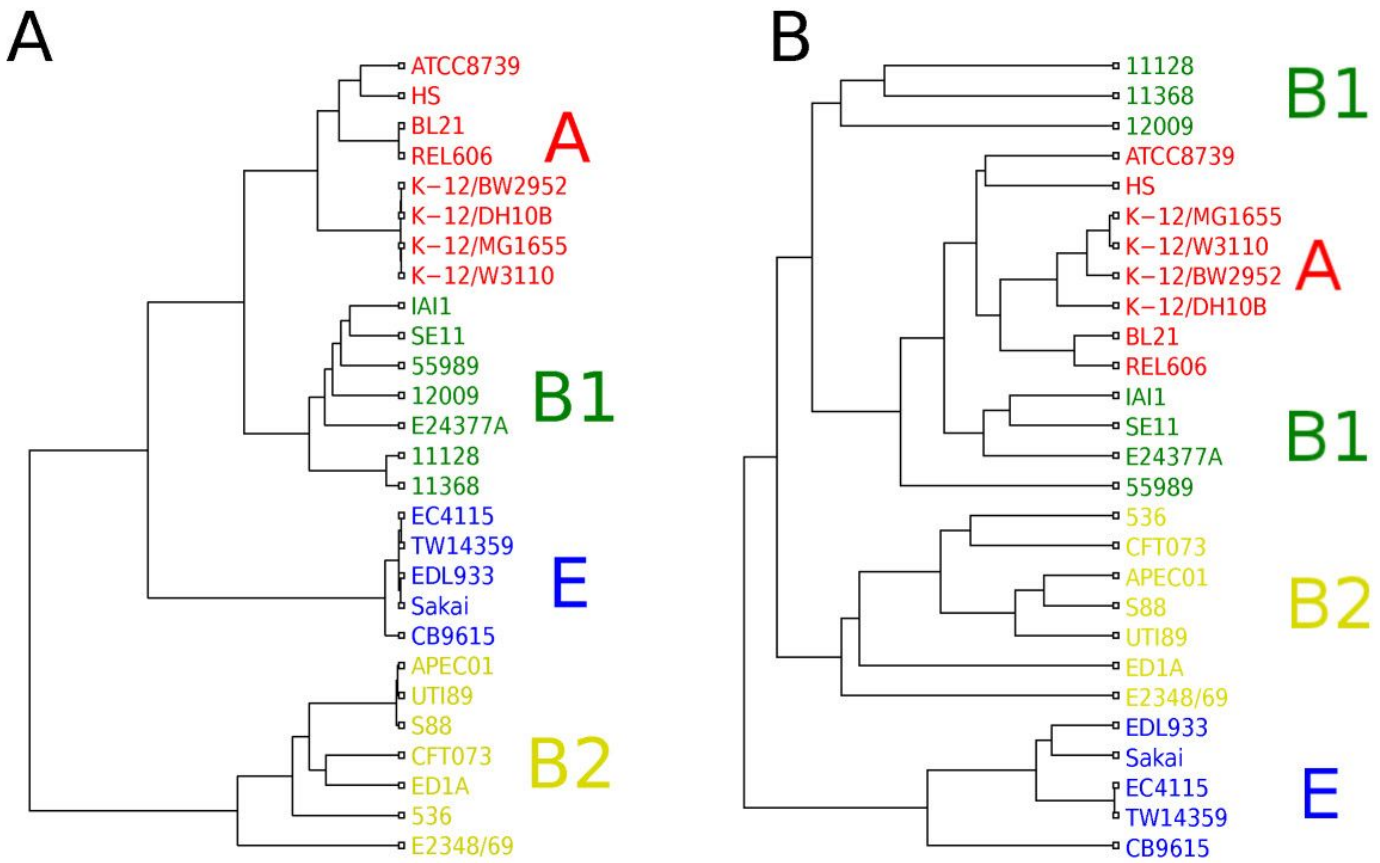
- A single genome contains ~5000 genes
- The pangenome contains ~90,000 genes
- Most genes are rare

Relatedness in the core and accessory



Didelot et al., BMC Genomics (2012)

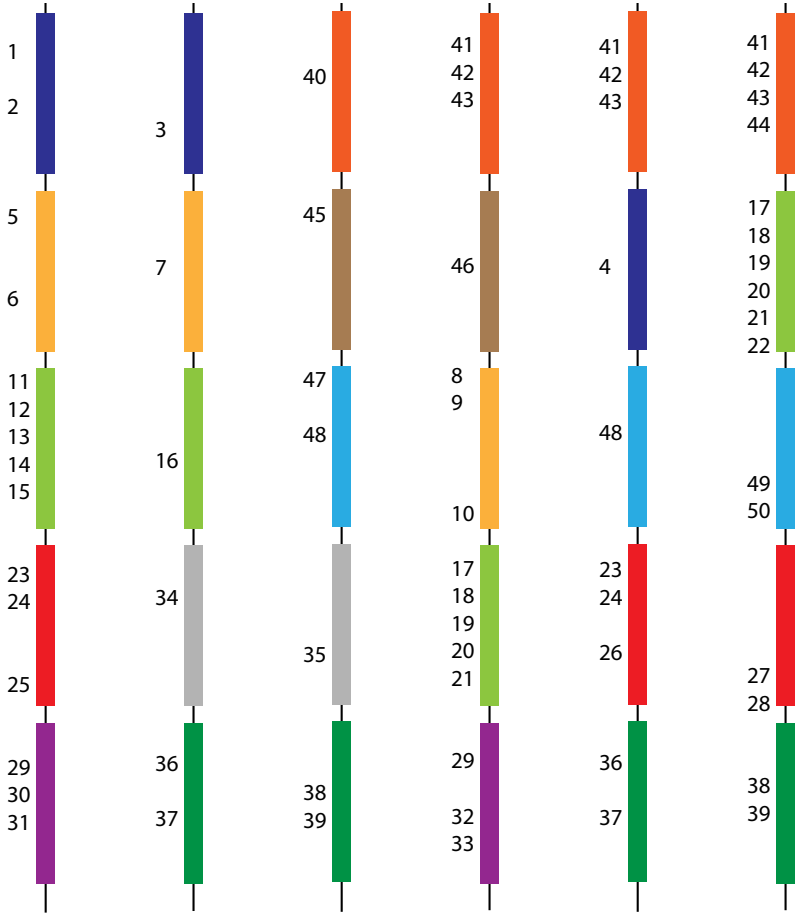
Relatedness in the core and accessory



Two genomes distant on the core tree can have more similar gene repertoires than two genomes which are close on the core tree

Didelot et al., BMC Genomics (2012)

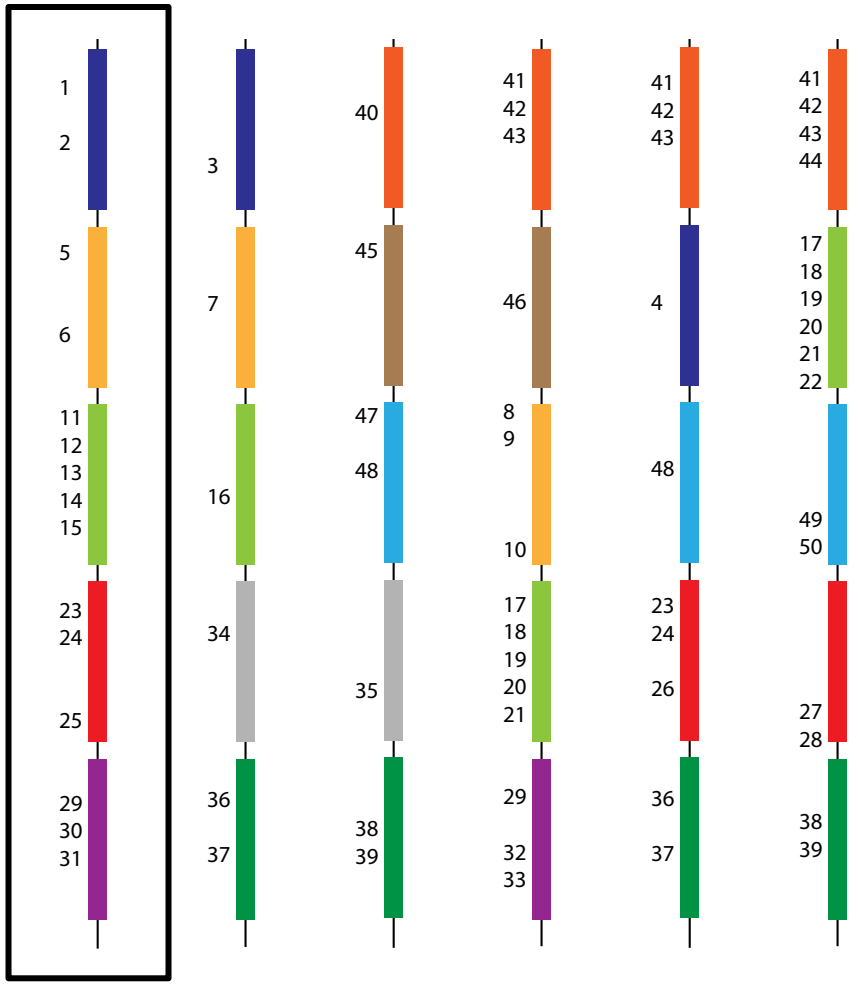
Example of the problem



Example of the problem

Suppose this is the reference

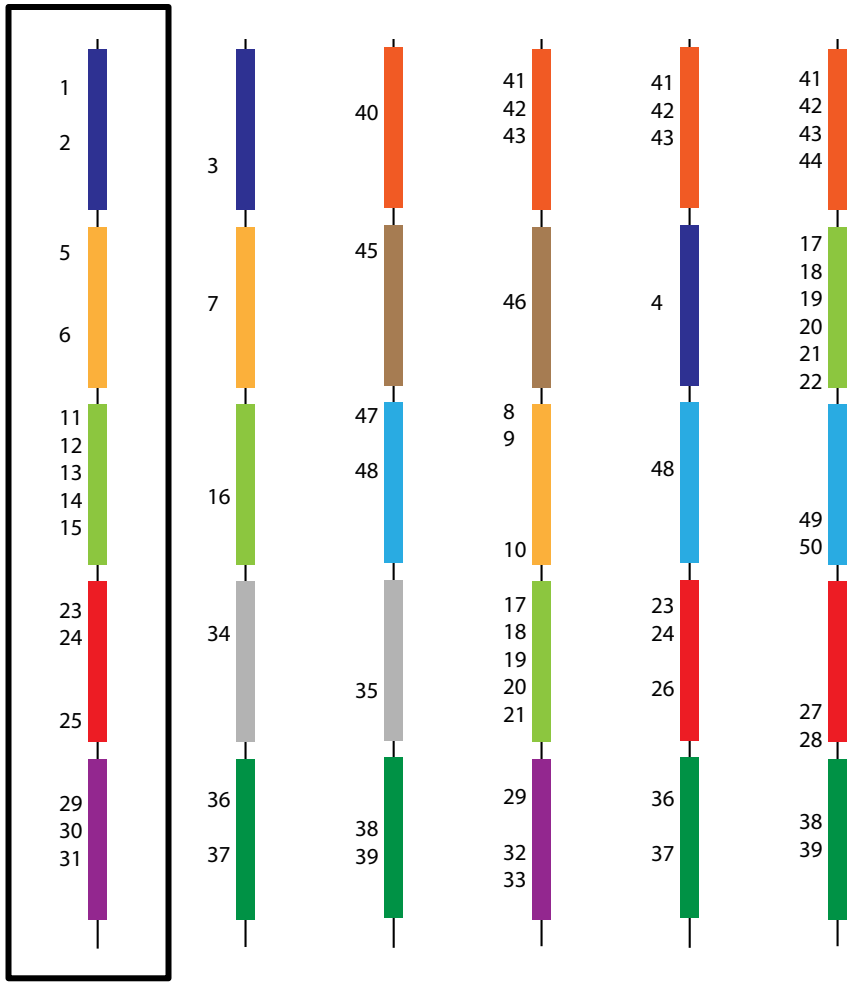
If we take perfect reads from the other genomes, and map them to this reference, how many of the 50 SNPs can we call?



Example of the problem

Look at the navy gene

We can call SNPs
1,2,3,4

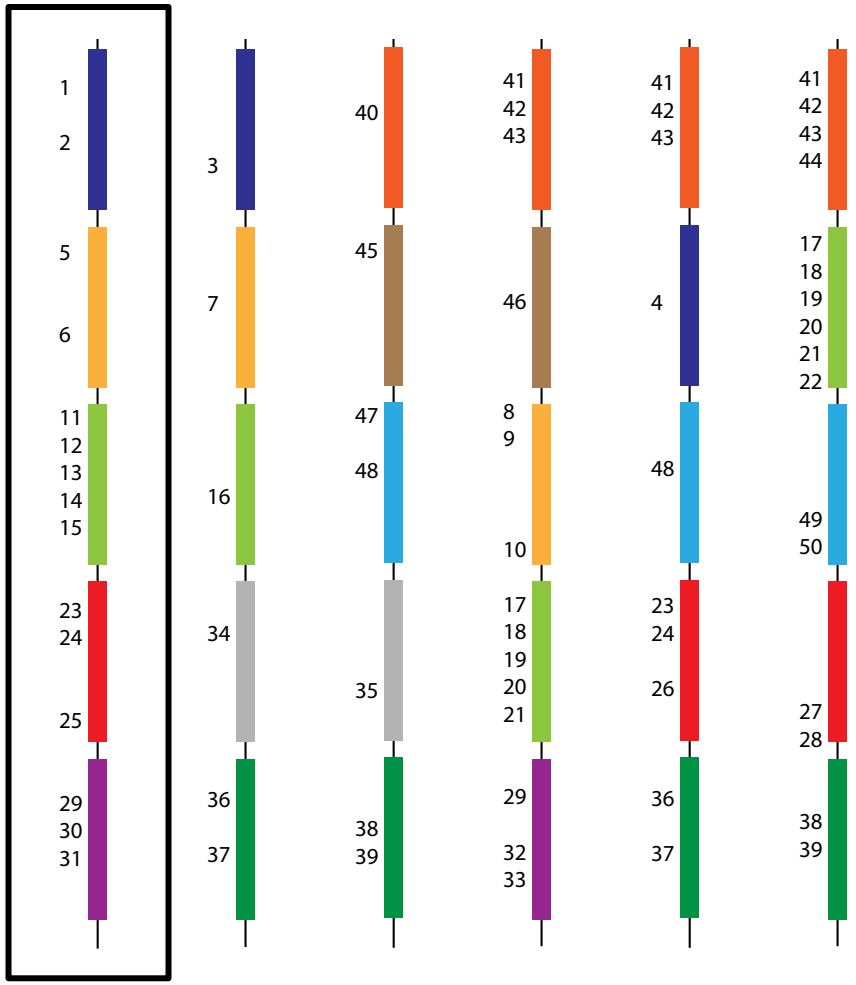


Total = 4

Example of the problem

Look at the yellow gene

We can call SNPs
5,6,7,8,9,10



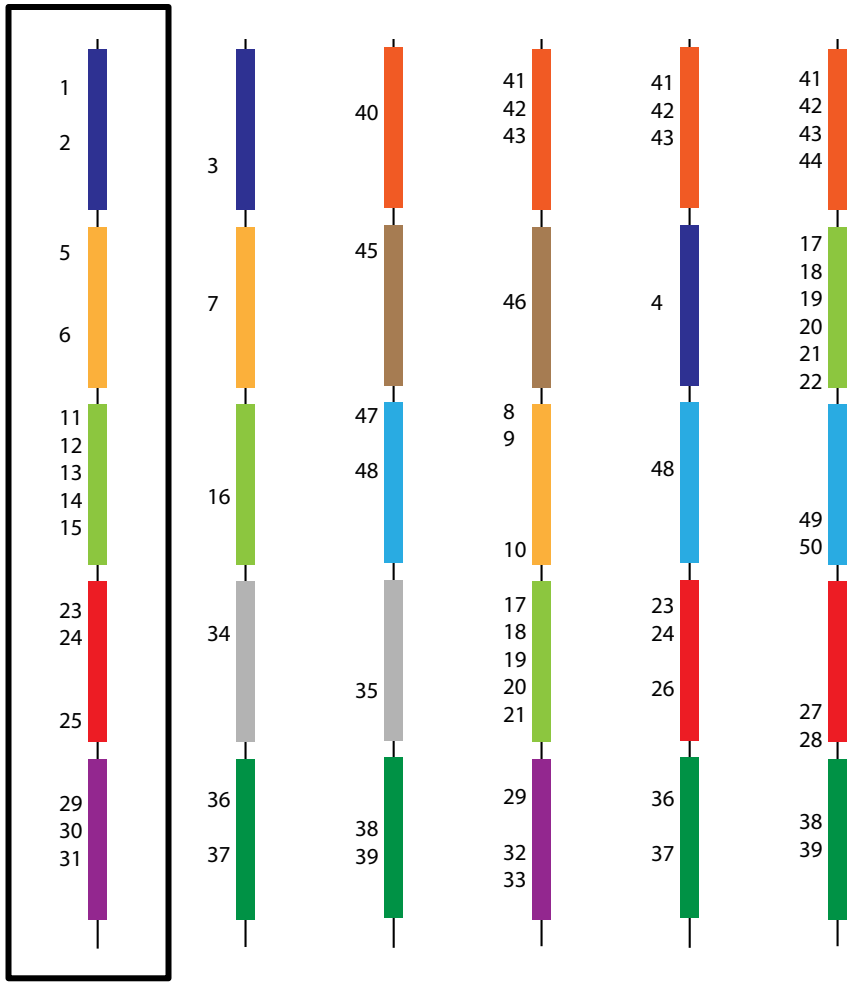
Total = 4+6

Example of the problem

Look at the green gene

We can call SNPs 11,12,13,14,15,16

However SNPs 17-22 are from a recombination event and are densely clustered. No reads map. So we cannot detect them.

