Computational Pangenomics

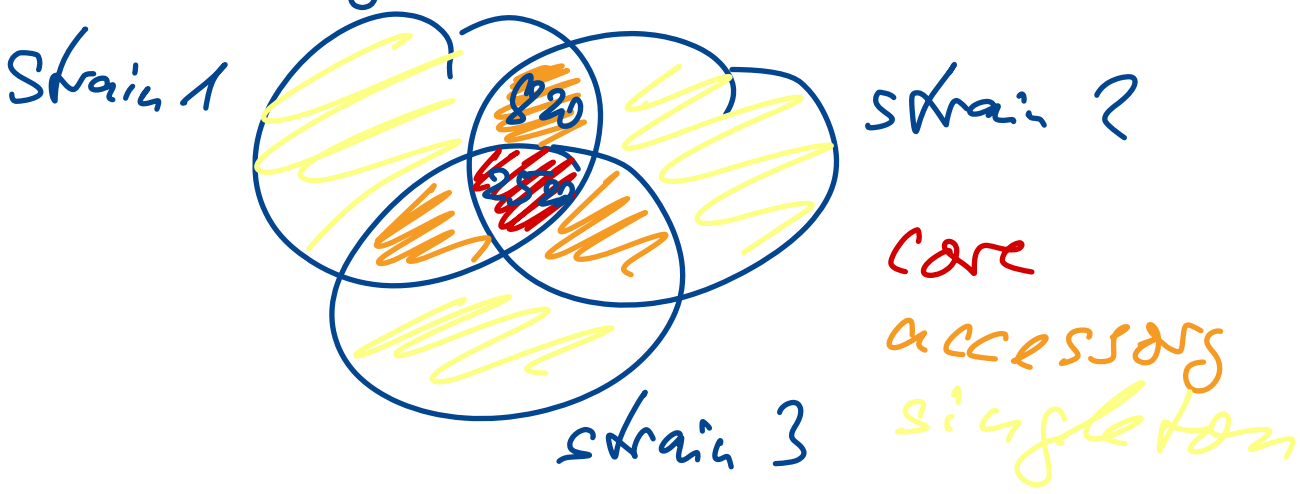
Jan. 15, 2021

goal: understand the genomic variation inside a species.

observation:
- core genome ← every genome has it
- accessory genome ← some
- singleton genome ← only few

I. Gene-based approach

- Analysis based on (protein coding) genes.
- visualized e.g. as a Venn diagram:

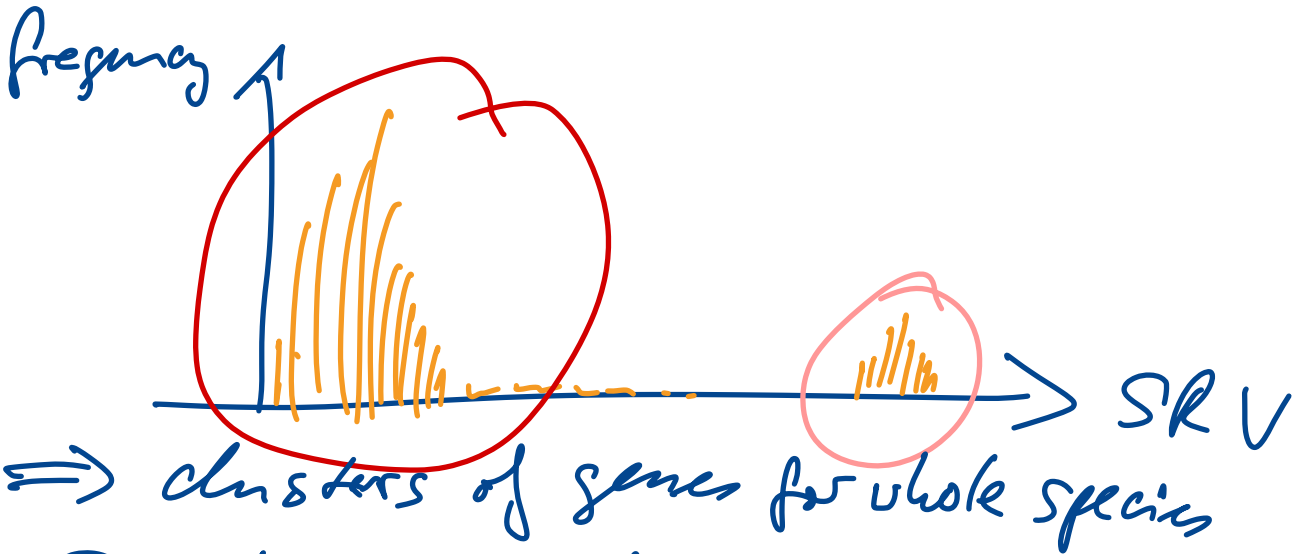


- typical protocol (EDGAR, 2009ff):
  1. select a set of genomes (strains) to be analyzed
  2. annotate genomes (gene finding)

