Computational PANGenomics 2022

#CPANG22

Instituto Gulbenkian de Ciência, Portugal
Day 1 - 2022/05/23

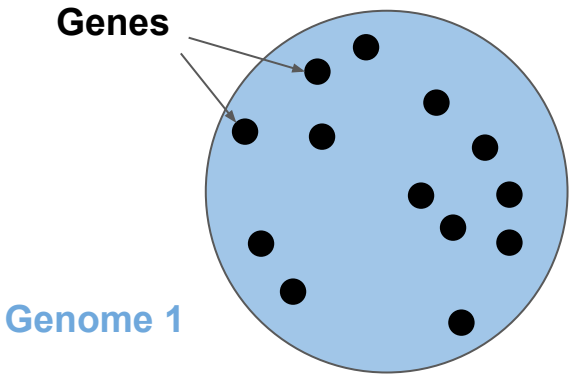
Erik Garrison and Andrea Guarracino

What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).

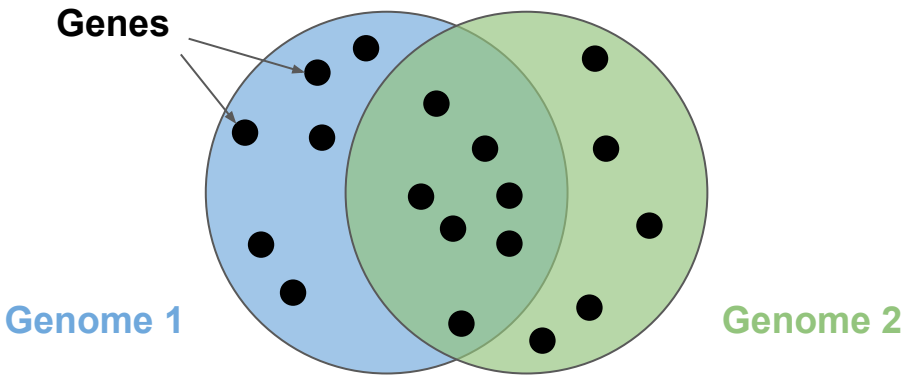
What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).



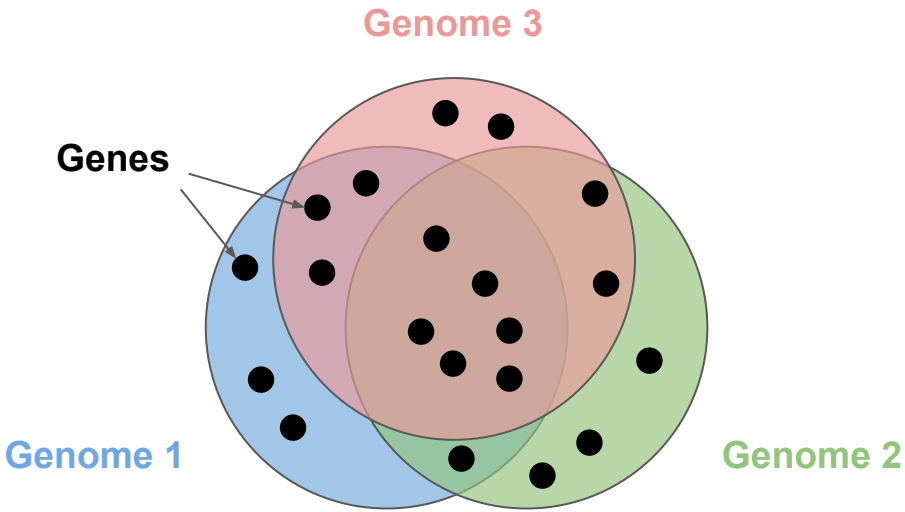
What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).



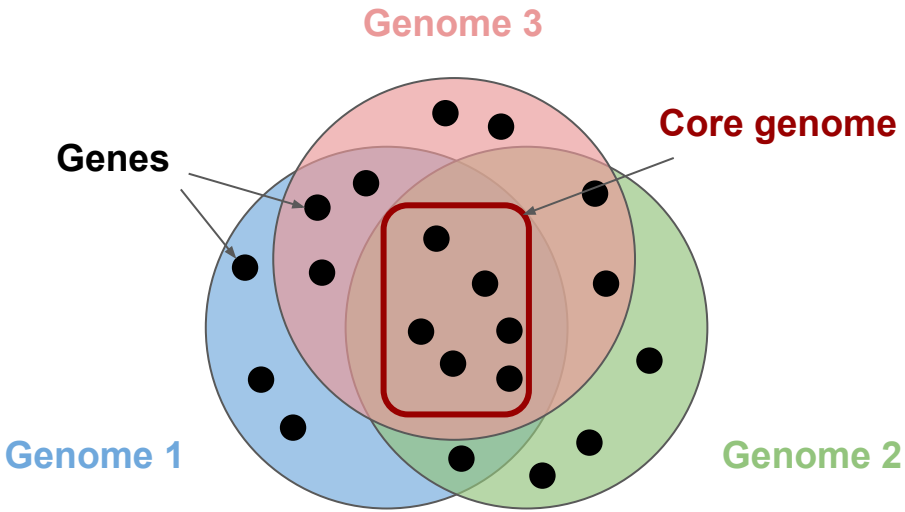
What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).