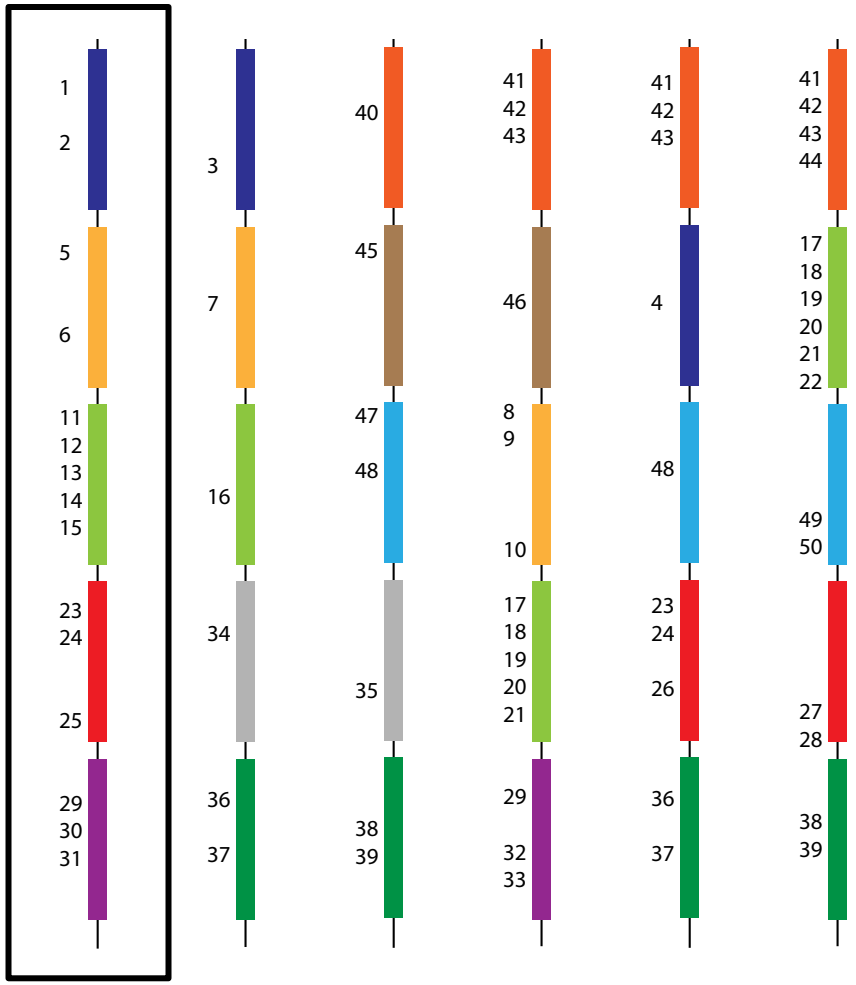


Total = 4+6+6

Example of the problem

Look at the red gene

We can call SNPs
23,24,25,26,27,28

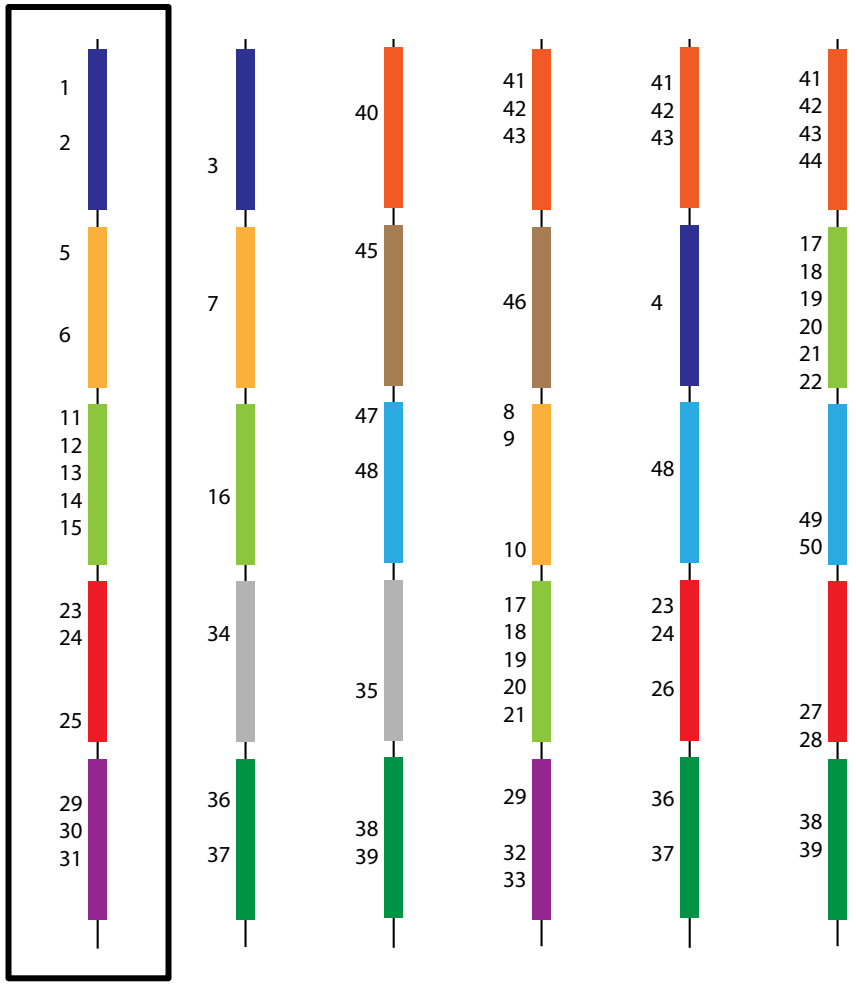


Total = 4+6+6+6

Example of the problem

Look at the purple gene

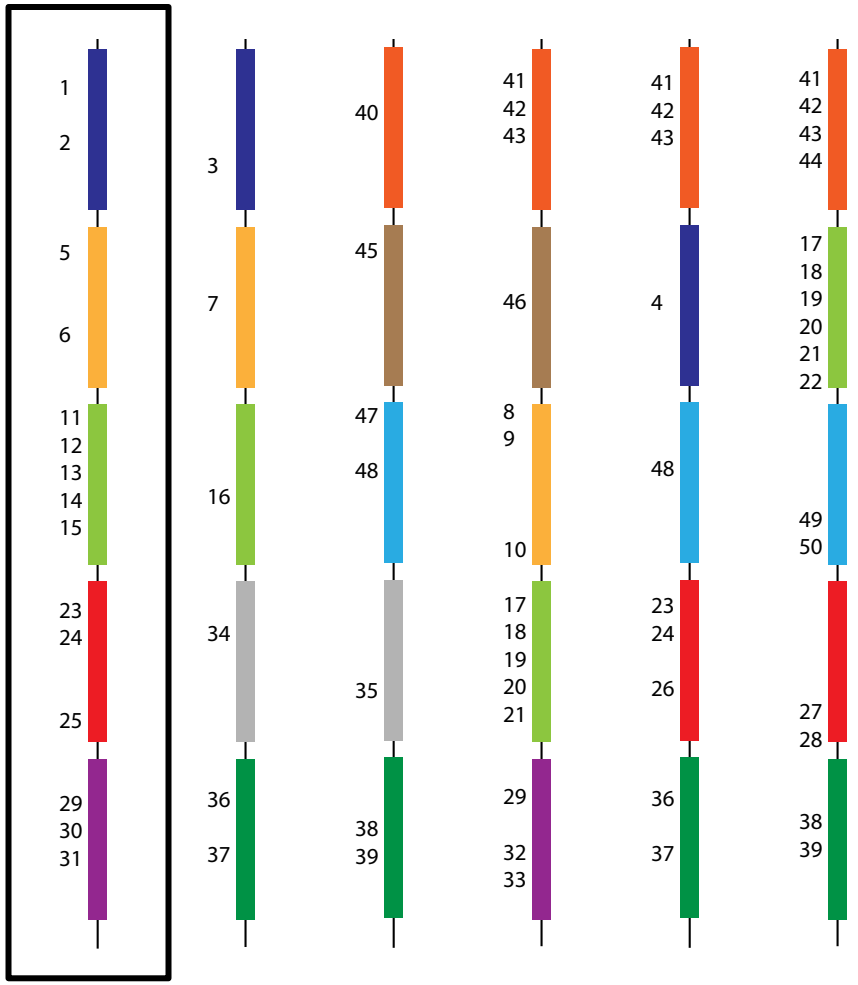
We can call SNPs
29,30,31,32,33



Total = 4+6+6+6+5

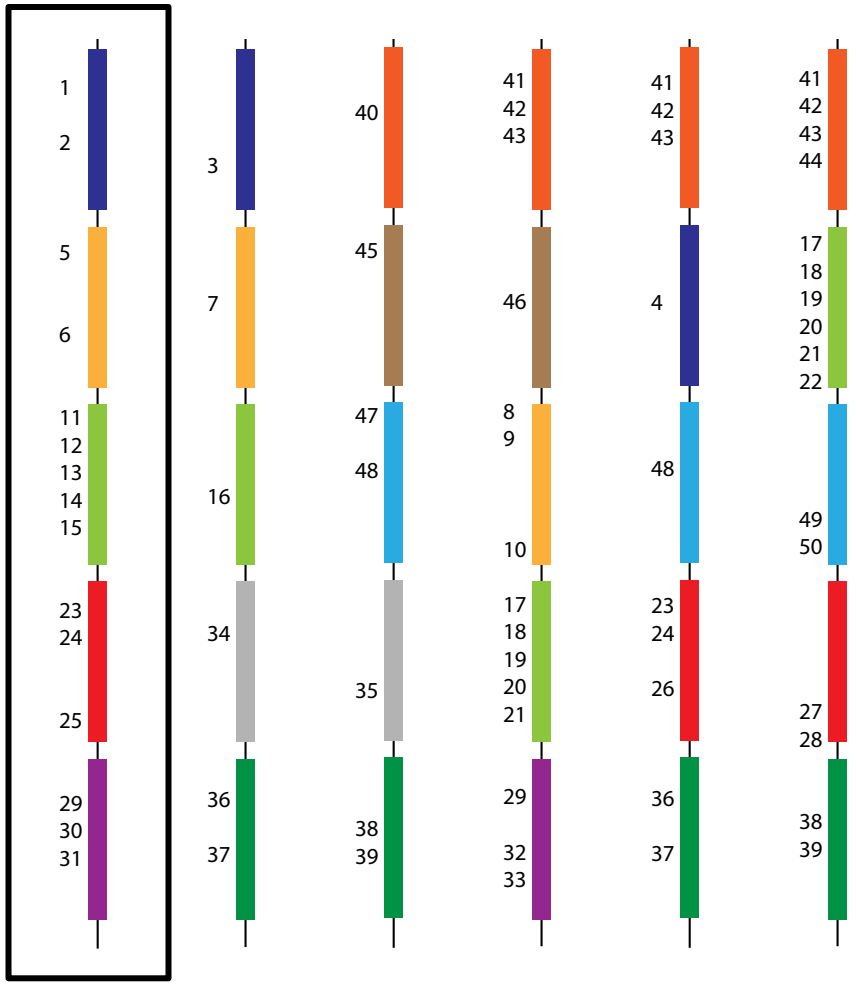
Example of the problem

We cannot detect SNPs on grey, orange brown, light blue or dark green genes



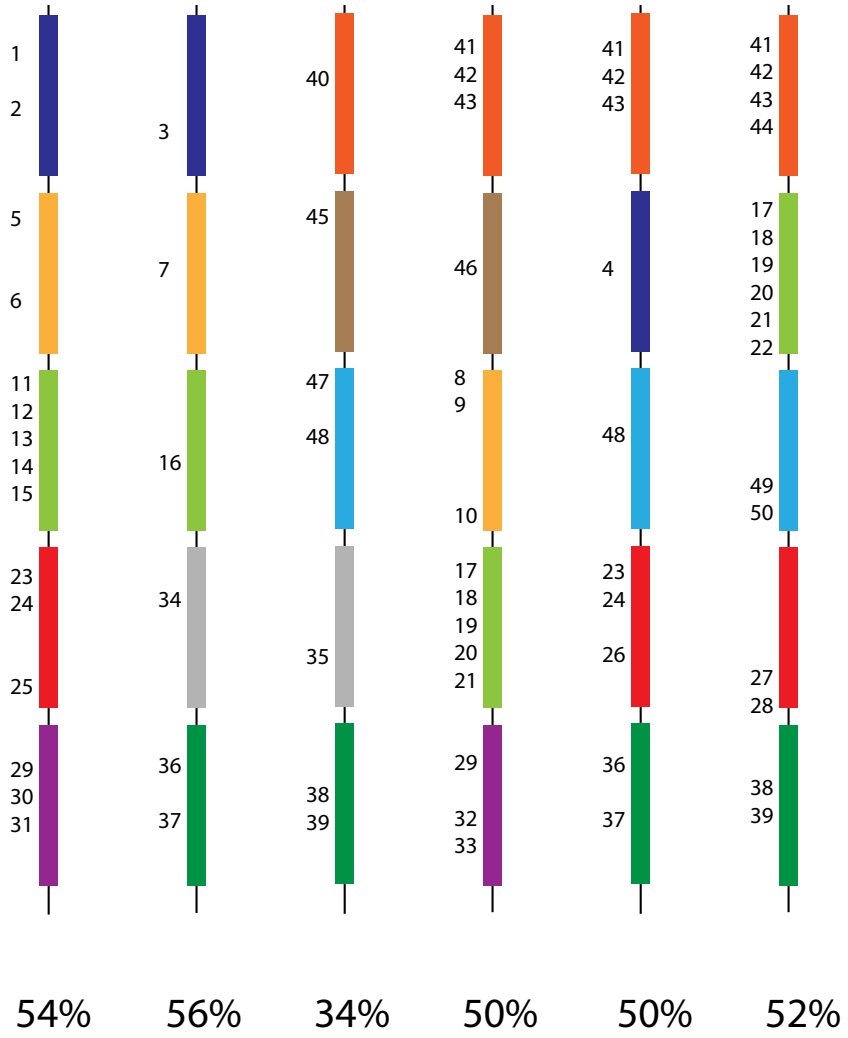
Total = 4+6+6+6+5

Example of the problem



Total = 4+6+6+6+5
= 27/50
= 54%

Example of the problem



The key problem

There is a lack of correlation between:

- detailed (core SNP/tree) distance
- coarse (repertoire) distance.

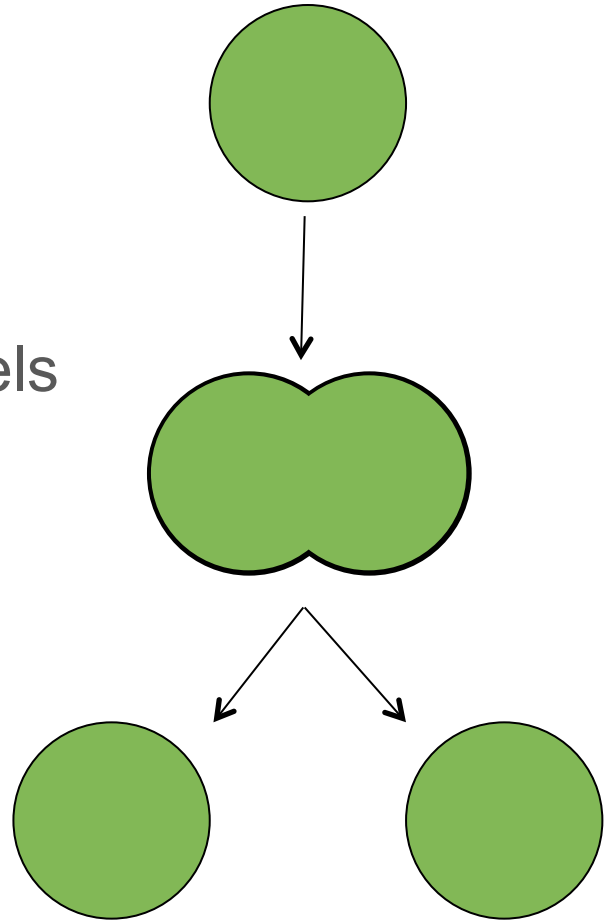
At present for diverse genomes:

- We cannot get SNP calls outside the core genome with reference-based variant calling
- Multiple sequence alignments do not scale to many whole genomes (plus is nightmare to determine if a SNP in one place on one genome, is “the same” as another SNP at a different place in another)

Bacterial Inheritance

Vertically inherited variation:

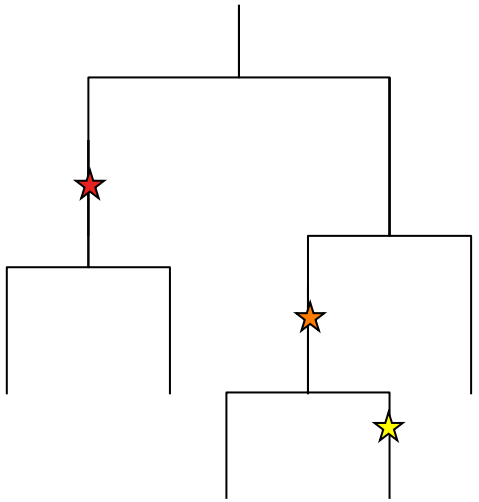
- Errors during replication -> SNPs
- Strand slippage during replication ->small indels
- Errors due to DNA damage and repair -> SNPs and indels

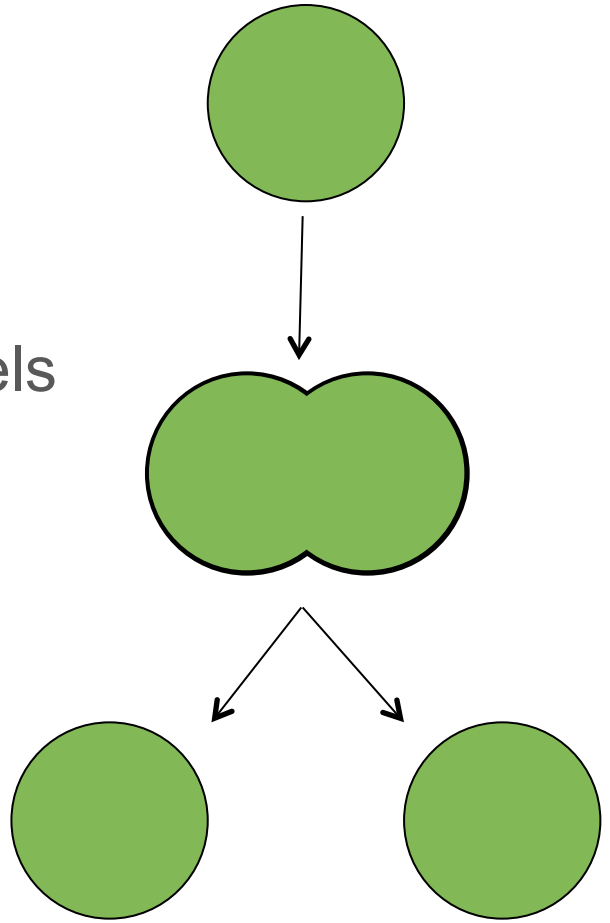


Bacterial Inheritance

Vertically inherited variation:

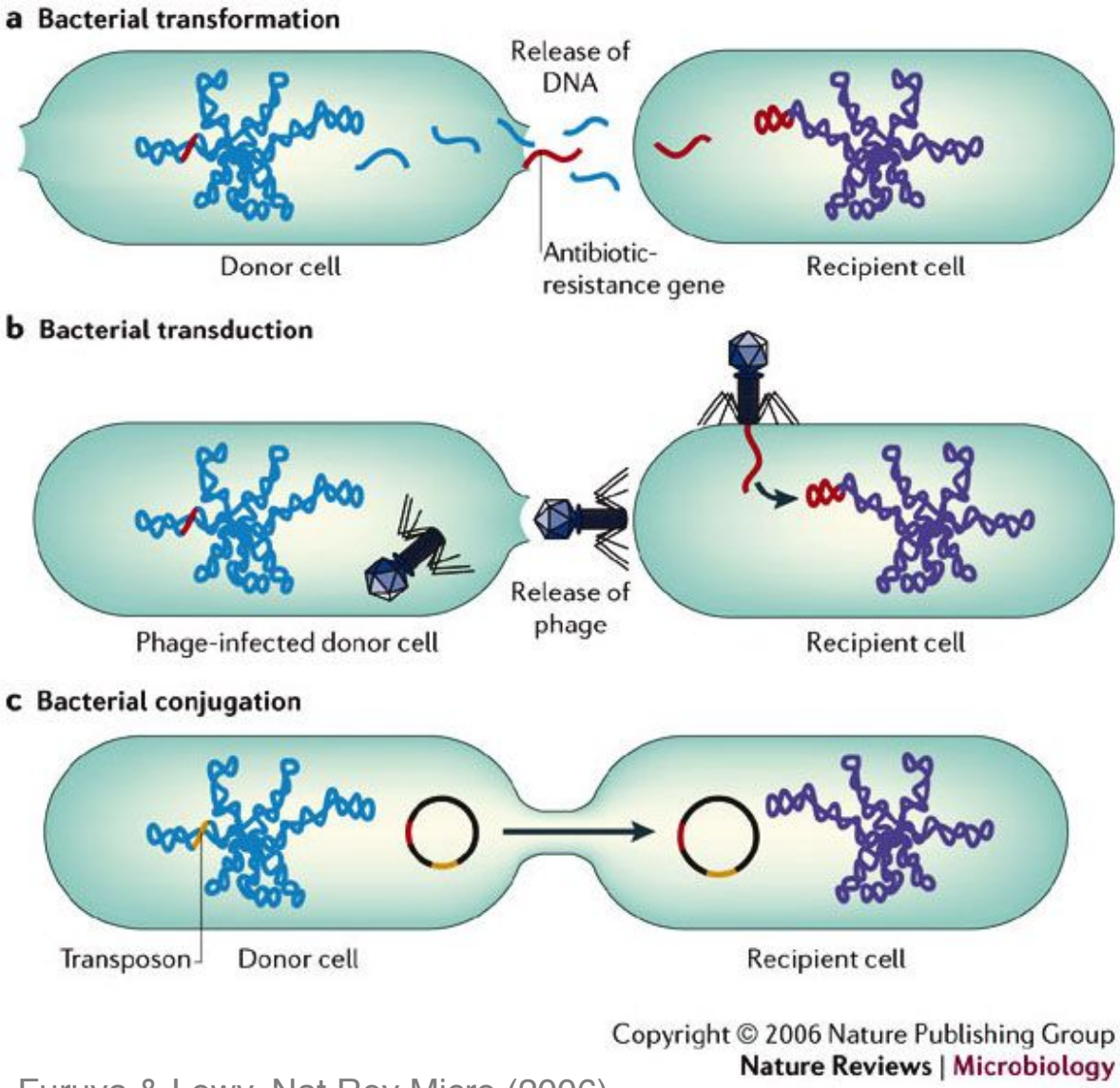
- Errors during replication -> SNPs
- Strand slippage during replication ->small indels
- Errors due to DNA damage and repair -> SNPs and indels





Bacterial Inheritance

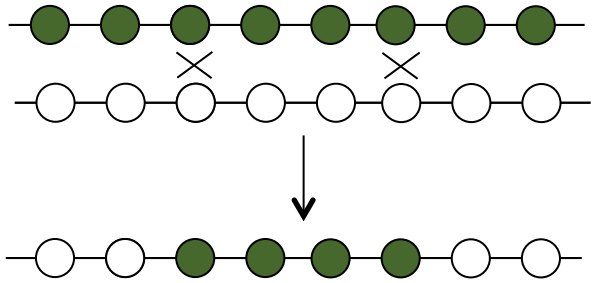
Horizontally acquired variation:



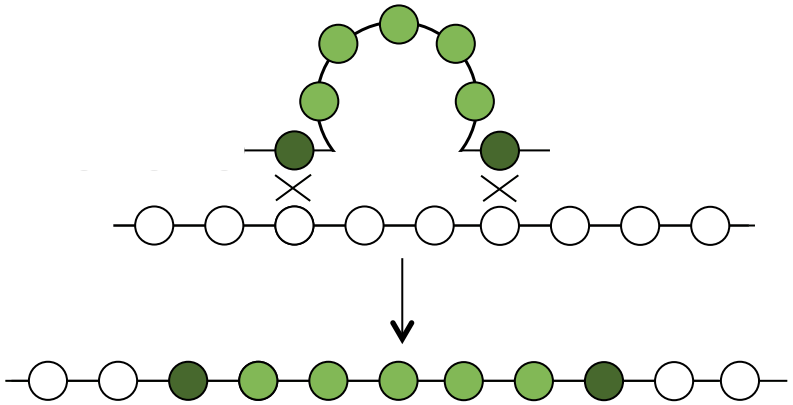
Furuya & Lowy, Nat Rev Micro (2006)

Bacterial Inheritance

Horizontally acquired variation:



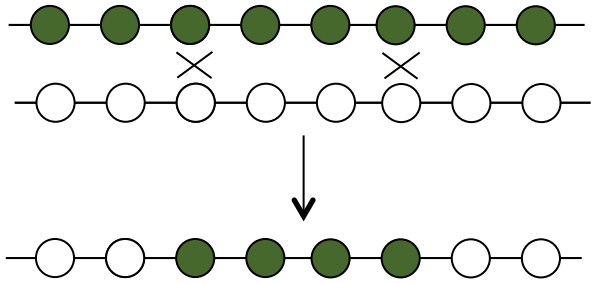
Homologous recombination



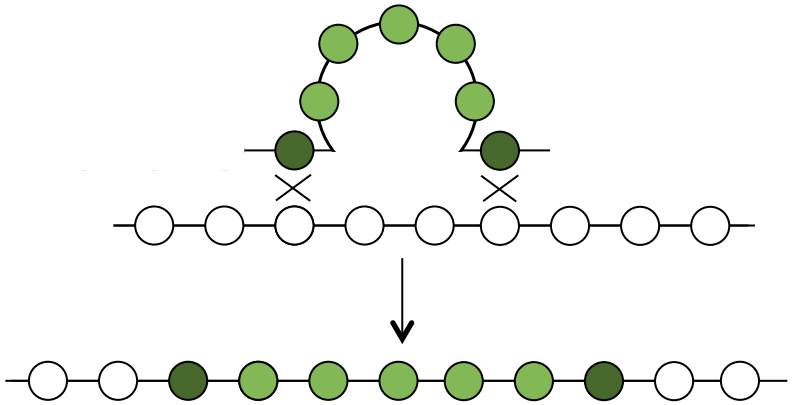
Allelic recombination and HGT

Bacterial Inheritance

Horizontally acquired variation:



Homologous recombination



Allelic recombination and HGT

Locally (within genes) sequences look like mosaics of those seen before

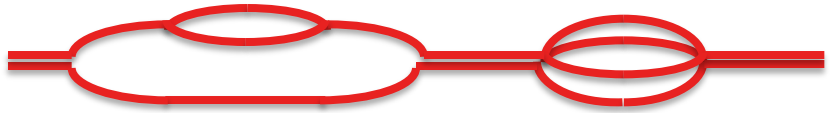
Globally genomes look like mosaics of those seen before

Goals

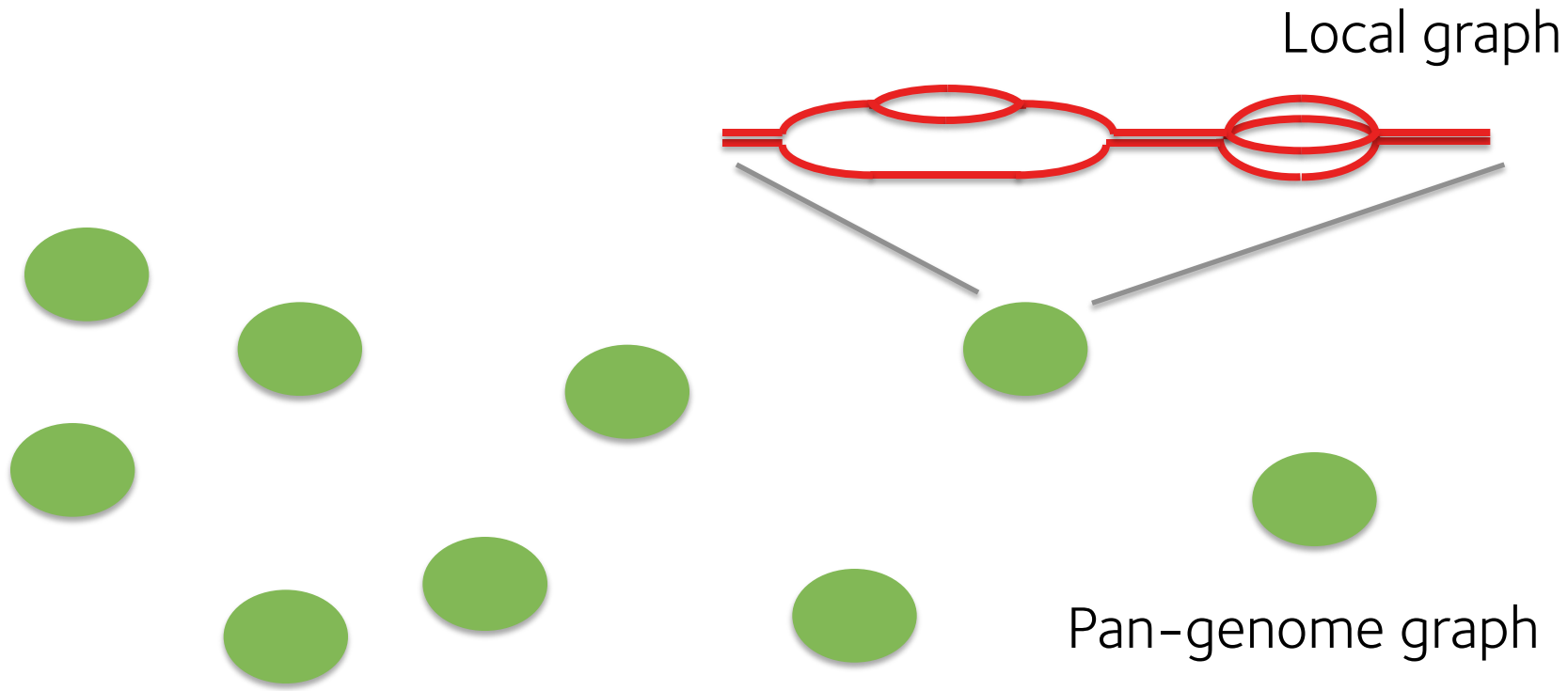
- Detect SNP variation between genomes in any gene/intergenic region shared between them
- Detect gene/allele presence in variety of contexts
- Compatible with long Nanopore or short Illumina reads
- Allow analysis of genome organization
- Flexible enough to cope with plasmid/phage/MGE
- Extensible to mixed read datasets

A Pangenome Reference Graph (PanRG)

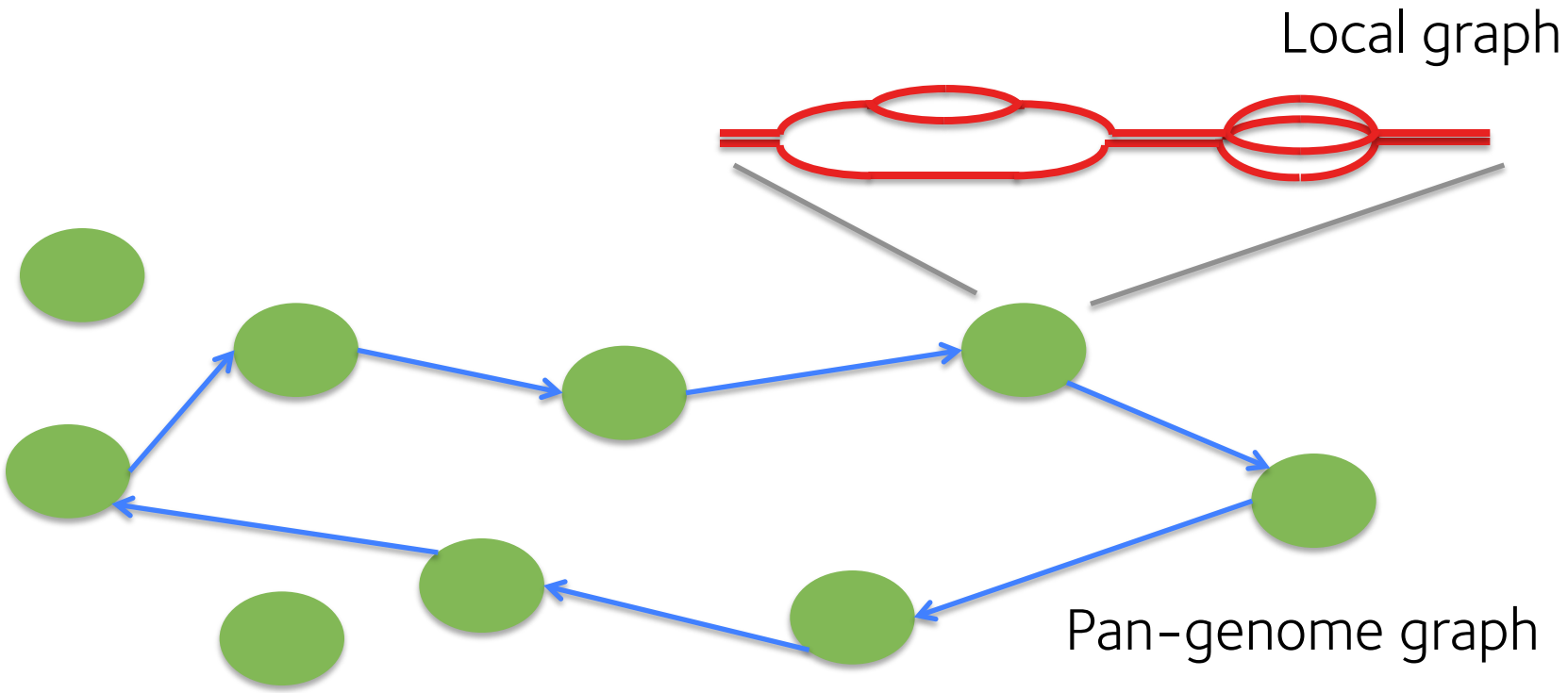
Local graph



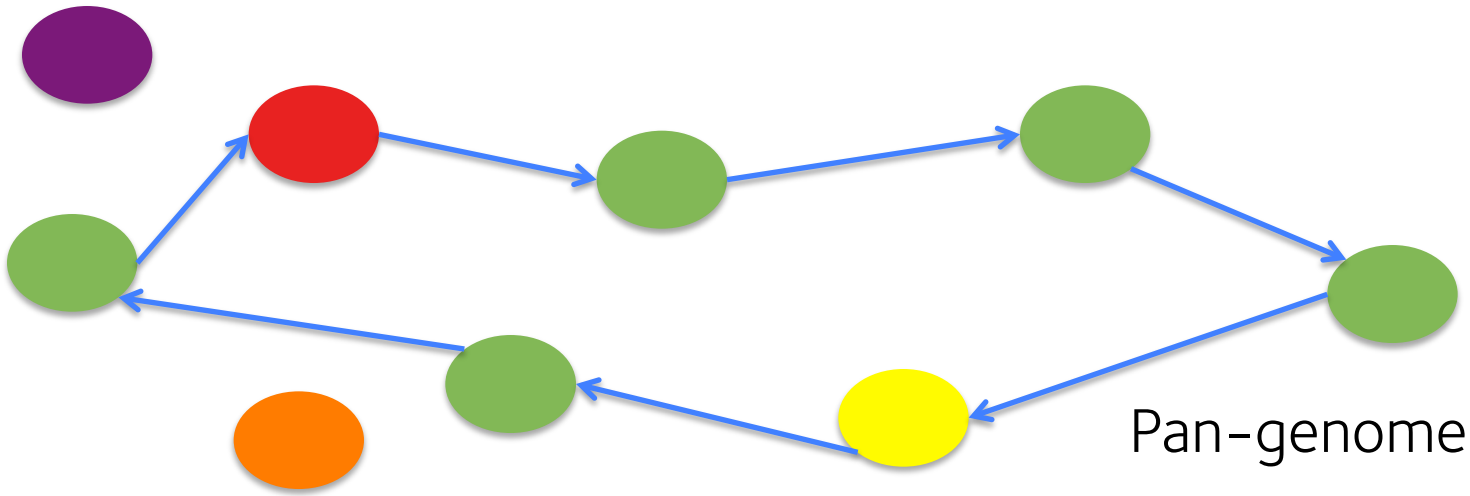
A Pangenome Reference Graph (PanRG)



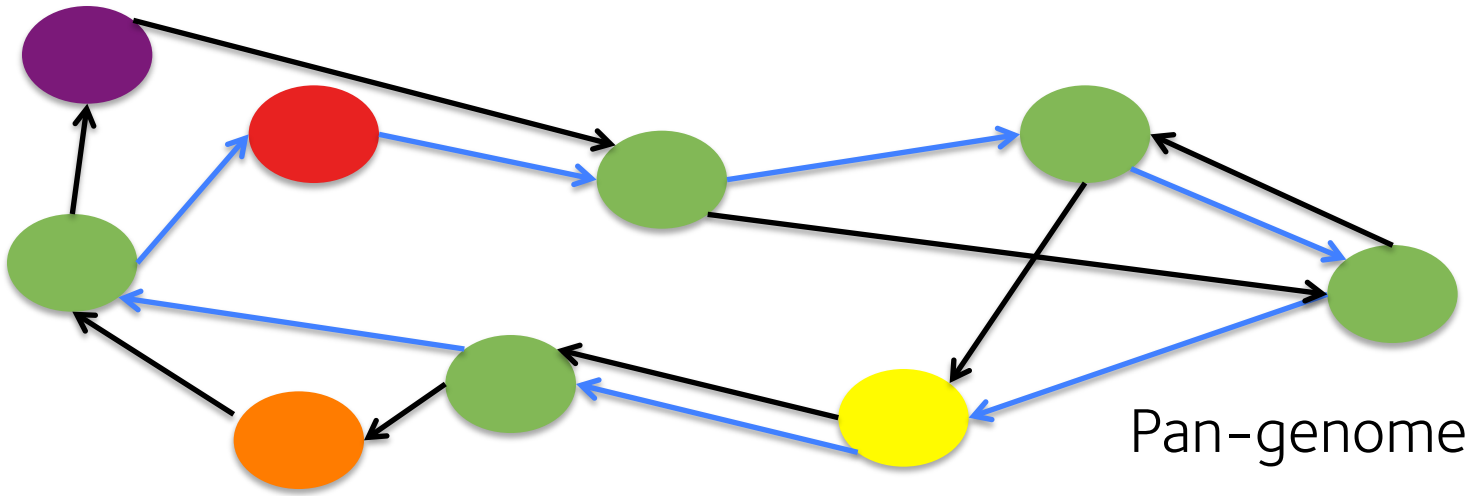
A Pangenome Reference Graph (PanRG)