3. gene clustering (many methods exist)

for example (EDGAR):

- all-against-all BLAST comparisons
- calculate score-ratio values (SRVs)
- histogram with bimodal distribution



⇒ clusters of genes for whole species

4. Production of Venn diagram.

Easy to see:

"open" or "closed" pangenome

large accessory + singleton genomes

large core fraction

II. Genome-based approach

disadvantage of approach I: Not all of the genome is studied

prokaryotes: ≈90% coding

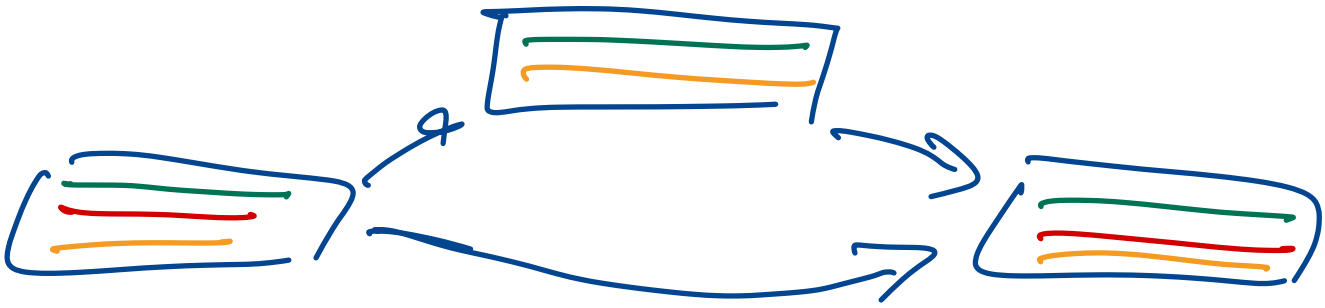
eukaryotes: 1-2% coding

for the complete picture: look at DNA!

Data structures:

(1) based on alignments
(2) variation graph
(3) colored de Bruijn graph

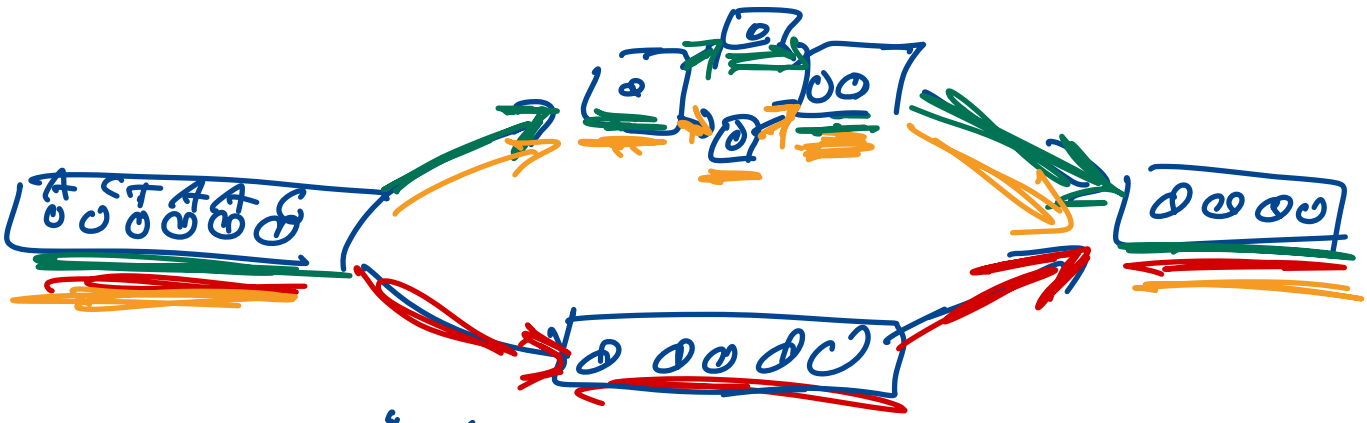
ad (1): Pan genome alignment graph (PanCake)



each vertex is a multiple alignment of "shared" genome regions.