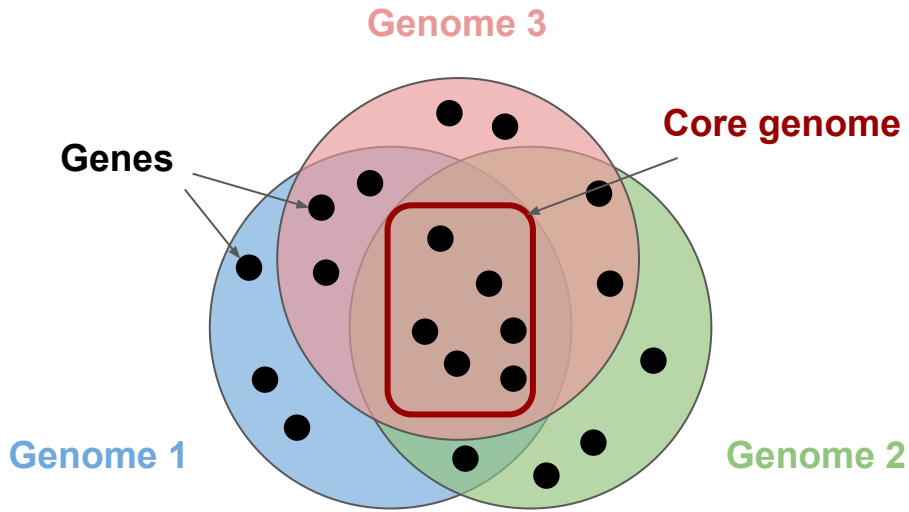
What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).



What is a pangenome?

Term pangenome popularized in microbiology in 2005 (Tettelin et al., 2005).



pangenome =
core genome + disposable genome

What is a pangenome?

We use the term **pangenome** to refer to any collection of genomic sequences to be analyzed **jointly** or to be used as a reference.

What is a pangenome?

We use the term **pangenome** to refer to any collection of genomic sequences to be analyzed **jointly** or to be used as a reference.

- Not restricted to taxonomic units
- Not restricted to full genomes
- Not tied to graphs

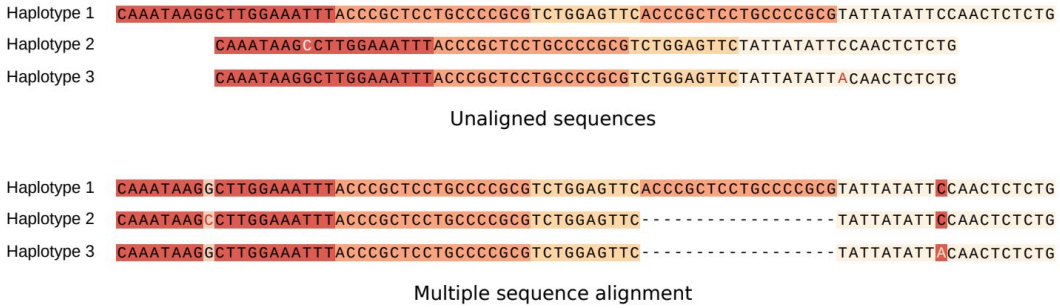


Figure adapted from [The Computational Pan-Genomics Consortium et al., 2018](#).

What is a pangenome?

We use the term **pangenome** to refer to any collection of genomic sequences to be analyzed **jointly** or to be used as a reference.