Comparison with Pandora

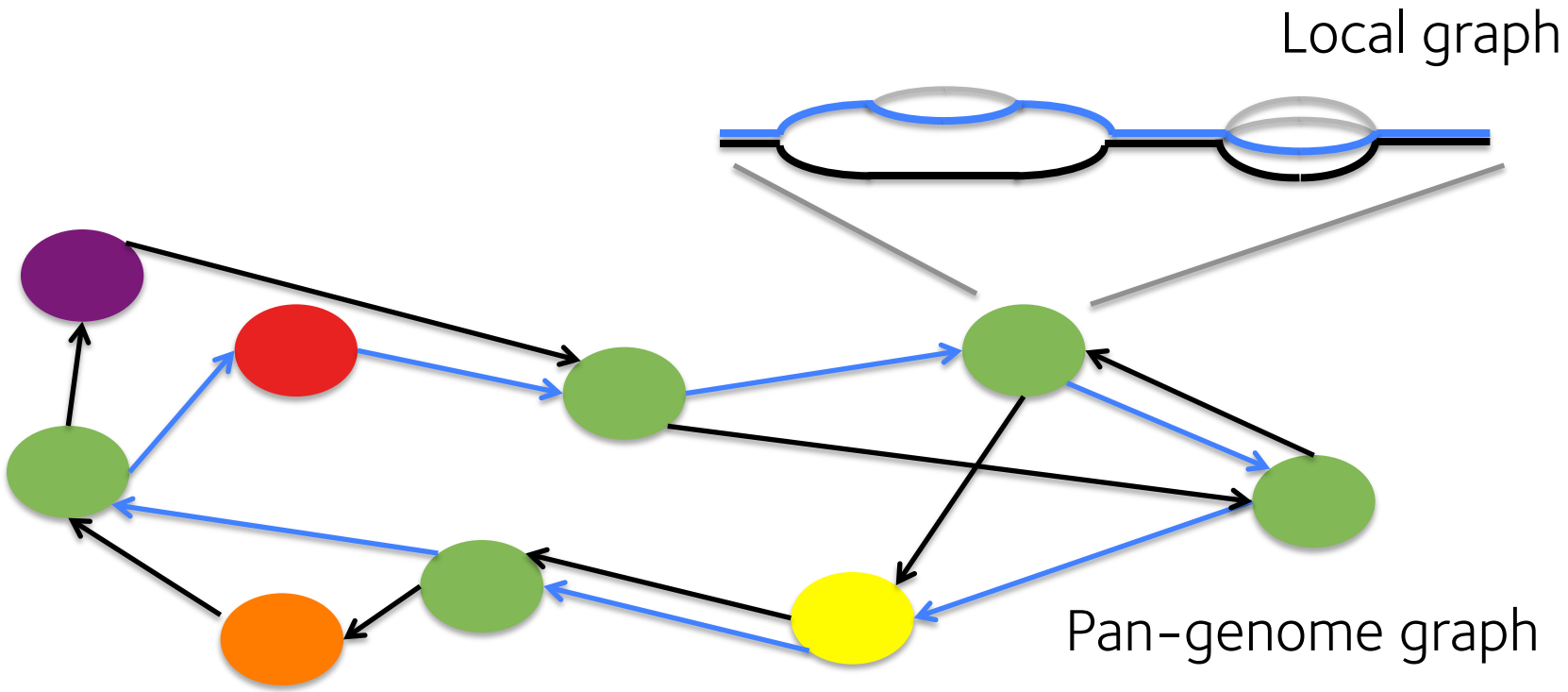


Comparison with Pandora

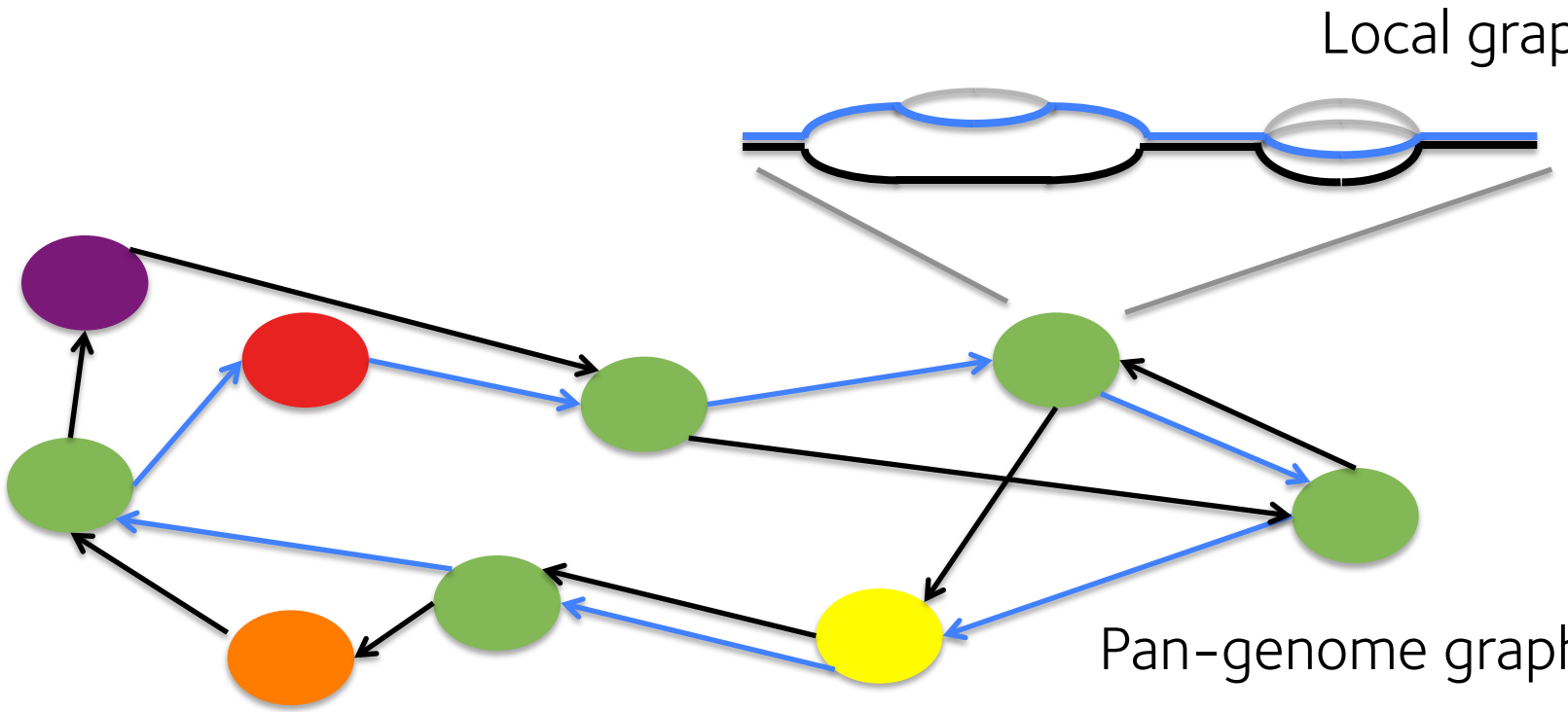


Pan-genome graph

Comparison with Pandora

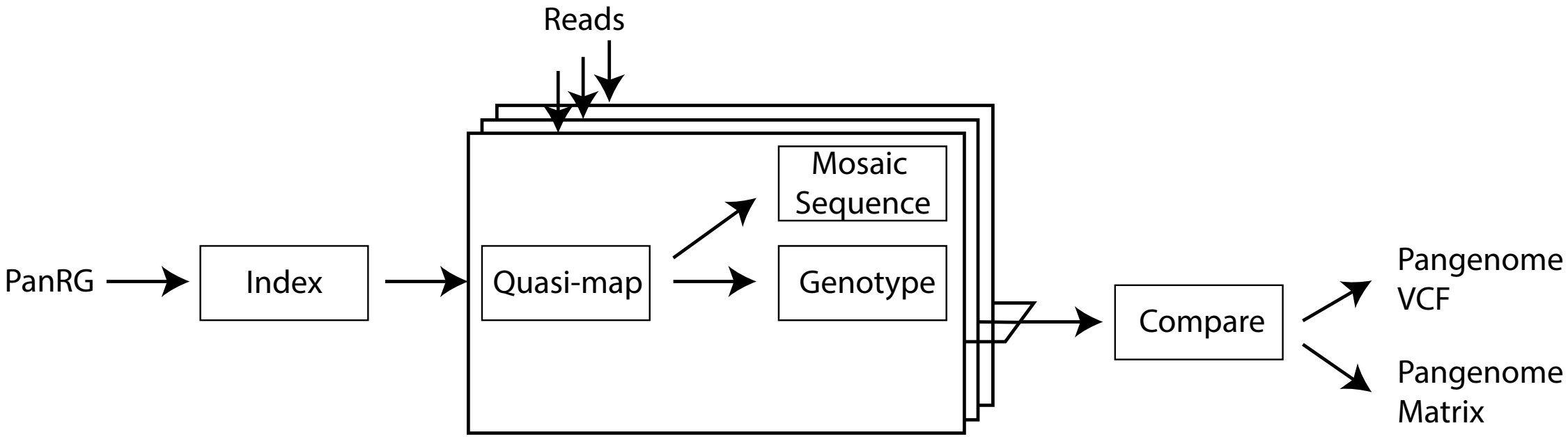


Comparison with Pandora

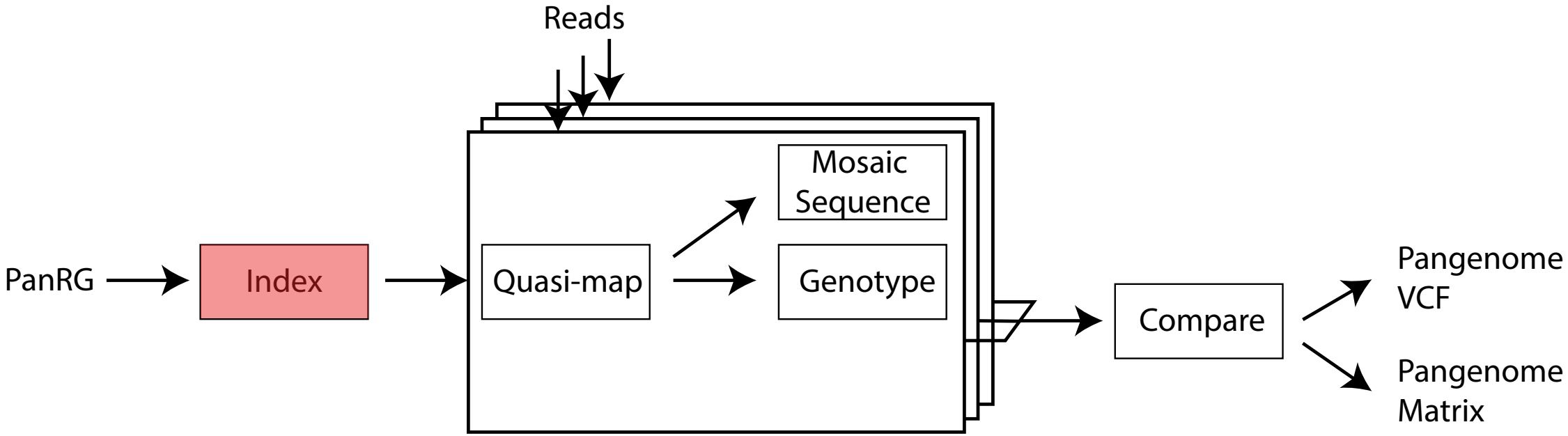


We choose the best reference path for each gene!