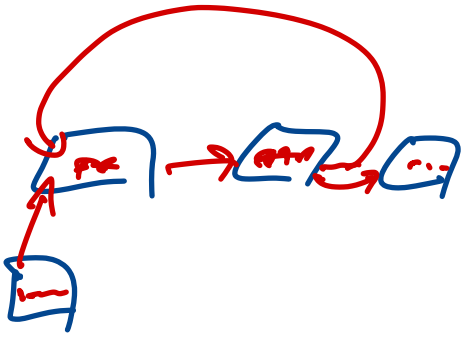
problems: - difficult to generate modeling + computationally
- greedy in memory

ad (2) variation graph:



"vg toolkit"

- directed acyclic graph (DAG)
  → genomes are uniquely stored in the graph



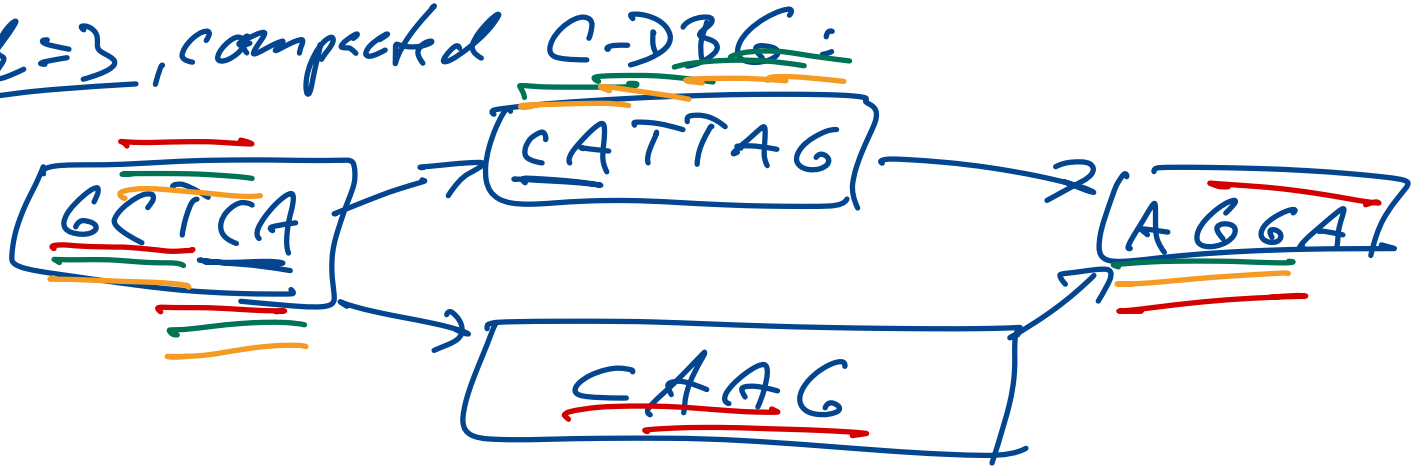
- with cycles: unique if also the genome sequences are stored.

genome sequence

ad (3): colored de Bruijn graph (C-DBG)
compacted de Bruijn graph

- vertices: k-mers (or longer) that occur in the genomes
- edges: overlaps of length $k-1$ (that occur in any of the input genomes)
- color annotation
  for each k-mer indicating the genomes in which it occurs.

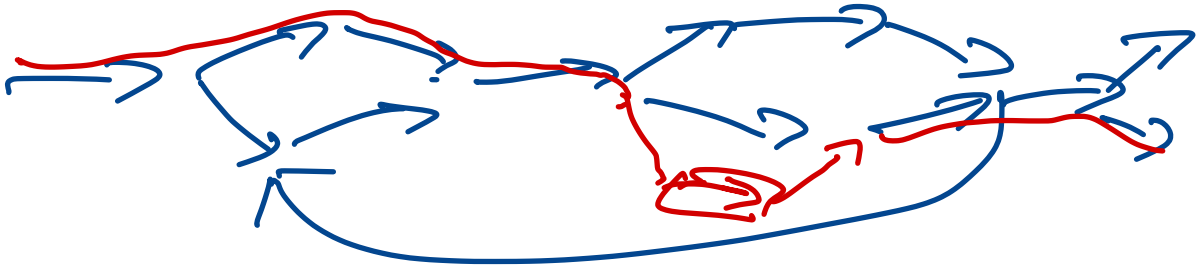
k=3, connected C-DBG:



advantage: uses little memory for graph
potential disadvantage: how to represent the colors

Applications for genome-based pangenomics

- variation detection:
  → via sequence-to-graph mapping/alignment



- core genome detection:
  → try to identify regions in the graph that are covered by a high percentage of colors.
- phylogenomics: (SANS-serif)
  → diverging arcs define splits between colors.

→ split (•,•)|(•)

→ splits → trees

"SplitsTree"

- haplotype inference → next week