- Not restricted to taxonomic units
- Not restricted to full genomes
- Not tied to graphs

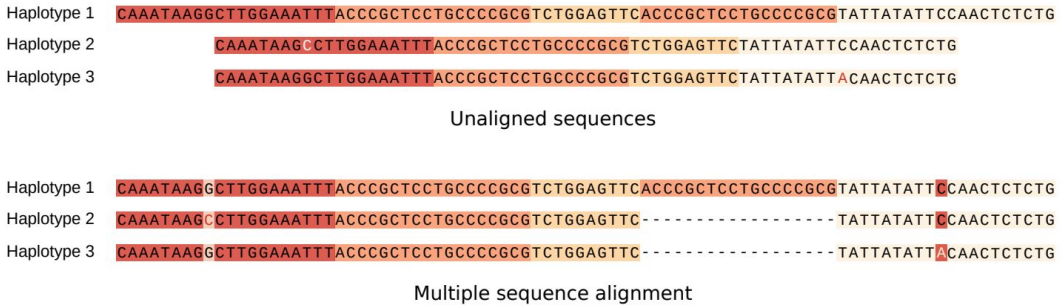


Figure adapted from The Computational Pan-Genomics Consortium et al., 2018.

Data structures | Algorithm | Statistical methods ⟸ **Computational pangenomics**

Why do we need pangenomes?

Thanks to advances in sequencing technology, new **telomere-to-telomere** genome assemblies are produced at a high rate.

Different data types can be combined to generate *de novo* assemblies that approach the high quality of reference genomes, but at a fraction of the cost.

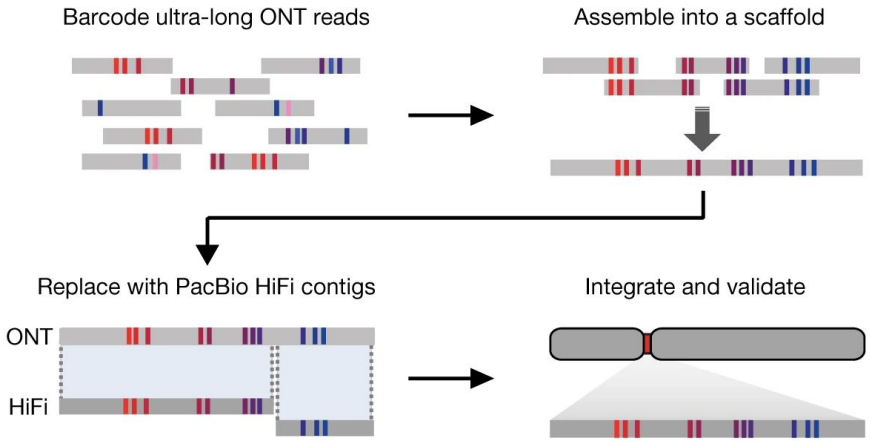
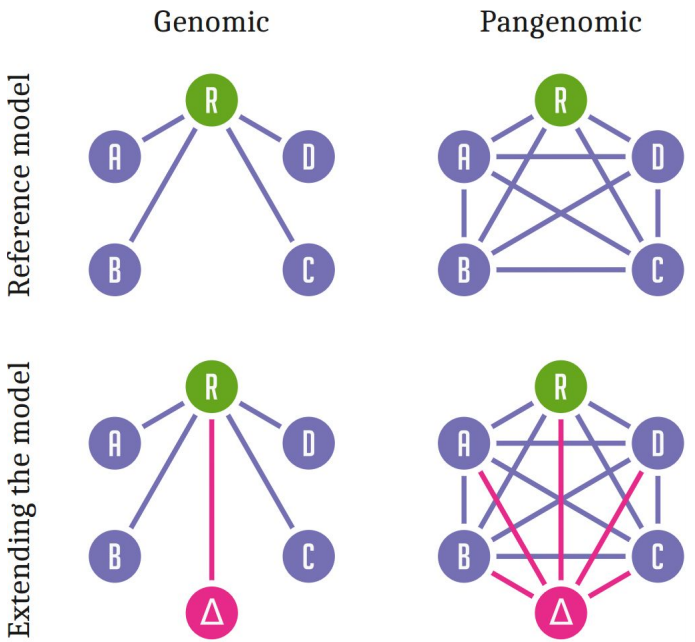


Figure adapted from Logsdon et al., 2021.

Why do we need pangenomes?

Alignments summarize the relationship between sequences, exposing putative evolutionary and functional information.

Pangenomes can **model** the full set of genomic elements in a given species or clade, reducing the **reference-bias**.



Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

Pangenome graph

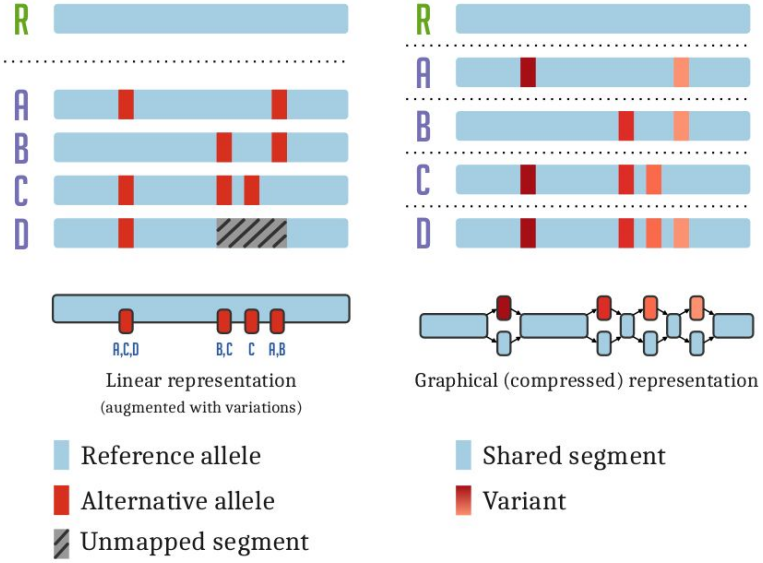


Figure from Eizenga et al., 2020.

Pangenomes can take many forms, including **graph-based** data structures.

Pangenome graphs compress redundant sequences into a smaller data structure that is still representative of the full set.