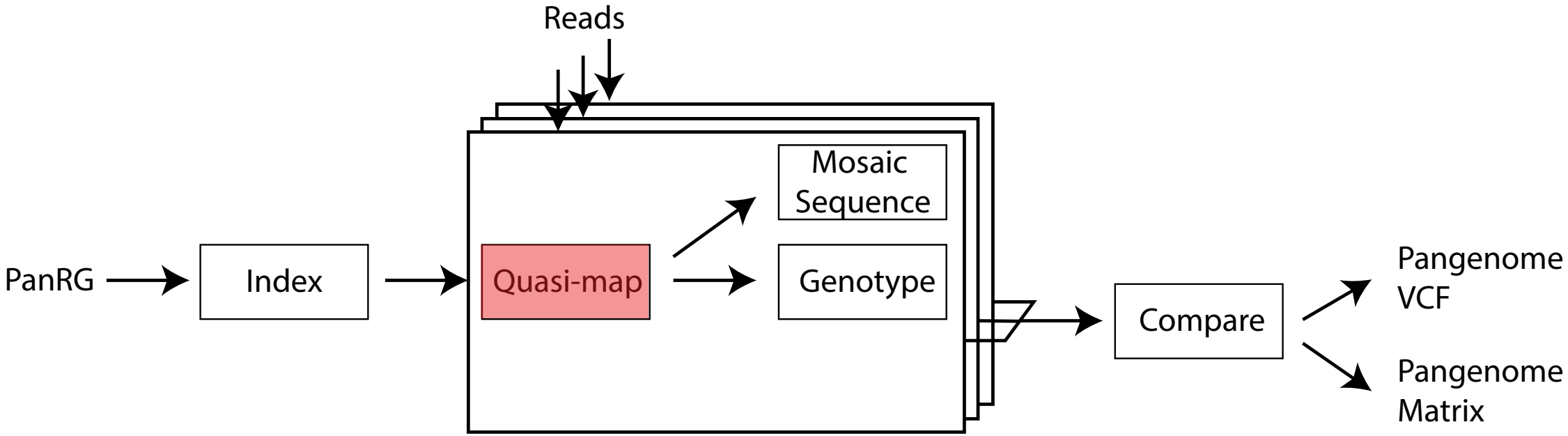
Pandora workflow



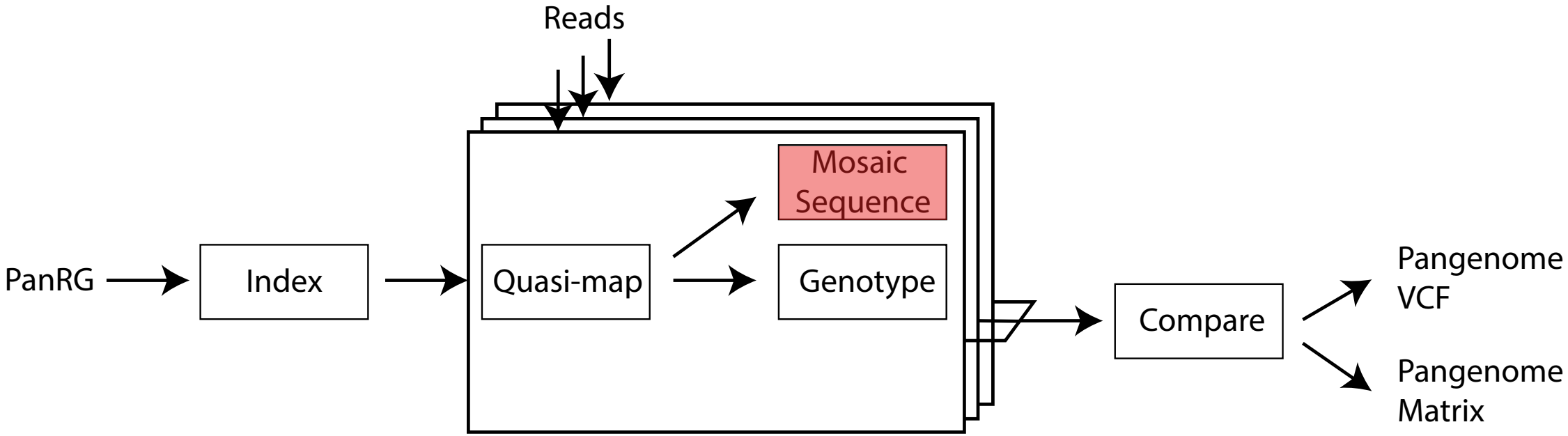
Pandora workflow



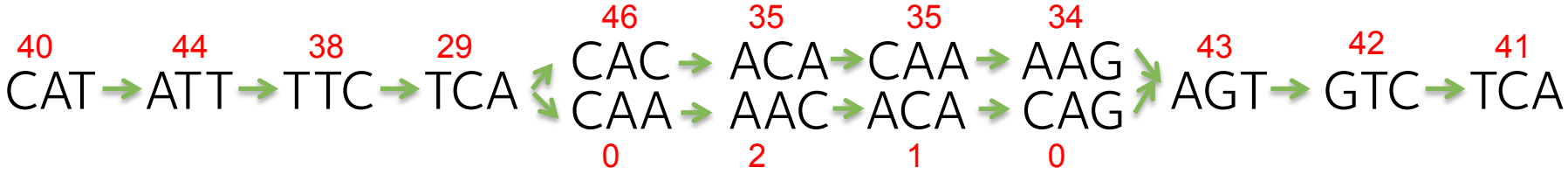
Pandora workflow



Pandora workflow

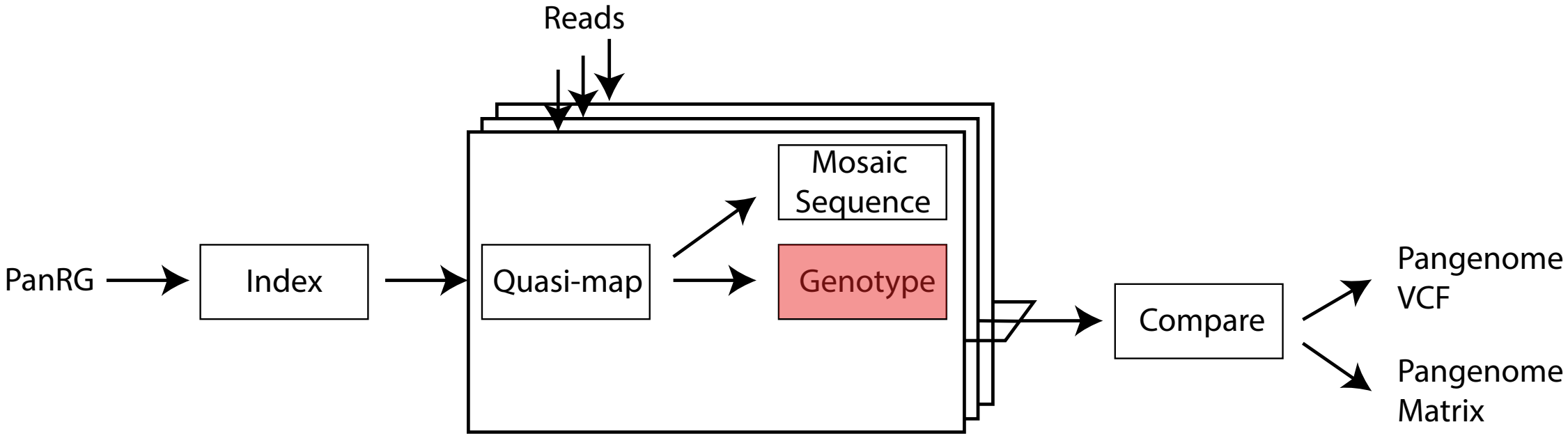


Mosaic sequence inference

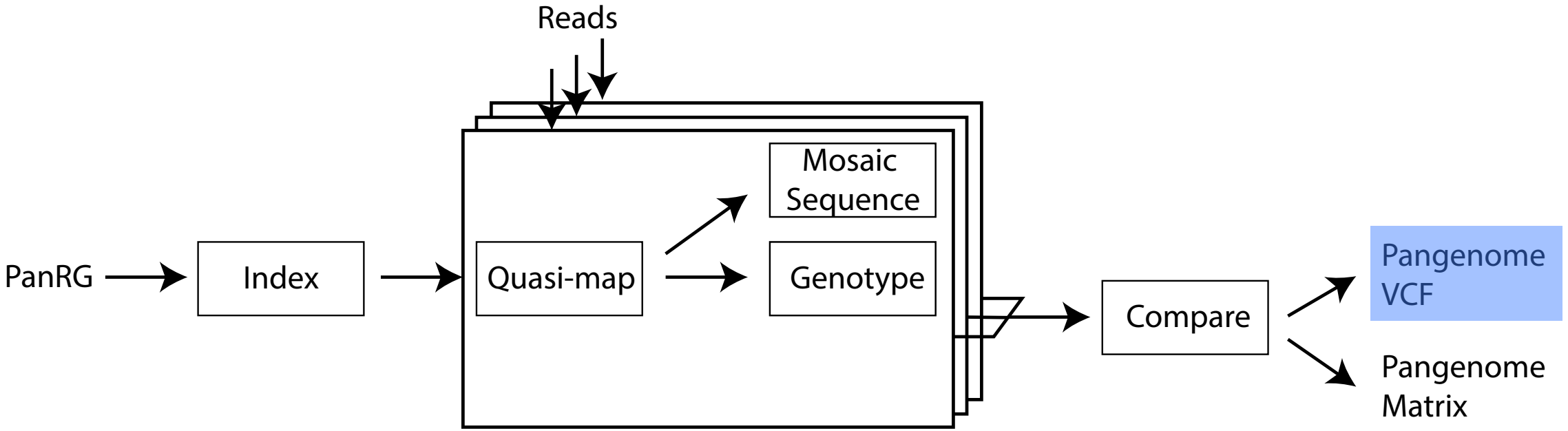


Pick the path with maximum likelihood

Pandora workflow



Pandora workflow



Experiment: Compare 4 diverse *E. coli*

- Take 4 E.coli strains:
  - 2 from a cardiothoracic unit (human) outbreak (ST216)
  - 1 reference strain (ST73)
  - 1 from cattle faeces (ST3858)
- Both Nanopore and Illumina data (300X) and high quality Illumina-polished PacBio assemblies (“truth”)

Experiment: Compare 4 diverse *E. coli*

- Build a PanRG for *E. coli*
  - Construct graphs for 23052 genes built from 350 RefSeq genomes using the PanX tool from Neher lab (Ding et al)
  - Construct graphs for 14374 intergenic regions from 228 E. coli from ST131 using the Piggy tool from Harry Thorpe (Ed Feil's lab).
  - 58.9% of gene graphs and 43.8% of intergenic regions consist of just a single sequence, no variation

Comparators

Nanopolish

- Only published variant caller on Nanopore data.
- Used in Ebola outbreak

Snippy (bwa+freebayes)

- Standard tool. Illumina-only. Gives us an illumina baseline.

Try 10 different reference assemblies for variant calling

Comparators

~~**Nanopolish**~~ *(Don't have signal level data for all samples)*

- ~~Only published variant caller on Nanopore data.~~
- ~~Used in Ebola outbreak~~

Snippy (bwa+freebayes)

- Standard tool. Illumina-only. Gives us an illumina baseline.

Try 10 different reference assemblies for variant calling

Metrics for evaluation

- Do all pairwise alignments between these 4, and use mummer/dnadiff to find a set of high quality SNPs between them.

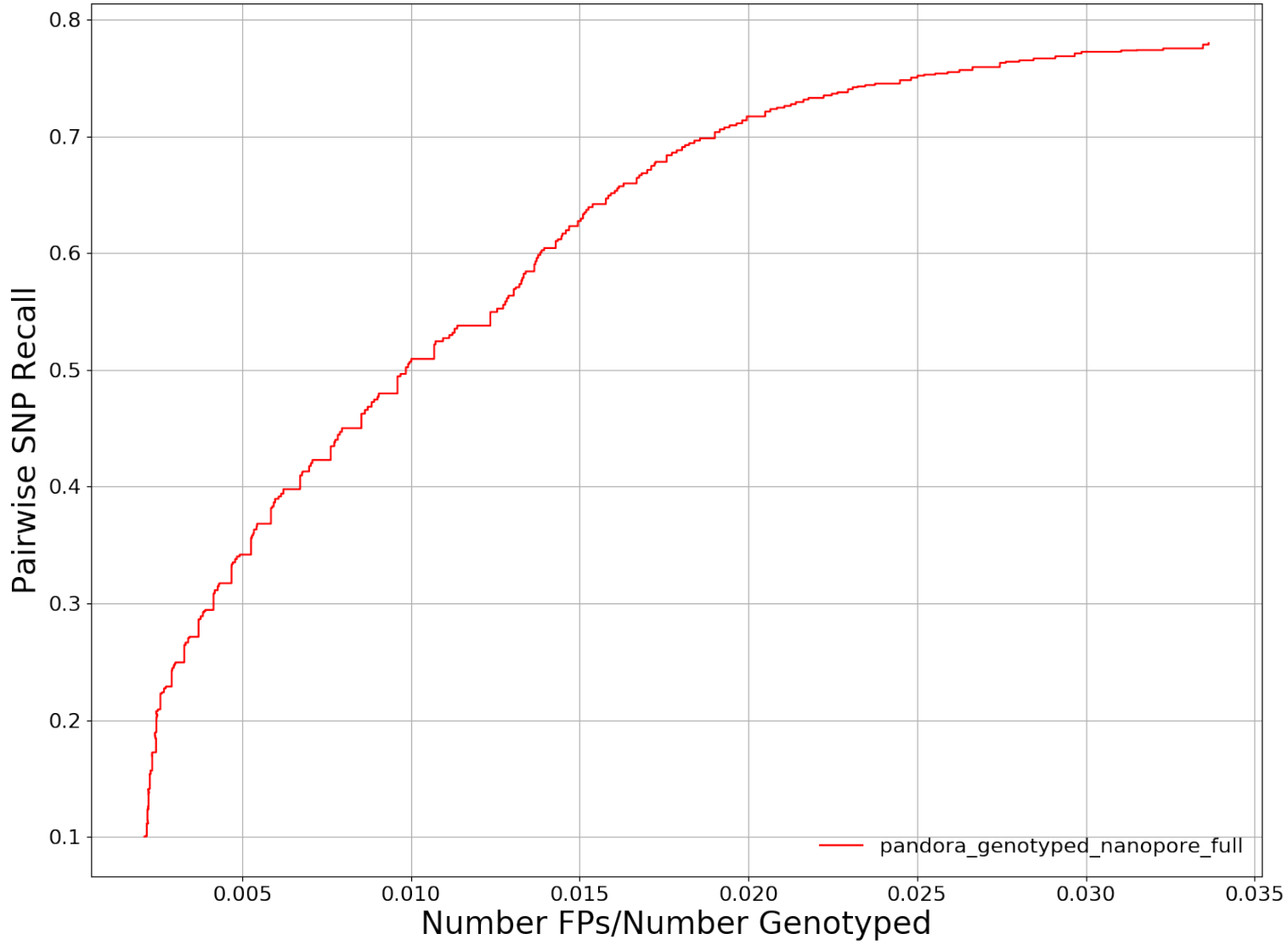
- Proxy for “recall”: what % of **all the pairwise dnadiff SNP**s are found?

  Note: If a SNP difference is found in 3 pairs, it is counted 3 times – weighted towards higher frequency variants.

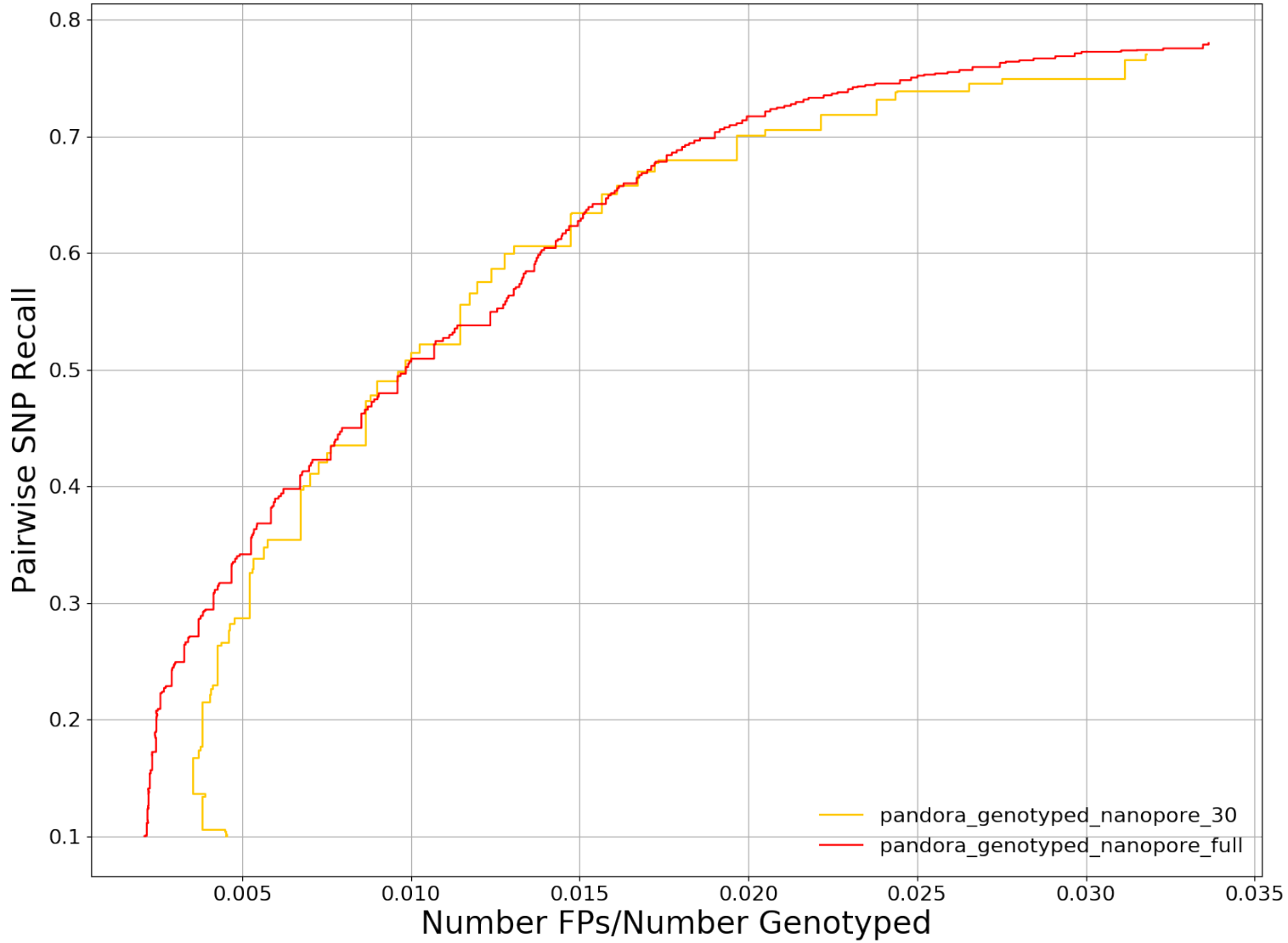
  Why do this? Hard to be sure if one SNP==another.

- Precision: what % of all calls made are correct (map variant and flanks to truth assembly)

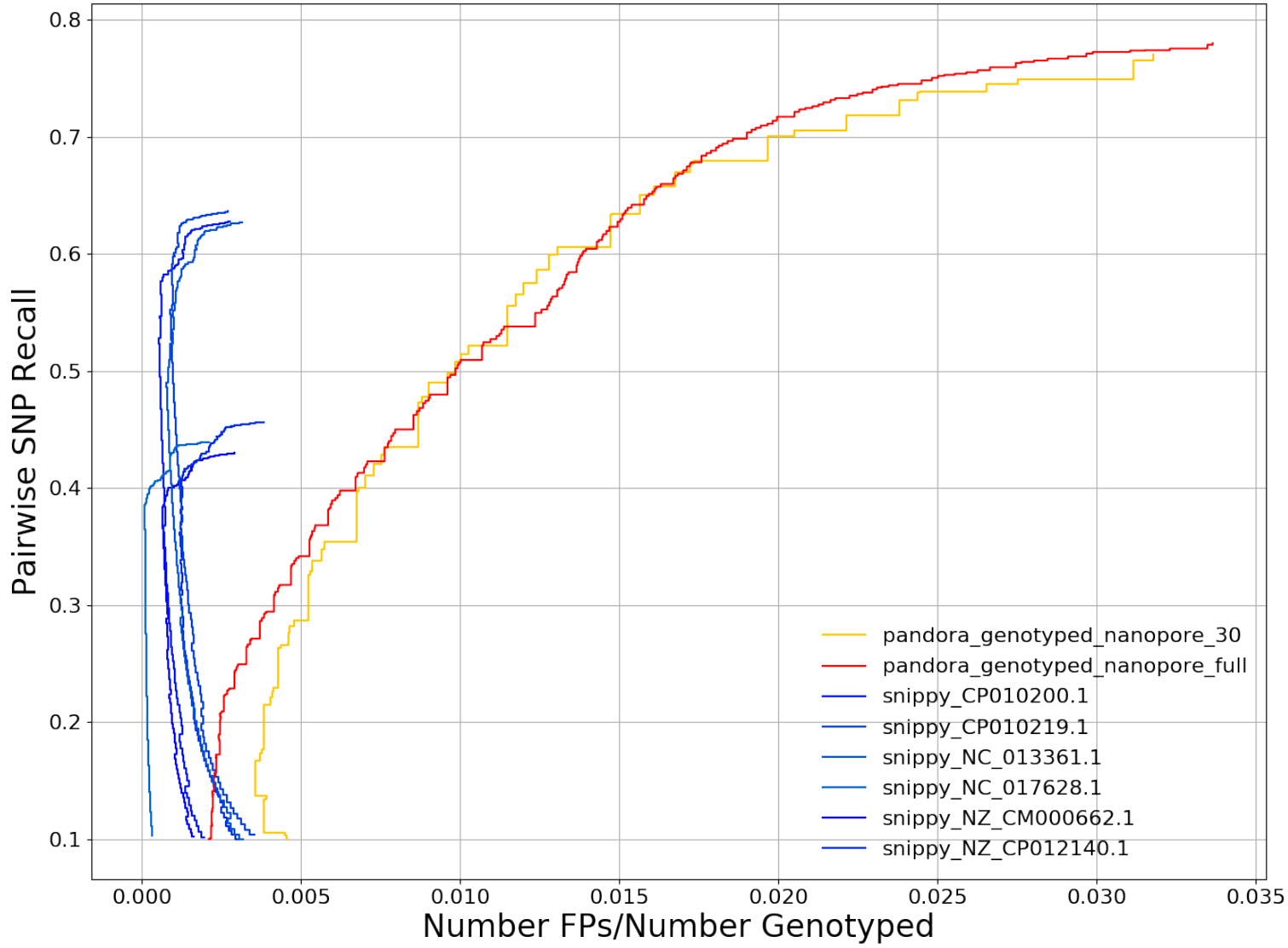
2 samples (1 human, 1 cattle)



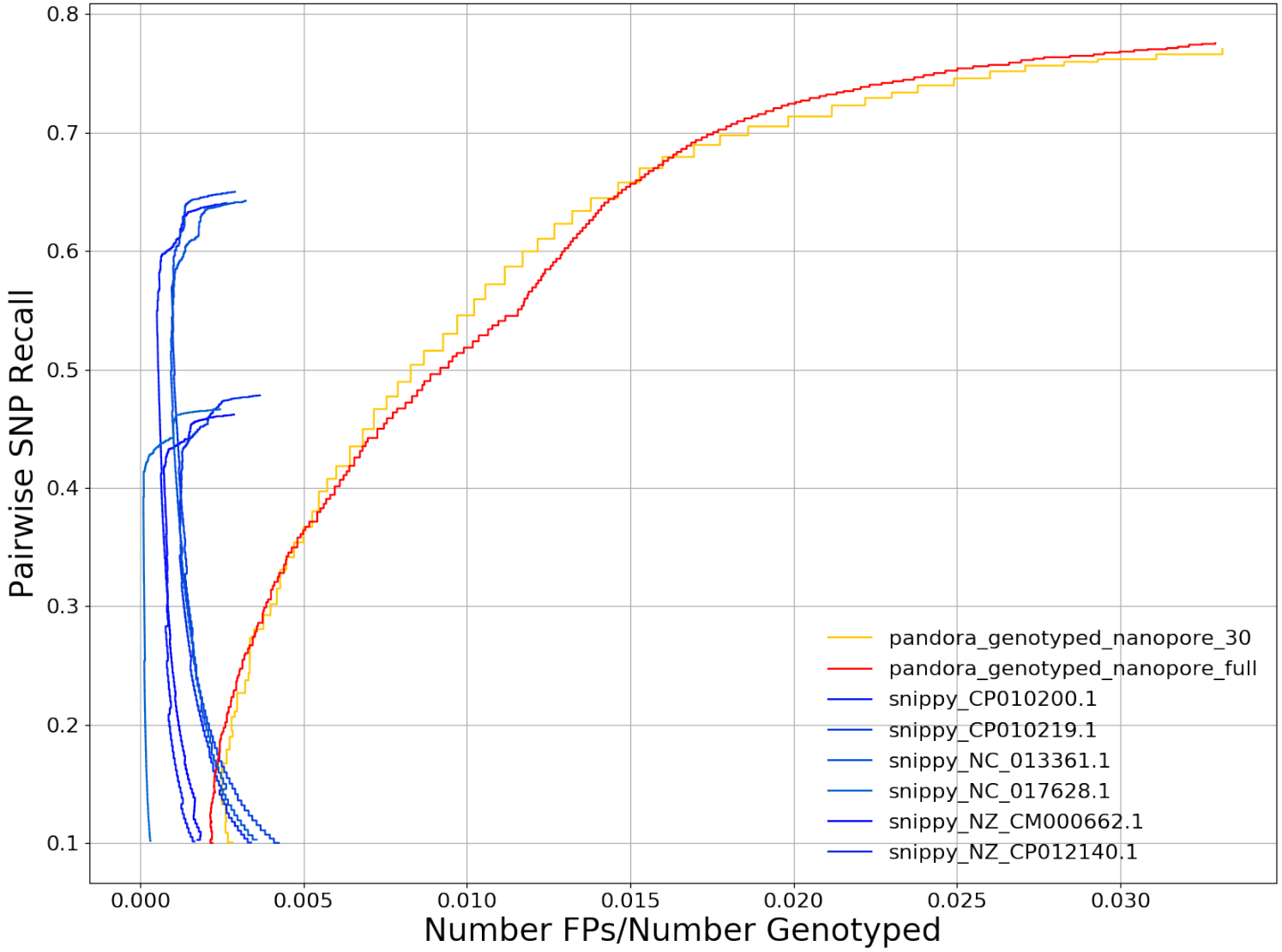
2 samples (1 human, 1 cattle)



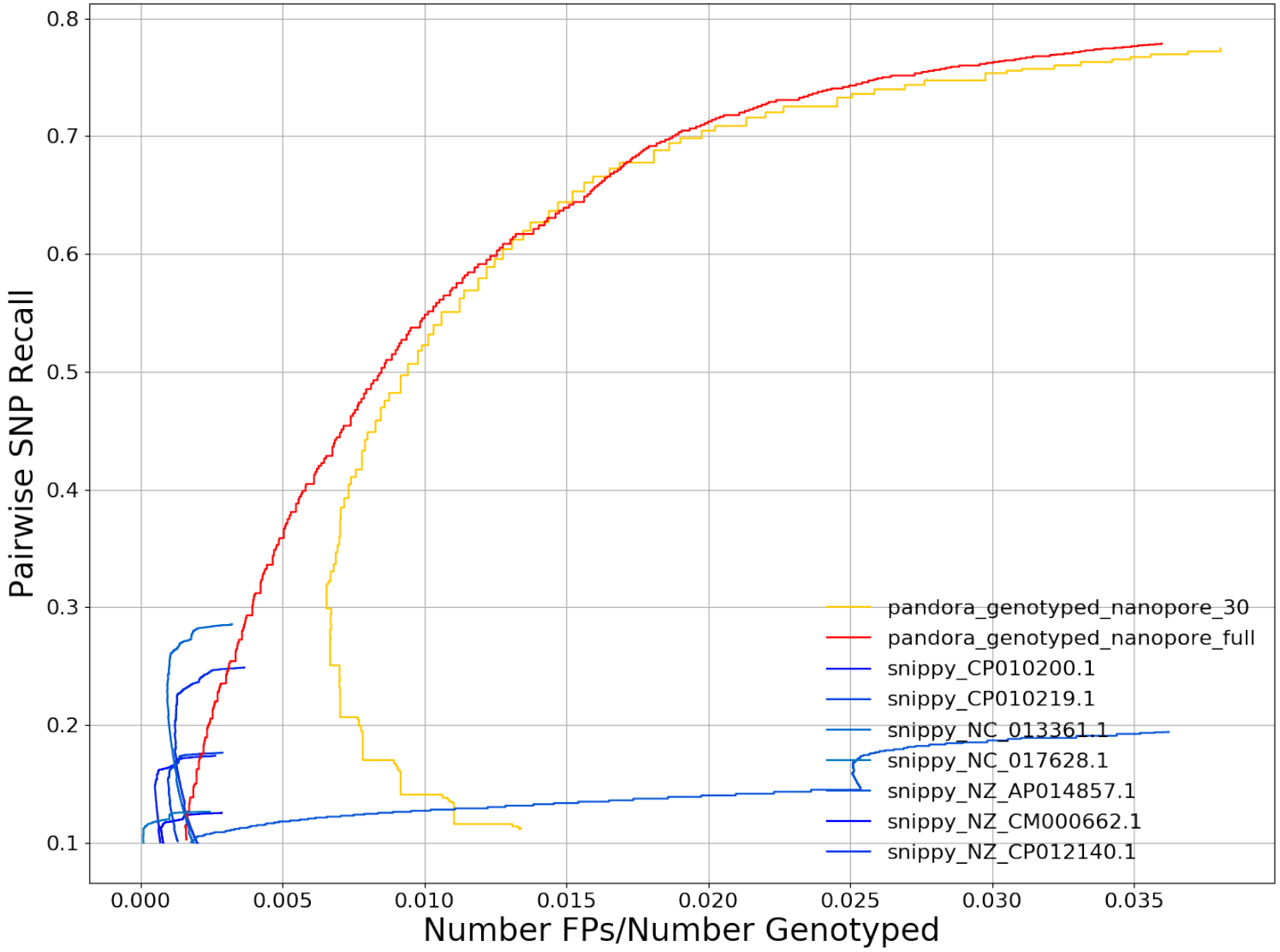
2 samples (1 human, 1 cattle)



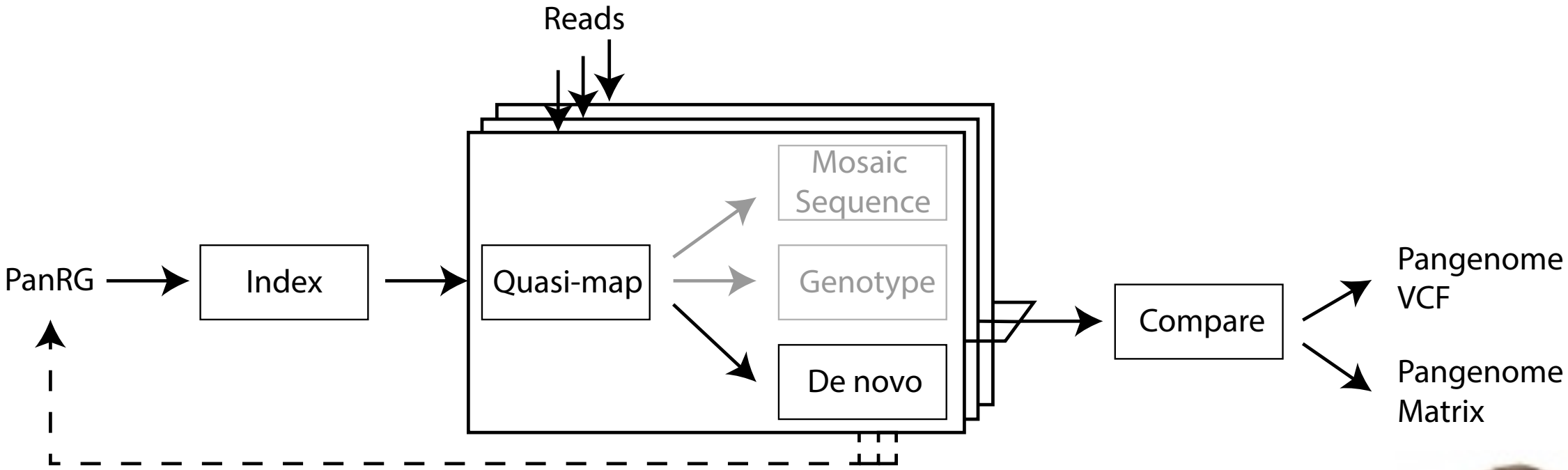
3 samples (2 human, 1 cattle)



4 samples (2 human, 1 cattle, 1 reference)



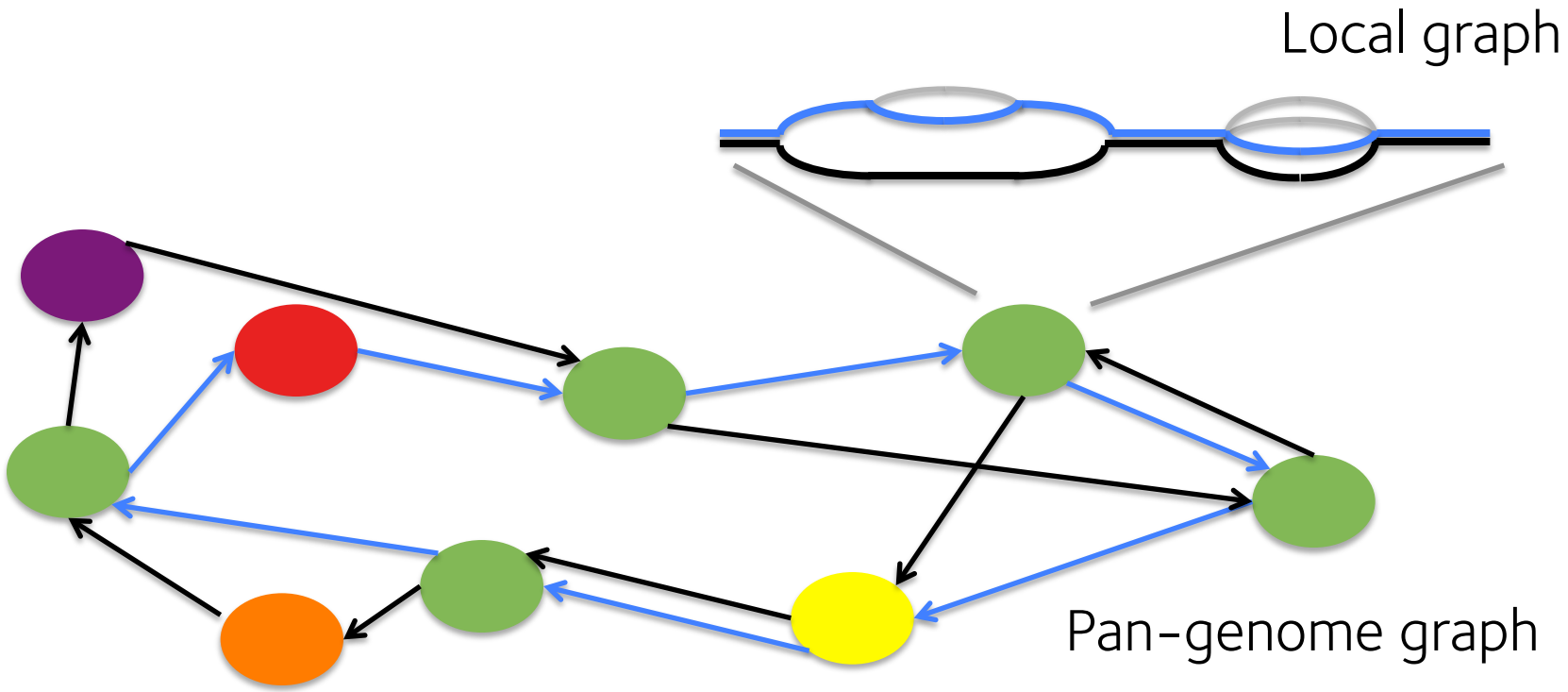
Add a local *de novo* assembly step



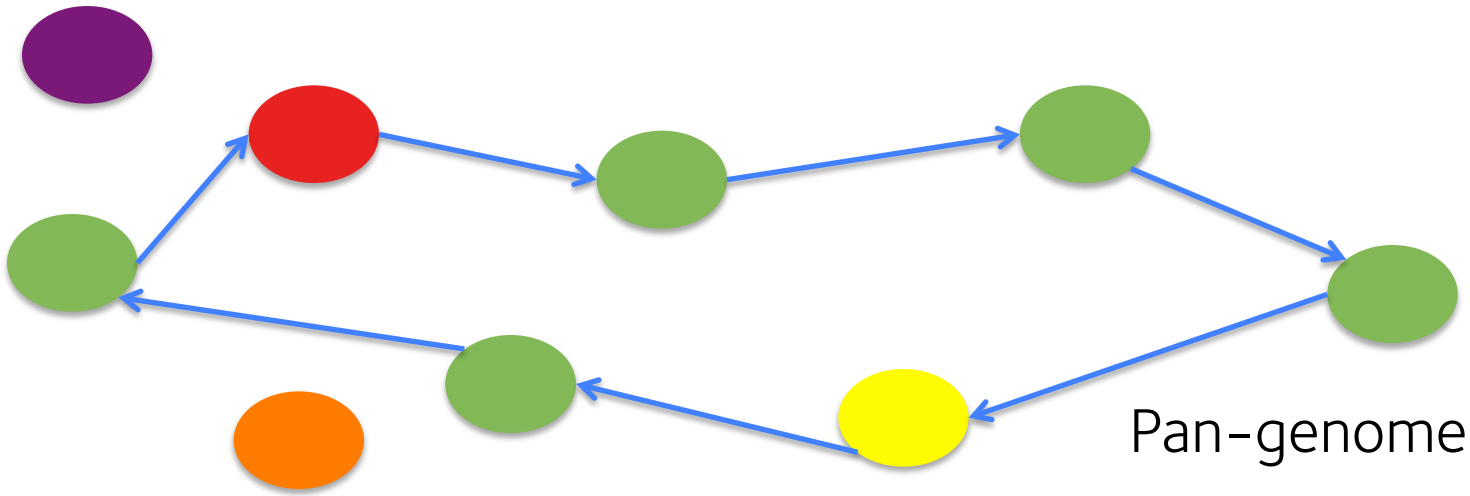
Identify regions of graph with low support. Cut out reads from that region, assemble candidate paths.
Implemented and currently being tested, by **Michael Hall**