Variation graph

- Genome 1: **ACTACAGTACTGGCAGT**
- Genome 2: **ACTACAGTAAAGTACAGT**

Variation graph

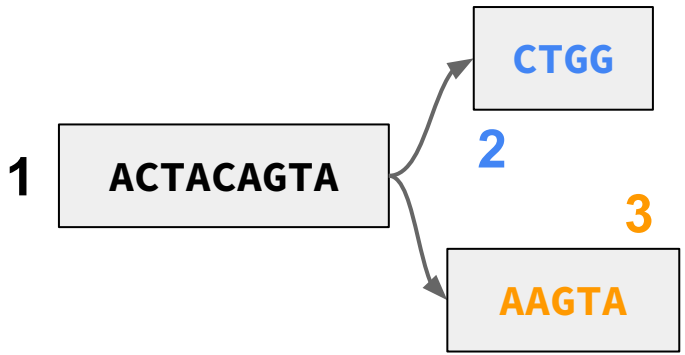
Genome 1: **ACTACAGTACTGGCAGT**

Genome 2: **ACTACAGTAAAGTACAGT**

Variation graph

Genome 1: **ACTACAGTACTGGCAGT**

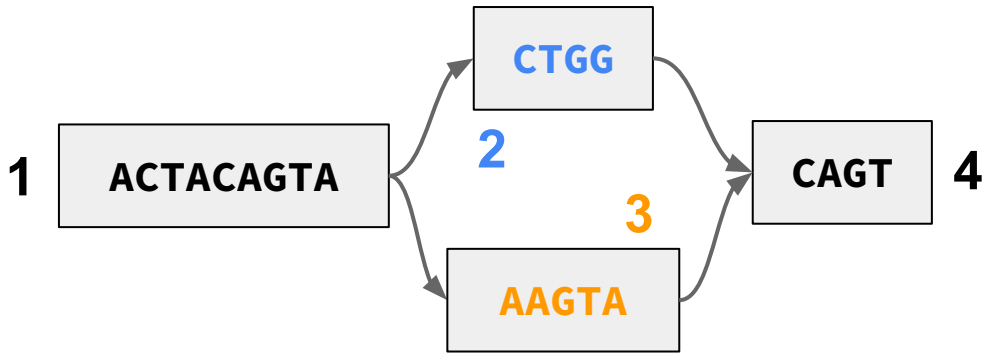
Genome 2: **ACTACAGTAAAGTACAGT**



Variation graph

Genome 1: **ACTACAGTACTGGCAGT**

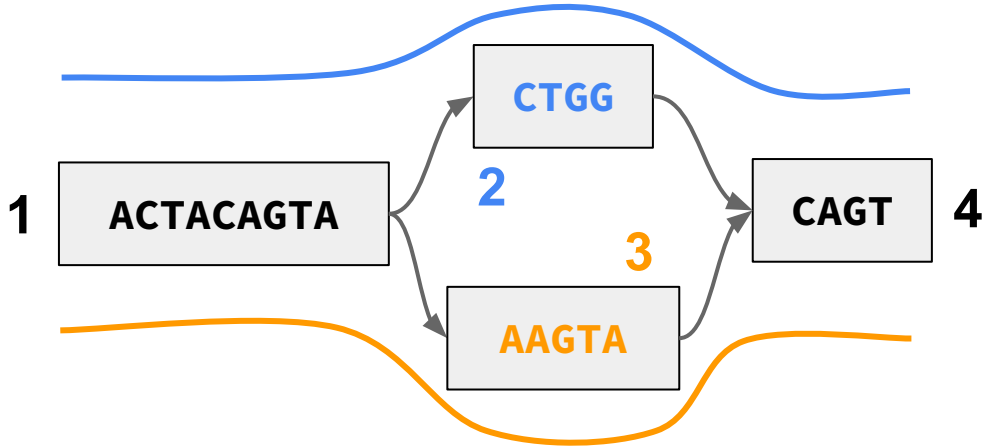
Genome 2: **ACTACAGTAAAGTACAGT**



Variation graph

Genome 1: **ACTACAGTACTGGCAGT**

Genome 2: **ACTACAGTAAAGTACAGT**



Linear sequences are **paths** through nodes.

Genome 1: 1 2 4

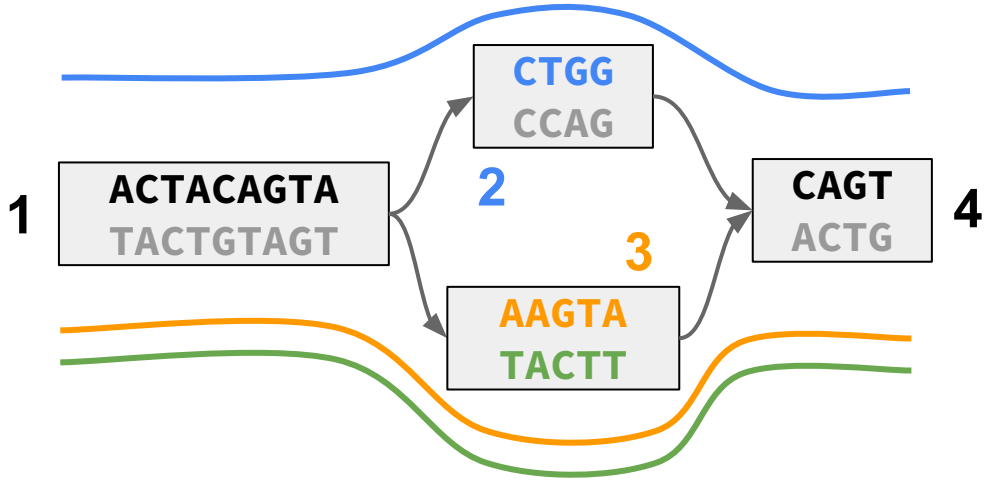
Genome 2: 1 3 4

Variation graph

Genome 1: **ACTACAGTACTGGCAGT**

Genome 2: **ACTACAGTAAAGTACAGT**

Genome 3: **ACTACAGTATACTTCAGT**



Linear sequences are **paths** through nodes.