Uses GATB: thanks Rayan Chikhi

A substrate for mixtures

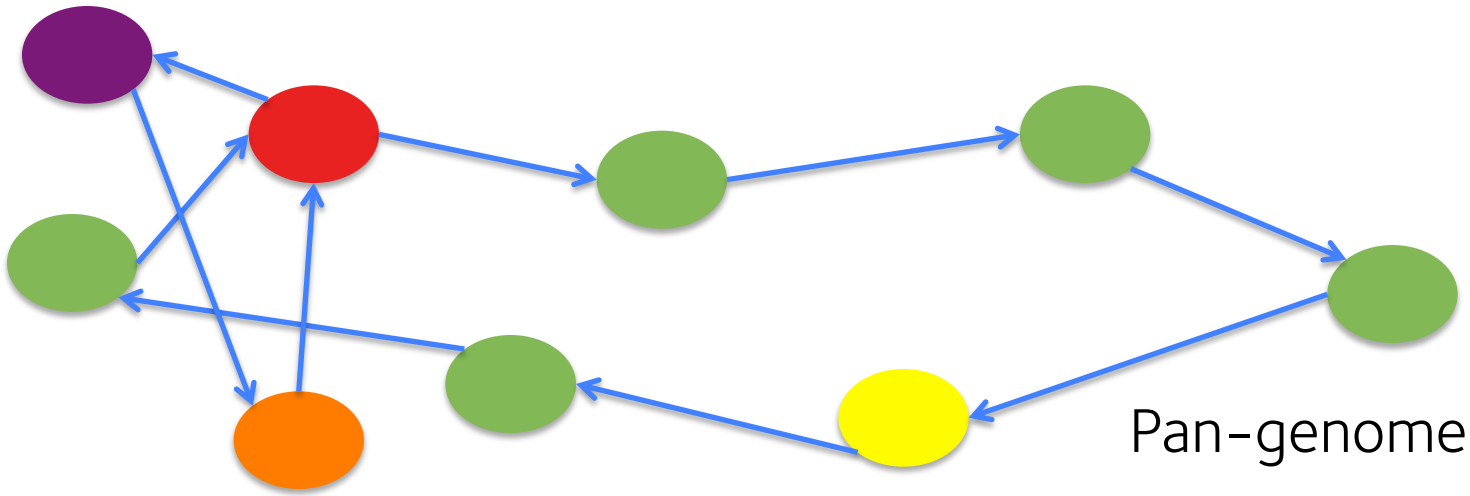


A single genome



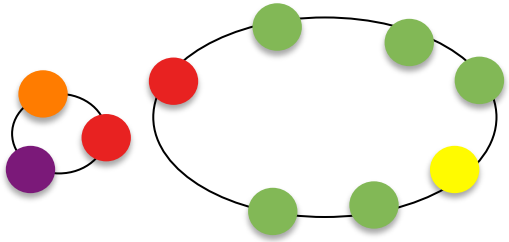
Pan-genome graph

A single genome + plasmids

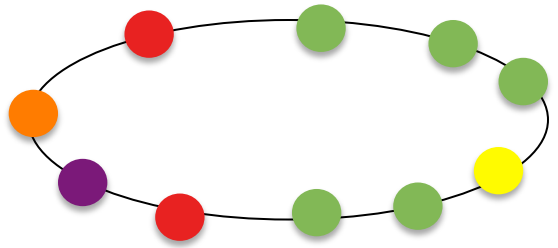


Pan-genome graph

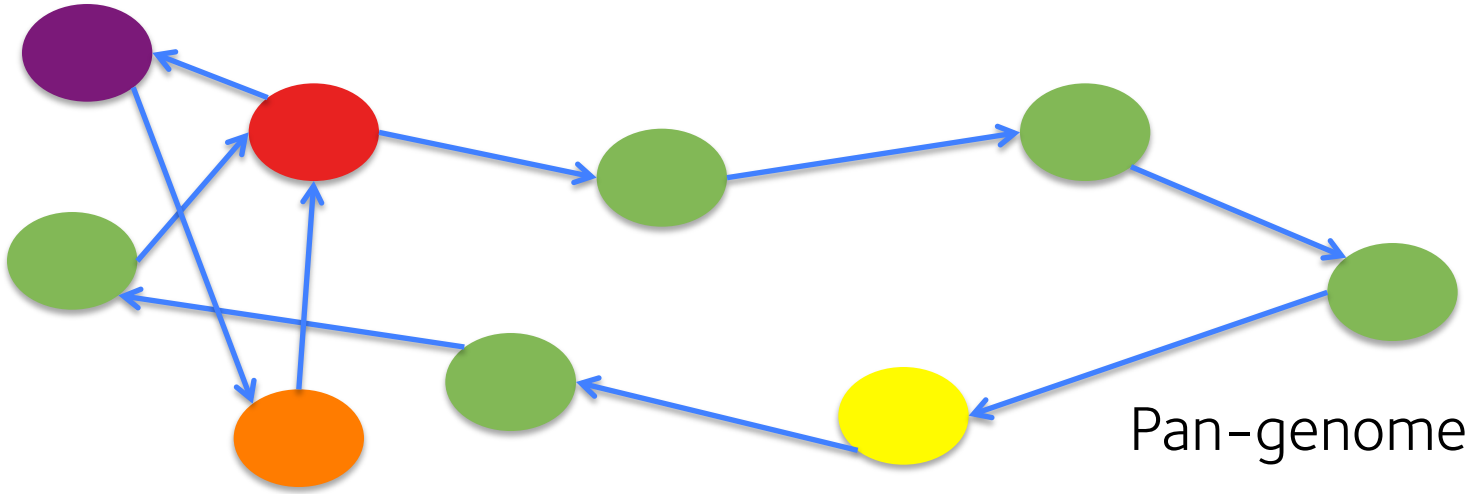
A single genome + plasmids



or

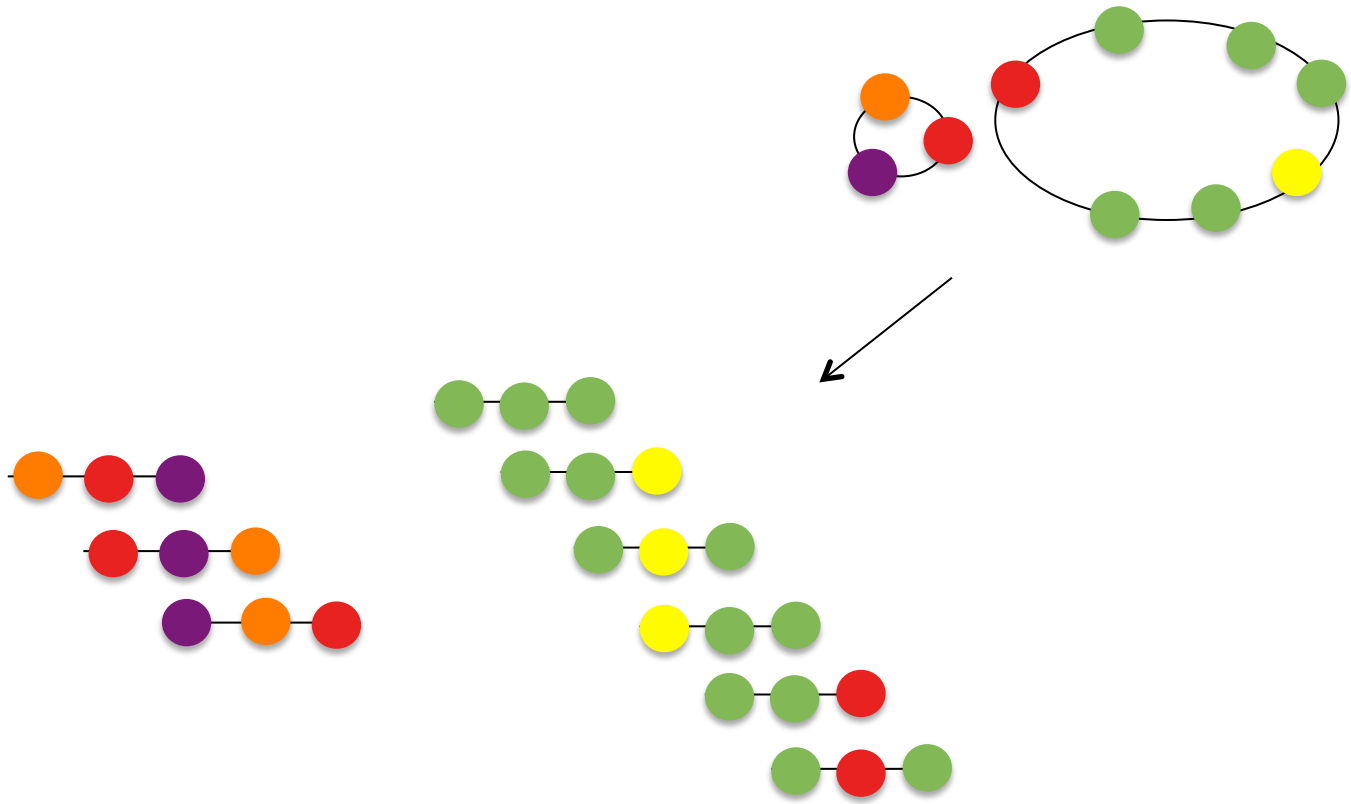


?

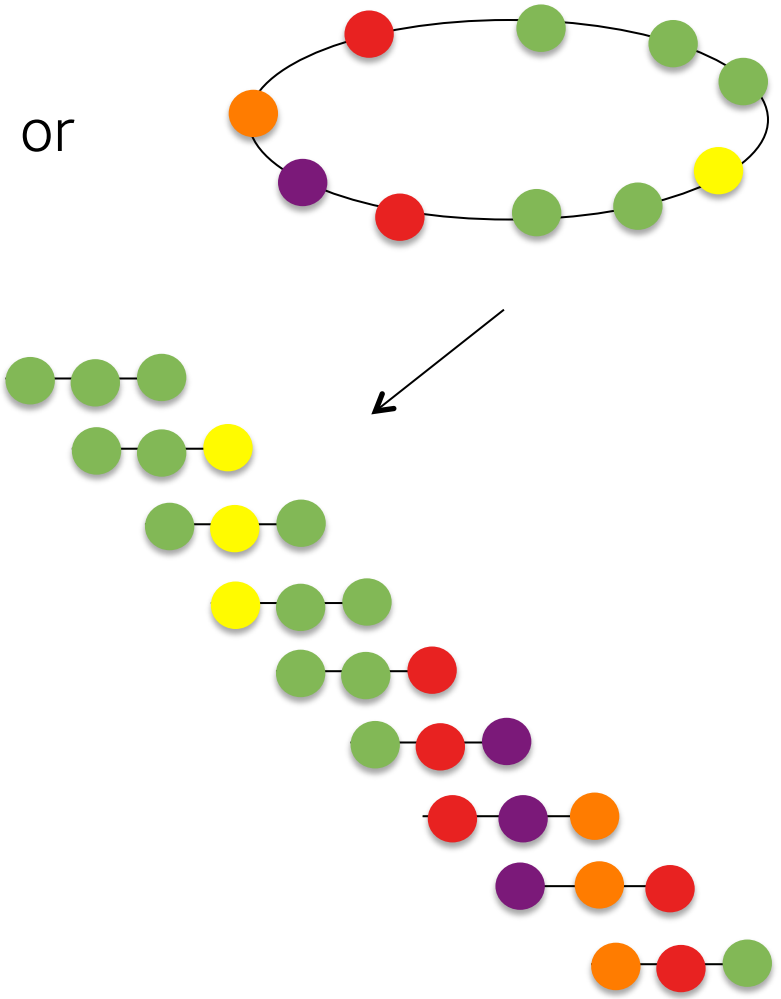


Pan-genome graph

A single genome + plasmids

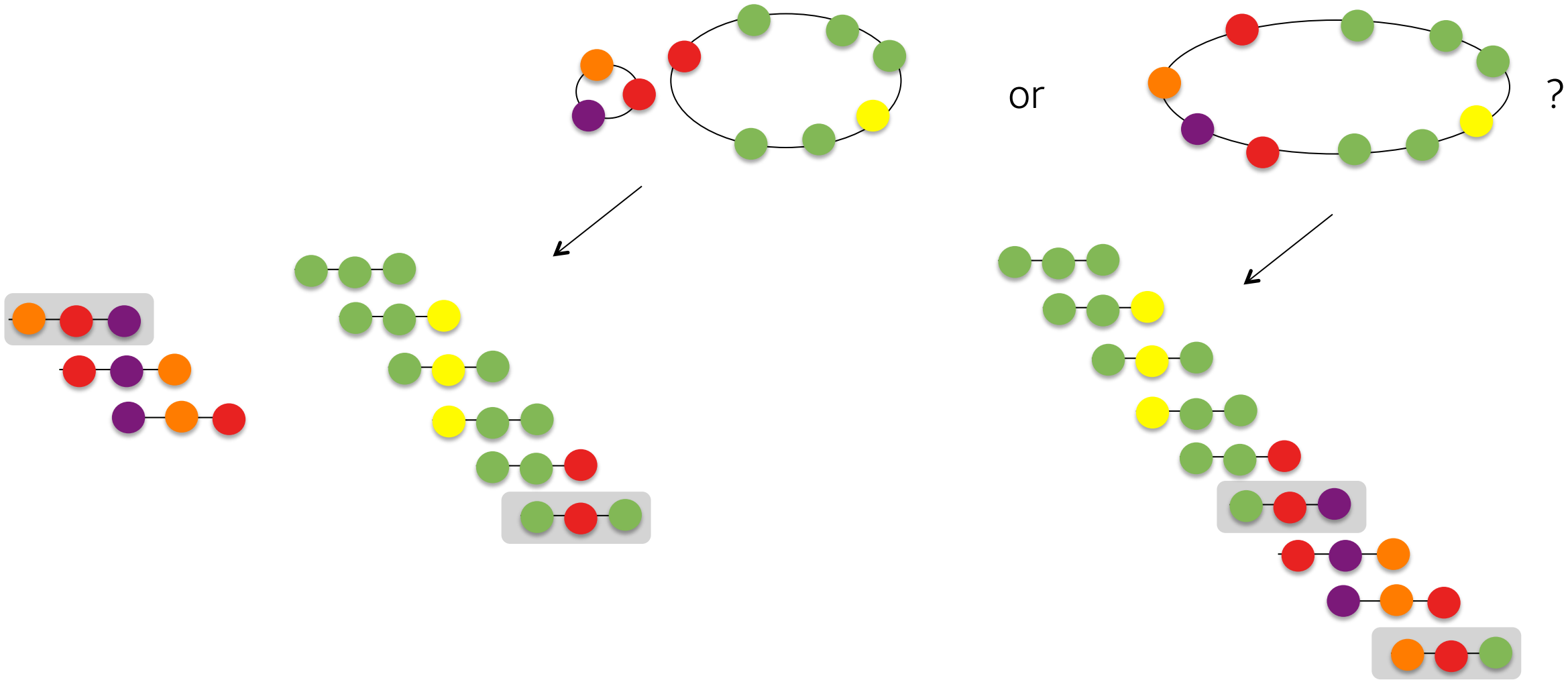


or

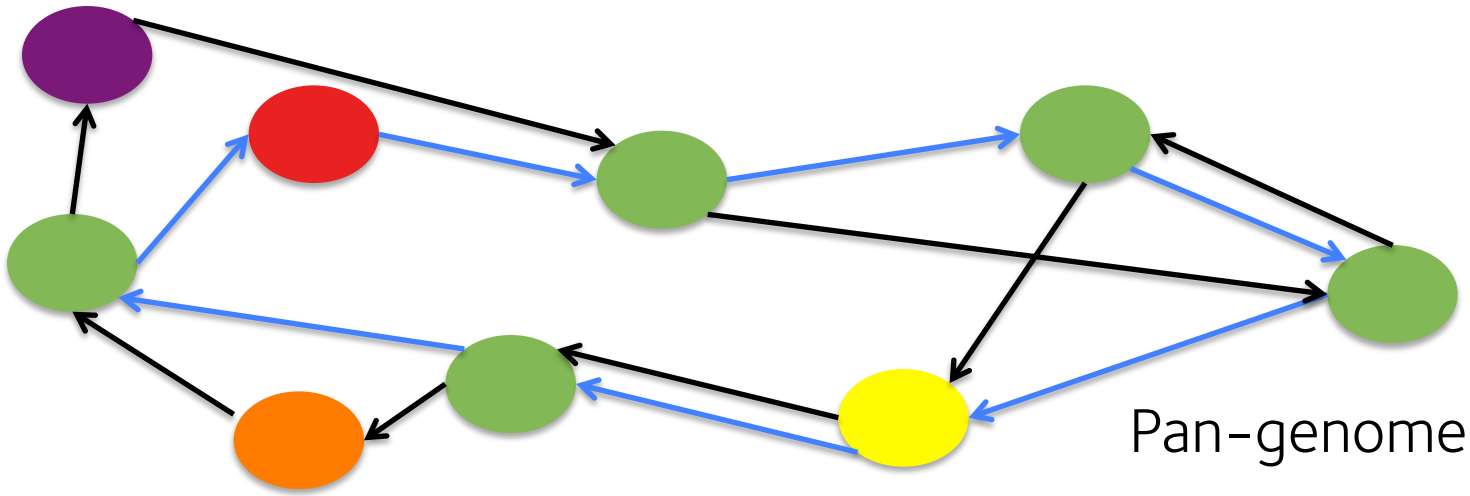


?