Genome 1: **1+ 2+ 4+**

Genome 2: **1+ 3+ 4+**

Genome 3: **1+ 3− 4+**

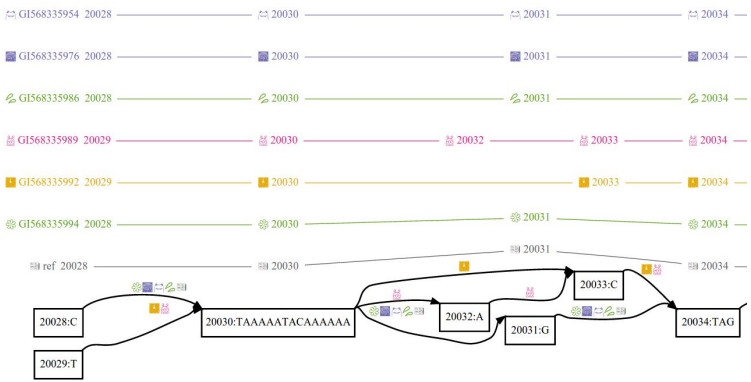
Node traversed in reverse

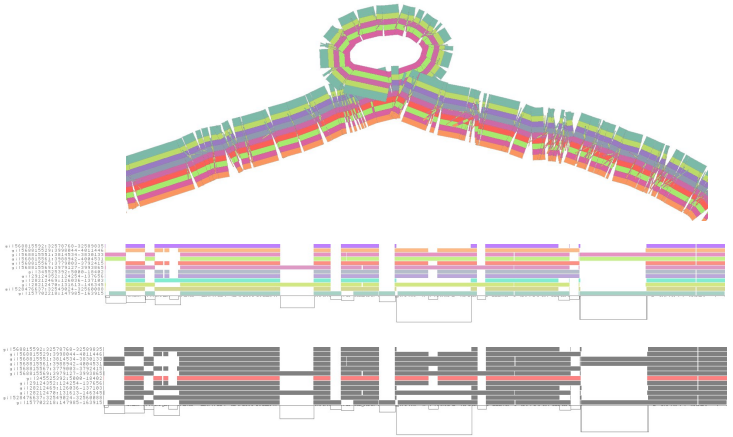
VG and ODGI toolkits

https://github.com/vgteam/vg

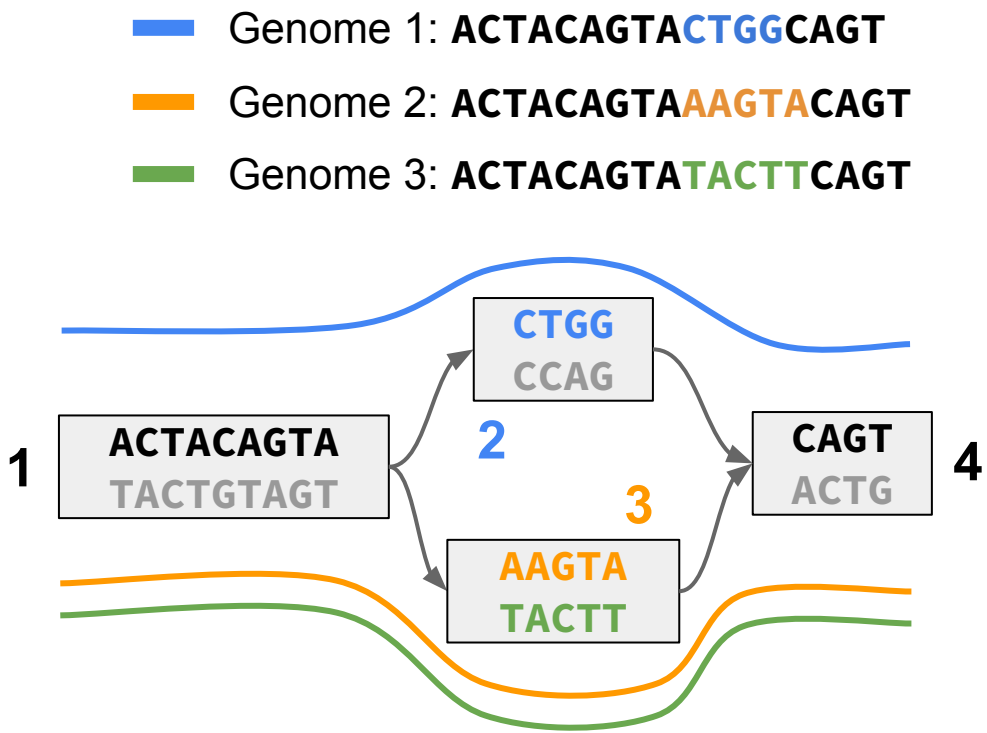
odgi

https://github.com/pangenome/odgi





Graphical Fragment Assembly version 1 (GFAv1)




```
H    VN:Z:1.0
S    1    ACTACAGTA
S    2    CTGG
S    3    AAGTA
S    4    CAGT
L    1    +    2    +    *
L    2    +    4    +    *
L    1    +    3    +    *
L    3    +    4    +    *
L    1    +    3    -    *
P    Genome 1    1+,2+,4+    *
P    Genome 2    1+,3+,4+    *
P    Genome 3    1+,3-,4+    *
```

http://gfa-spec.github.io/GFA-spec/GFA1.html

Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF    ALT        QUAL  FILTER  INFO       FORMAT  sample
ref     10   .   A      T          99    .       .          GT      1
ref     21   .   A      ATTAAGA    99    .       .          GT      1
ref     34   .   TCTTT  T          99    .       .          GT      1
```

```
TGGGAGAGAACTGGAACAAGAACCCAGTGCTCTTTCTGCTCTA
```

ref.fa

Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF    ALT        QUAL  FILTER  INFO       FORMAT  sample
ref     10   .   A      T          99    .       .          GT      1
ref     21   .   A      ATTAAGA    99    .       .          GT      1
ref     34   .   TCTTT  T          99    .       .          GT      1
```

TGGGAGAGAACTGGAACAAGAACCCAGTGCTCTTTCTGCTCTA

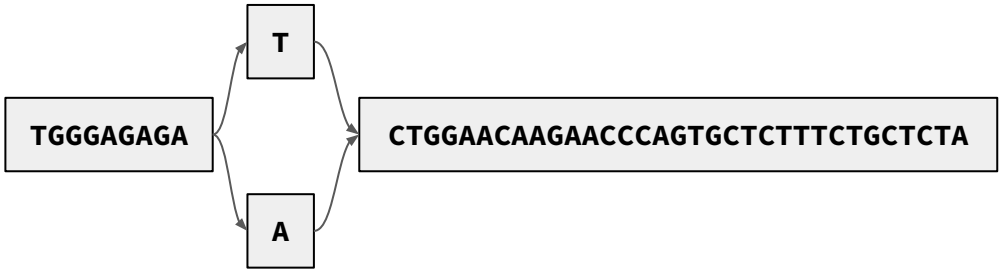
ref.fa

For each variant

1) cut the reference path around the variant
2) add the novel (ALT) sequence to the graph

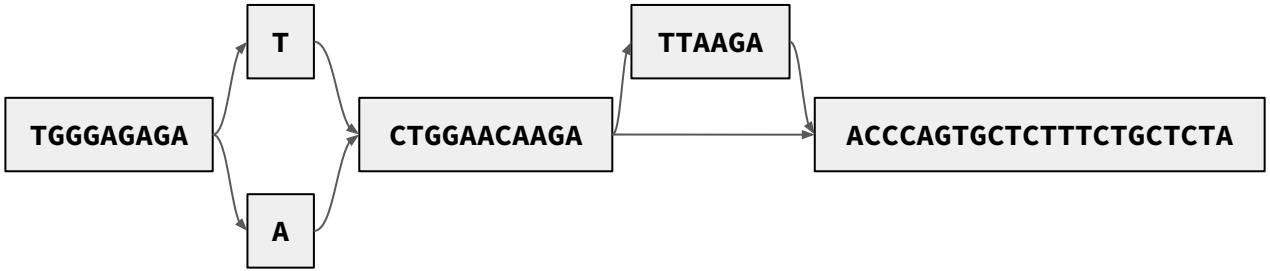
Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF    ALT      QUAL  FILTER  INFO  FORMAT  sample
ref     10   .   A      T        99    .       .     GT      1
ref     21   .   A      ATTAAGA  99    .       .     GT      1
ref     34   .   TCTTT  T        99    .       .     GT      1
```



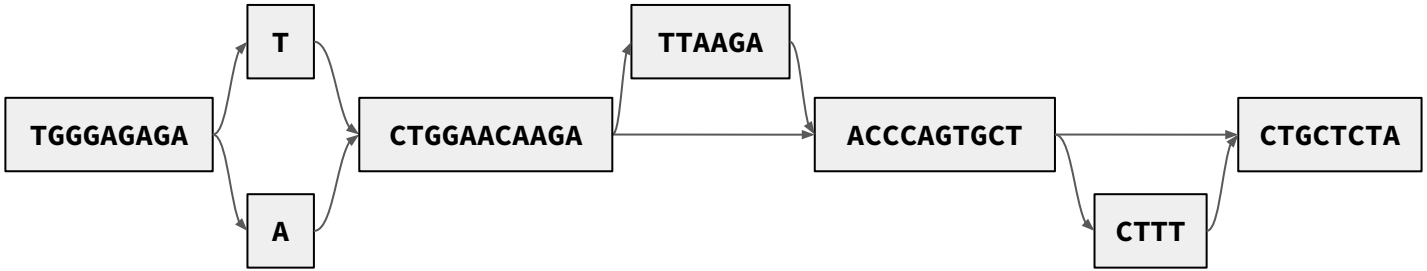
Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF   ALT      QUAL  FILTER  INFO  FORMAT  sample
ref     10   .   A     T        99    .       .     GT      1
ref     21   .   A     ATTAAGA  99    .       .     GT      1
ref     34   .   TCTTT T        99    .       .     GT      1
```