A single genome + plasmids

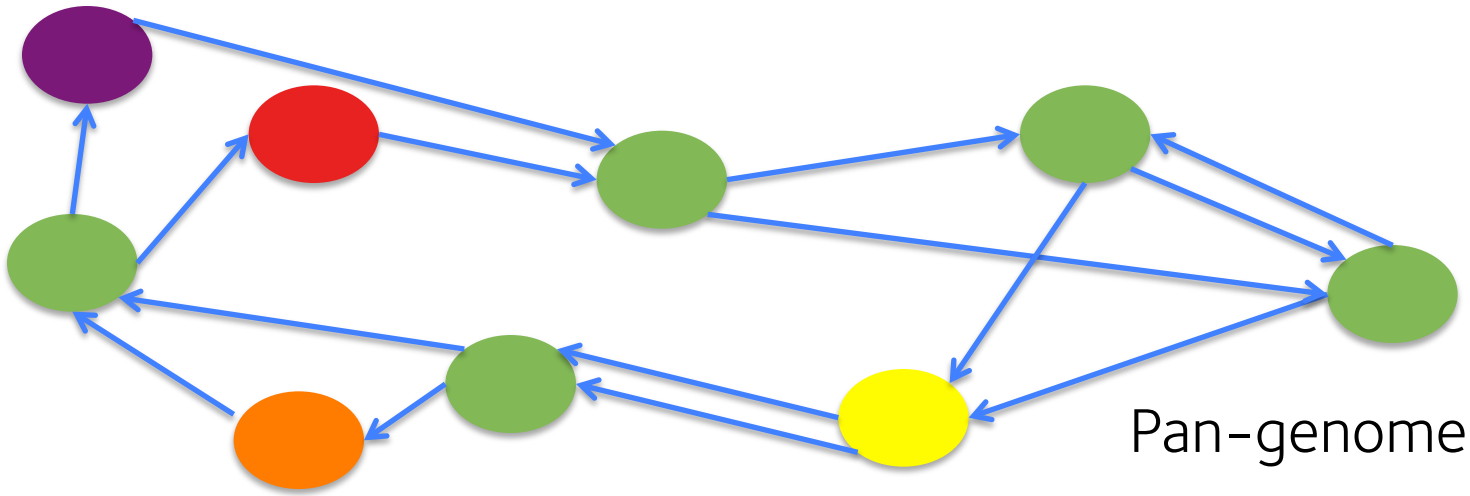


Mixed genomes



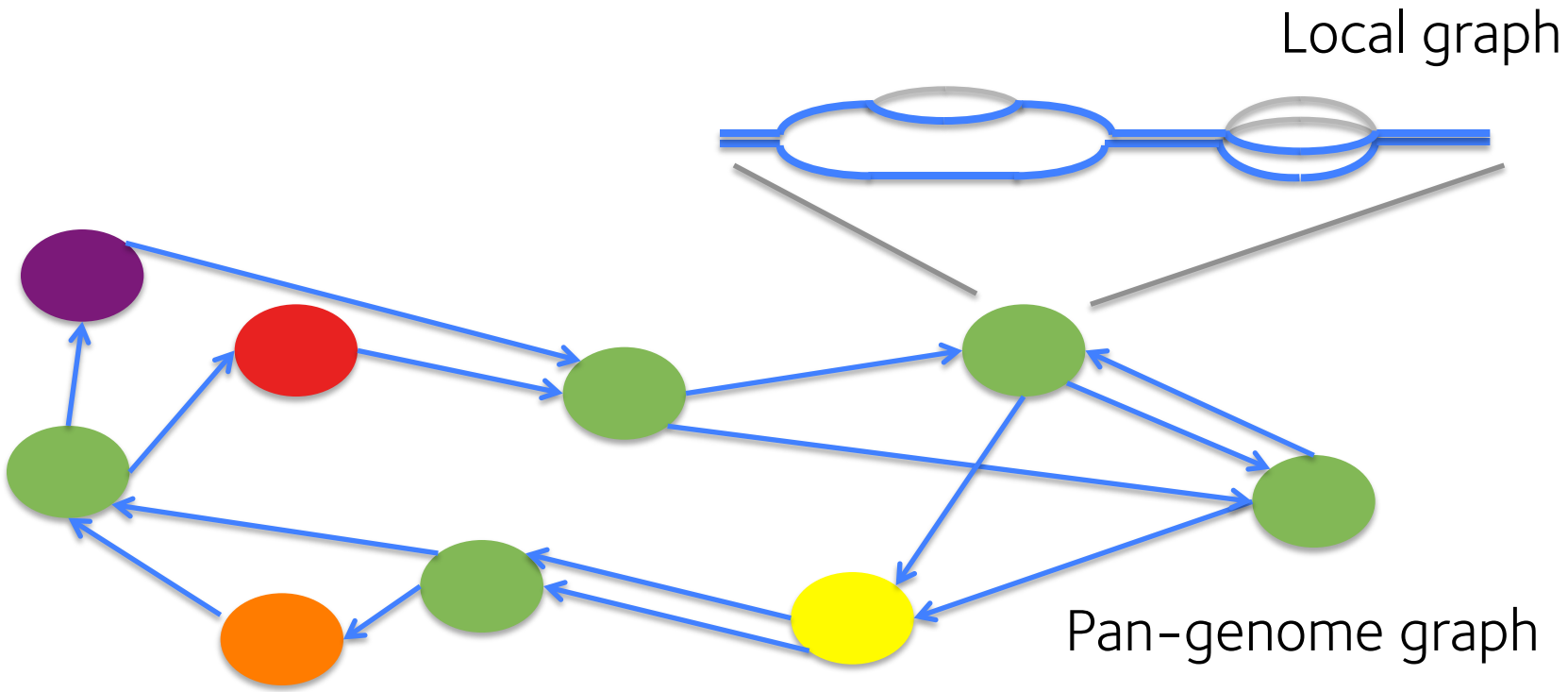
Pan-genome graph

Mixed genomes

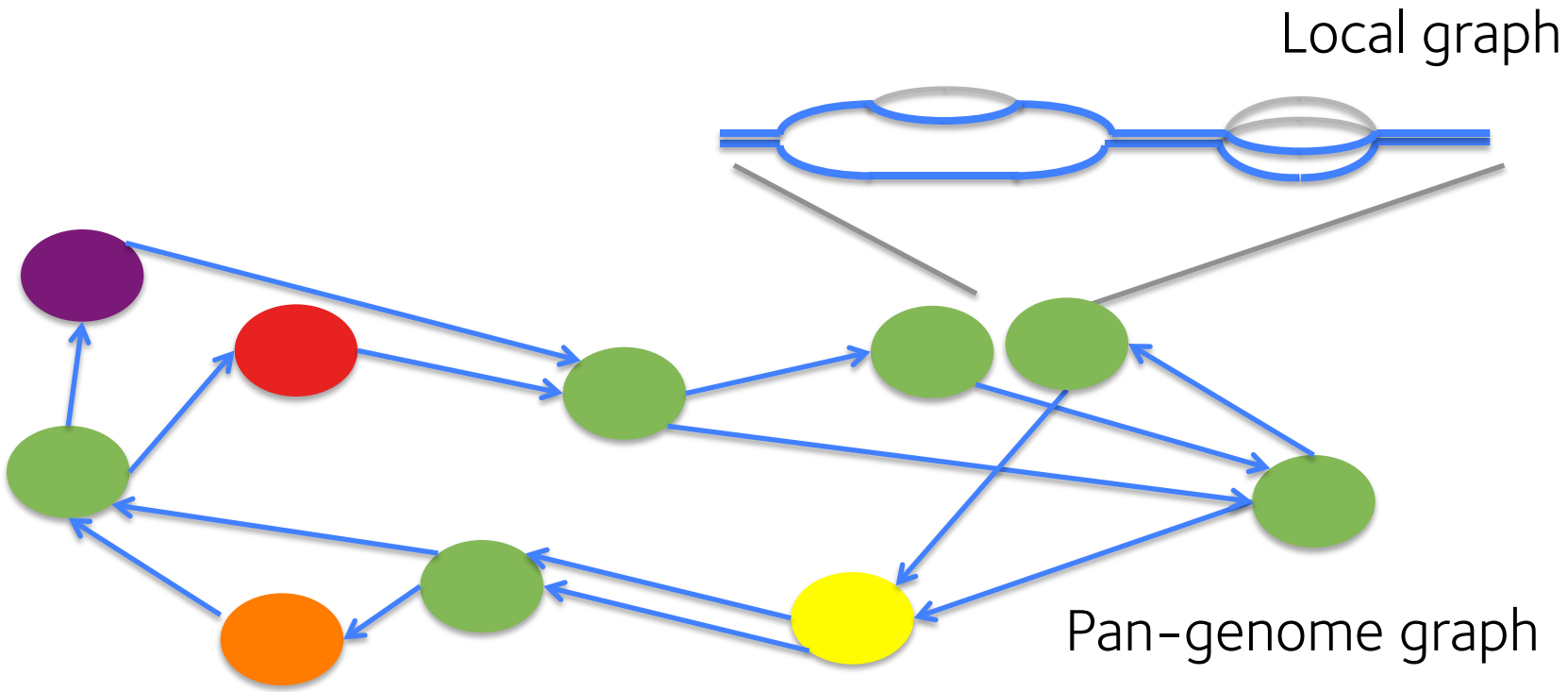


Pan-genome graph

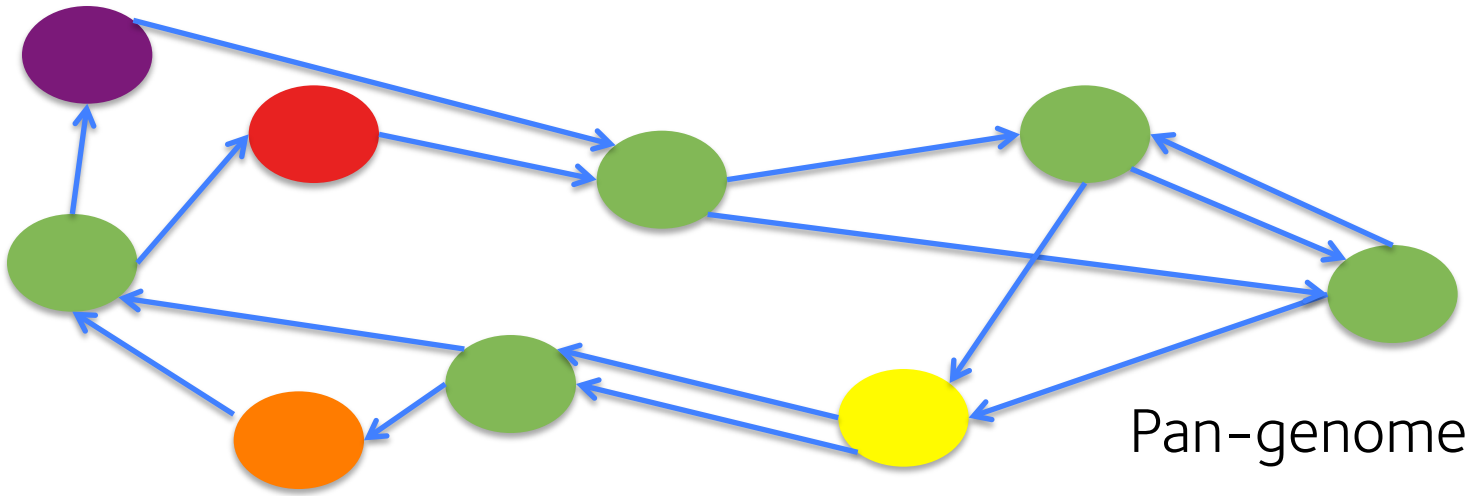
Mixed genomes



Mixed genomes

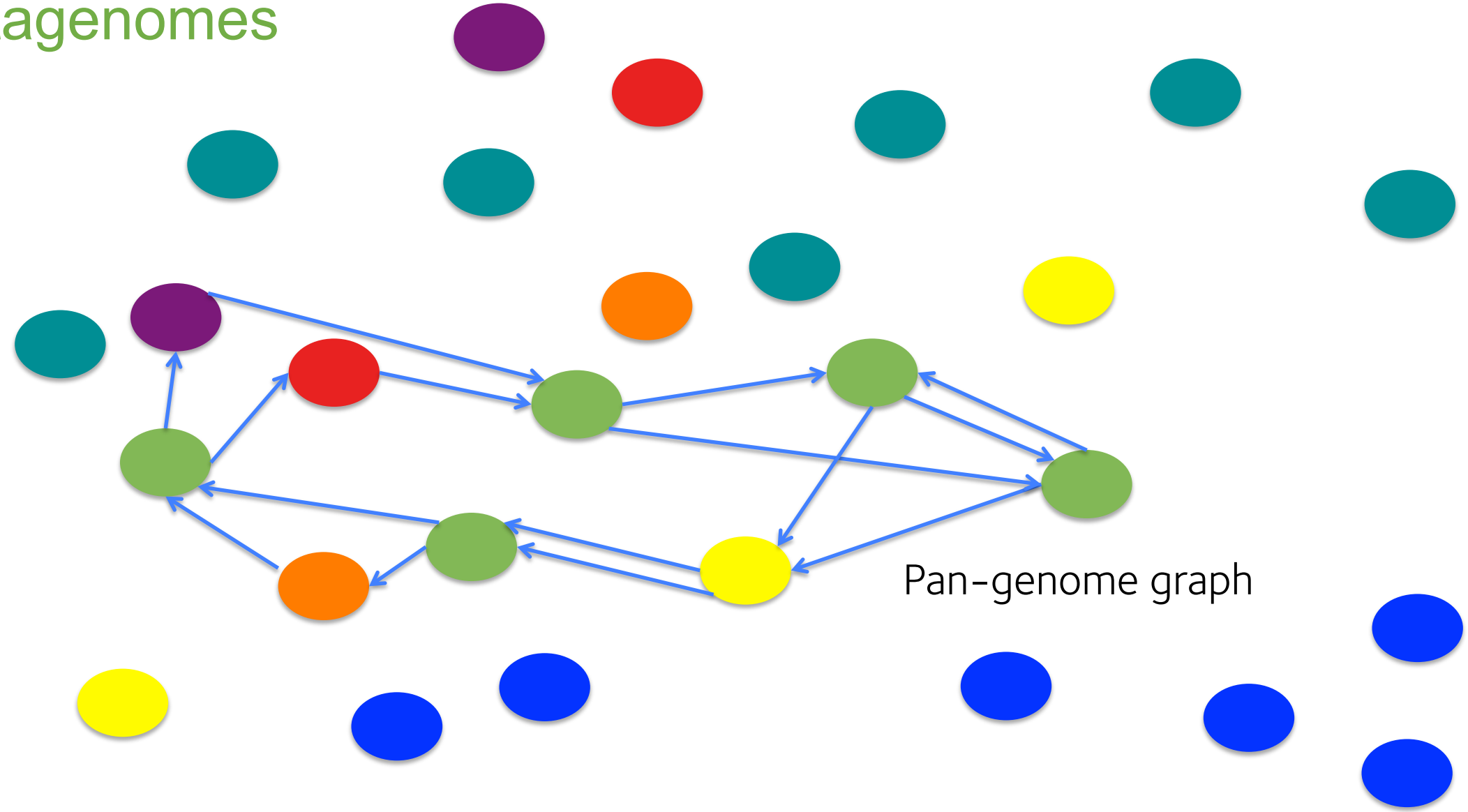


Metagenomes



Pan-genome graph

Metagenomes



Thank you!

Iqbal Lab
Michael Hall
Martin Hunt
Robyn Ffrancon

Manchester
Andrew Dodgson
Ryan George

MMM Group
Nicole Stoesser
Hang Phan
Sophie George
Louise Pankhurst

Biozentrum,
University of Basel
Richard Neher

Max Planck Institute for
Developmental Biology
Wei Ding

University of Bath
Harry Thorpe
Edward Feil

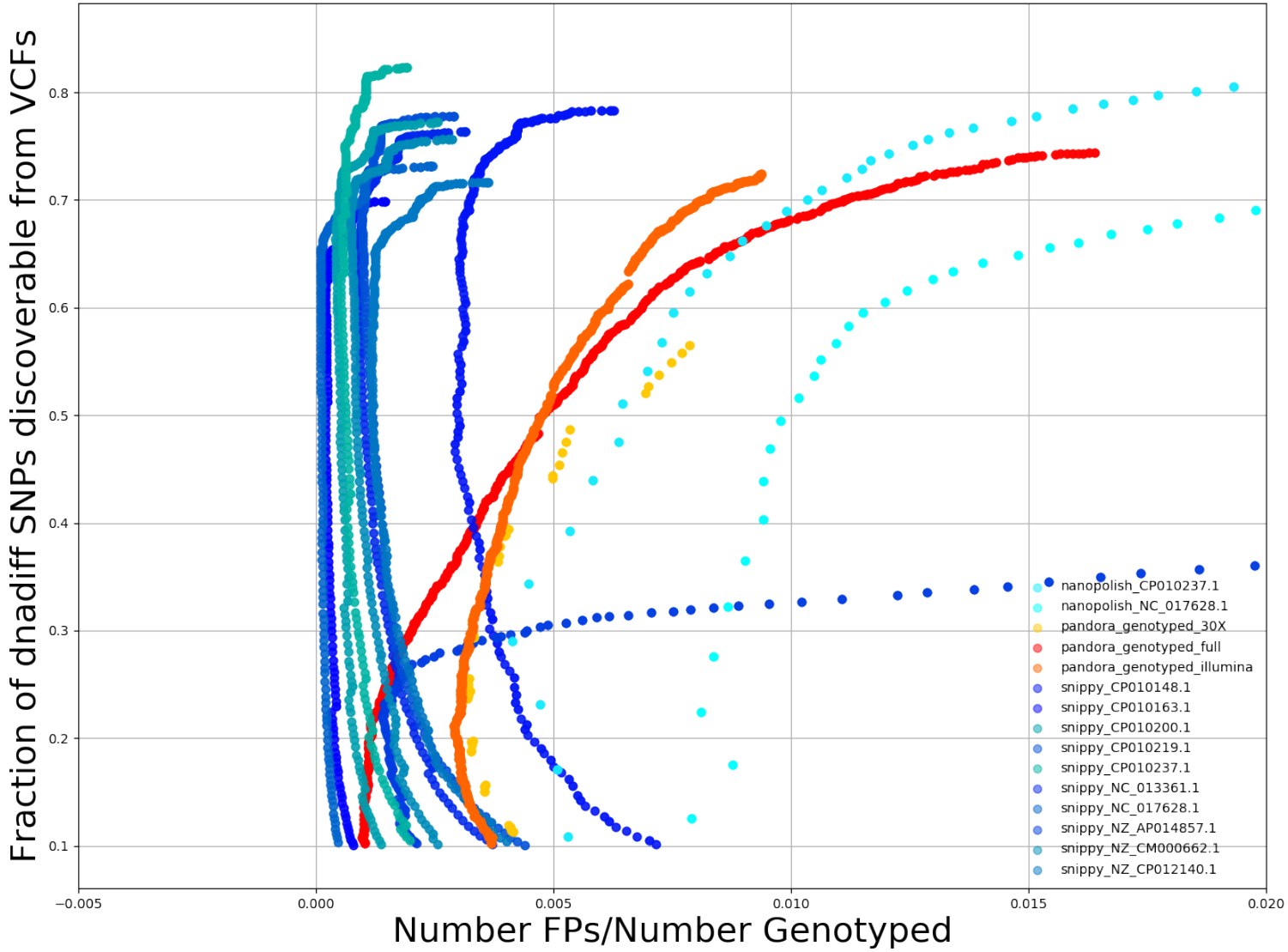
wellcome centre human genetics

@rmcolq

https://github.com/rmcolq/pandora

wellcome

2 samples including nanopolish (2 human)



Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order

AGGTGACACGT

Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order

Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order

AGGTGACACGT
AGGTG
GGTGA
GTGAC

Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order

Indexing in with (w,k)-minimizers

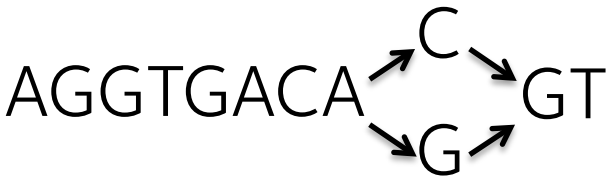
k=5, w=3
dictionary order

AGGTGACACGT
AGGTG
GGTGA
GTGAC
TGACA
GACAC
ACACG
CACGT

AGGTG → GGTGA → GACAC → ACACG

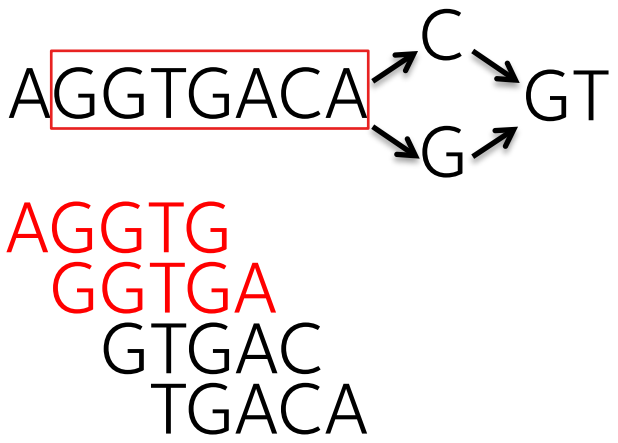
Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order



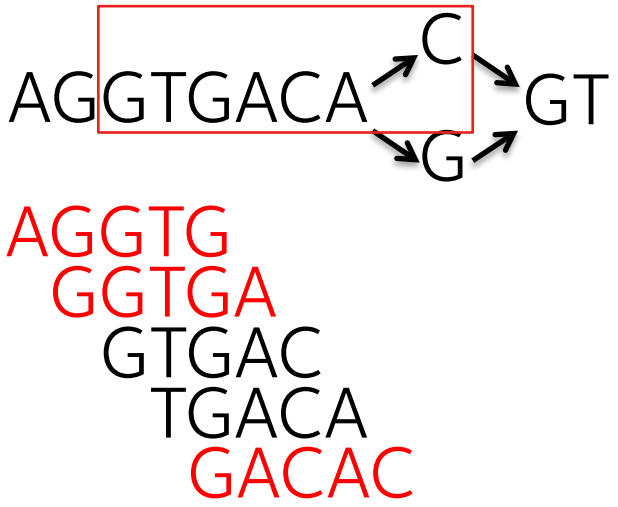
Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order



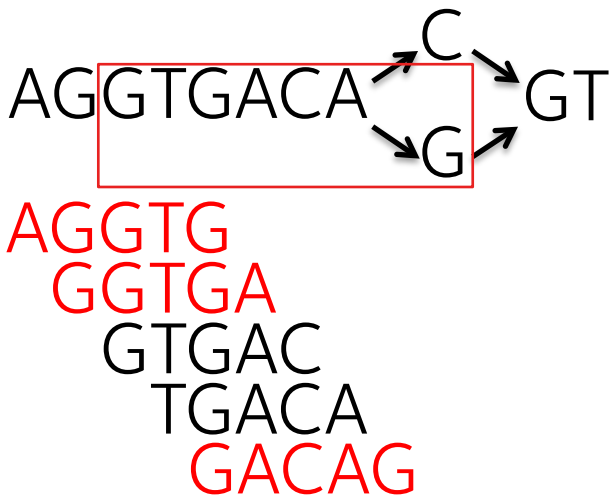
Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order



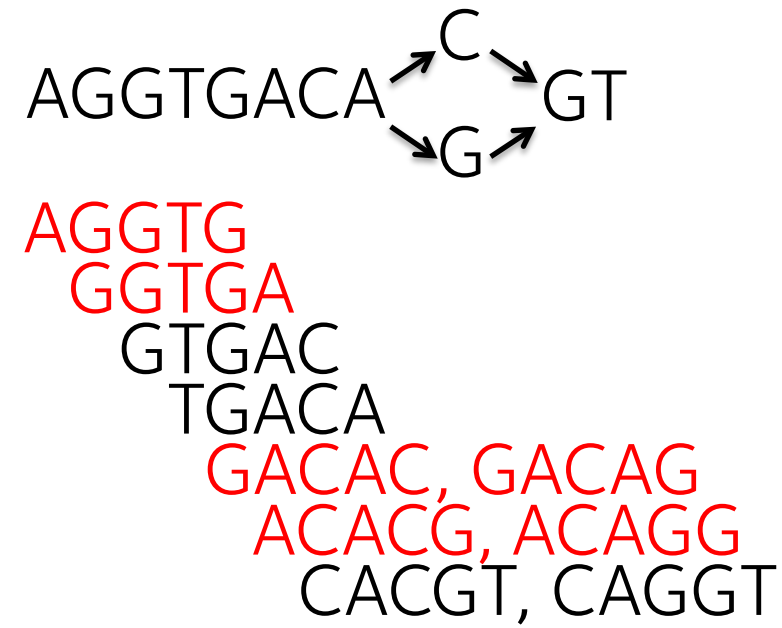
Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order



Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order



Indexing in with (w,k)-minimizers

k=5, w=3
dictionary order