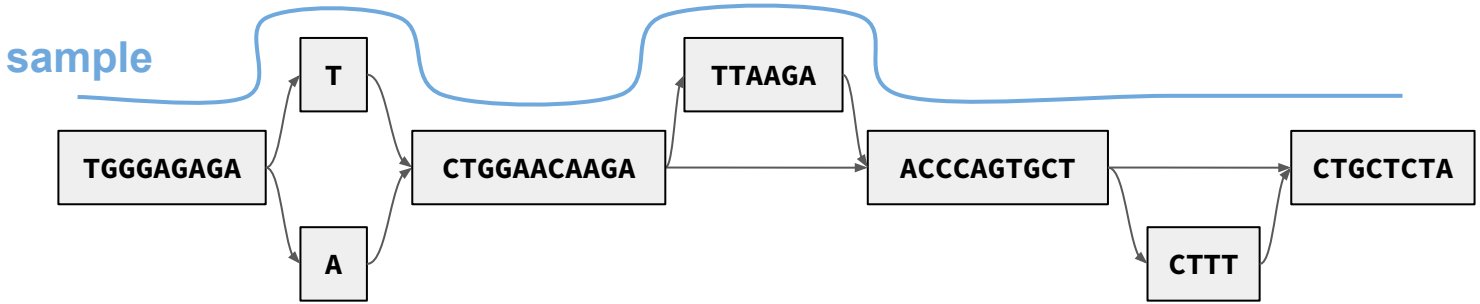
Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF    ALT        QUAL  FILTER  INFO        FORMAT  sample
ref     10   .   A      T          99    .       .           GT      1
ref     21   .   A      ATTAAGA    99    .       .           GT      1
ref     34   .   TCTTT  T          99    .       .           GT      1
```



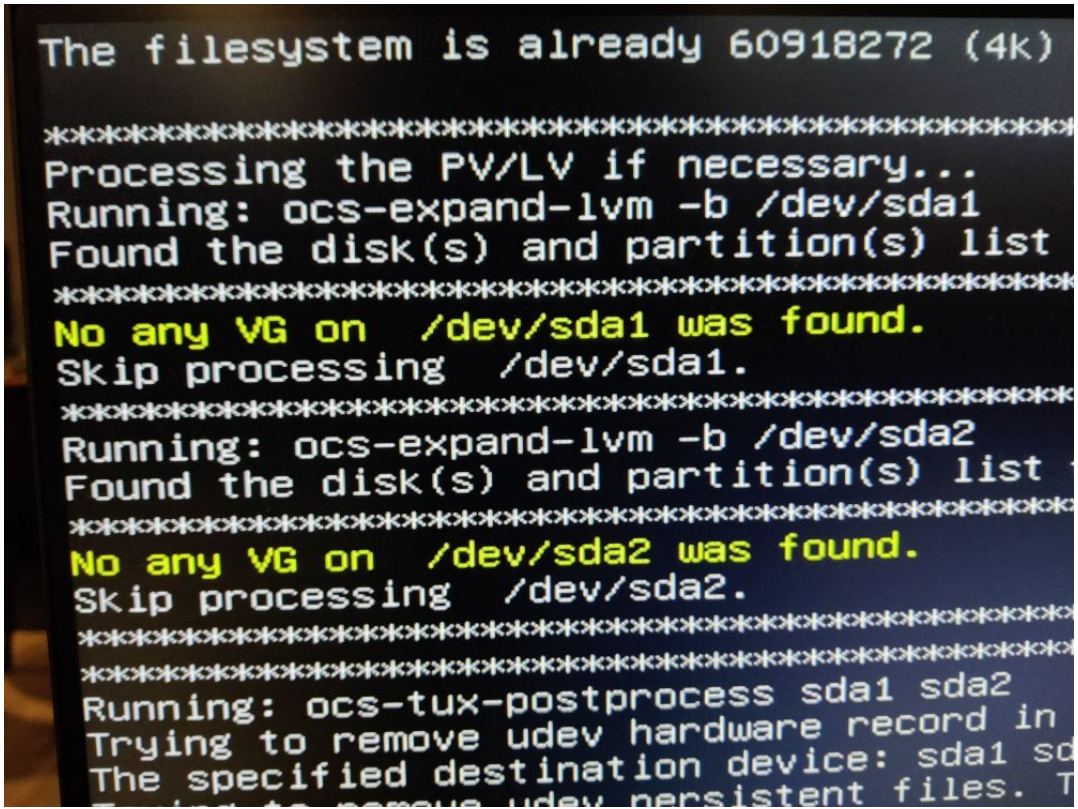
Pangenome building (from a VCF file)

```
##fileformat=VCFv4.3
##reference=ref.fa
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF    ALT        QUAL  FILTER  INFO        FORMAT  sample
ref     10   .   A      T          99    .       .           GT      1
ref     21   .   A      ATTAAGA    99    .       .           GT      1
ref     34   .   TCTTT  T          99    .       .           GT      1
```



Activities

https://hackmd.io/@AndreaGuarracino/SJQ1XPGD9


The filesystem is already 60918272 (4k)
Processing the PV/LV if necessary...
Running: ocs-expand-lvm -b /dev/sda1
Found the disk(s) and partition(s) list
No any VG on /dev/sda1 was found.
Skip processing /dev/sda1.
Running: ocs-expand-lvm -b /dev/sda2
Found the disk(s) and partition(s) list
No any VG on /dev/sda2 was found.
Skip processing /dev/sda2.
Running: ocs-tux-postprocess sda1 sda2
Trying to remove udev hardware record in
The specified destination device: sda1 sd