LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

GOBI PROJECT

Analyzing the Variance in a Pangenome

Aaron Roth

PLATZHALTER Für MOTIVATION

- WICHTIG!! FOLIE EINFÜGEN

Motivation

- Until now: Linear Reference Genome
- What could be the Problem with that?

Motivation

- It is missing the variation
- 3 million SNVs and 20 thousand structural variations between two humans

Source: Prof. Dr. Wolfgang Enard

Motivation

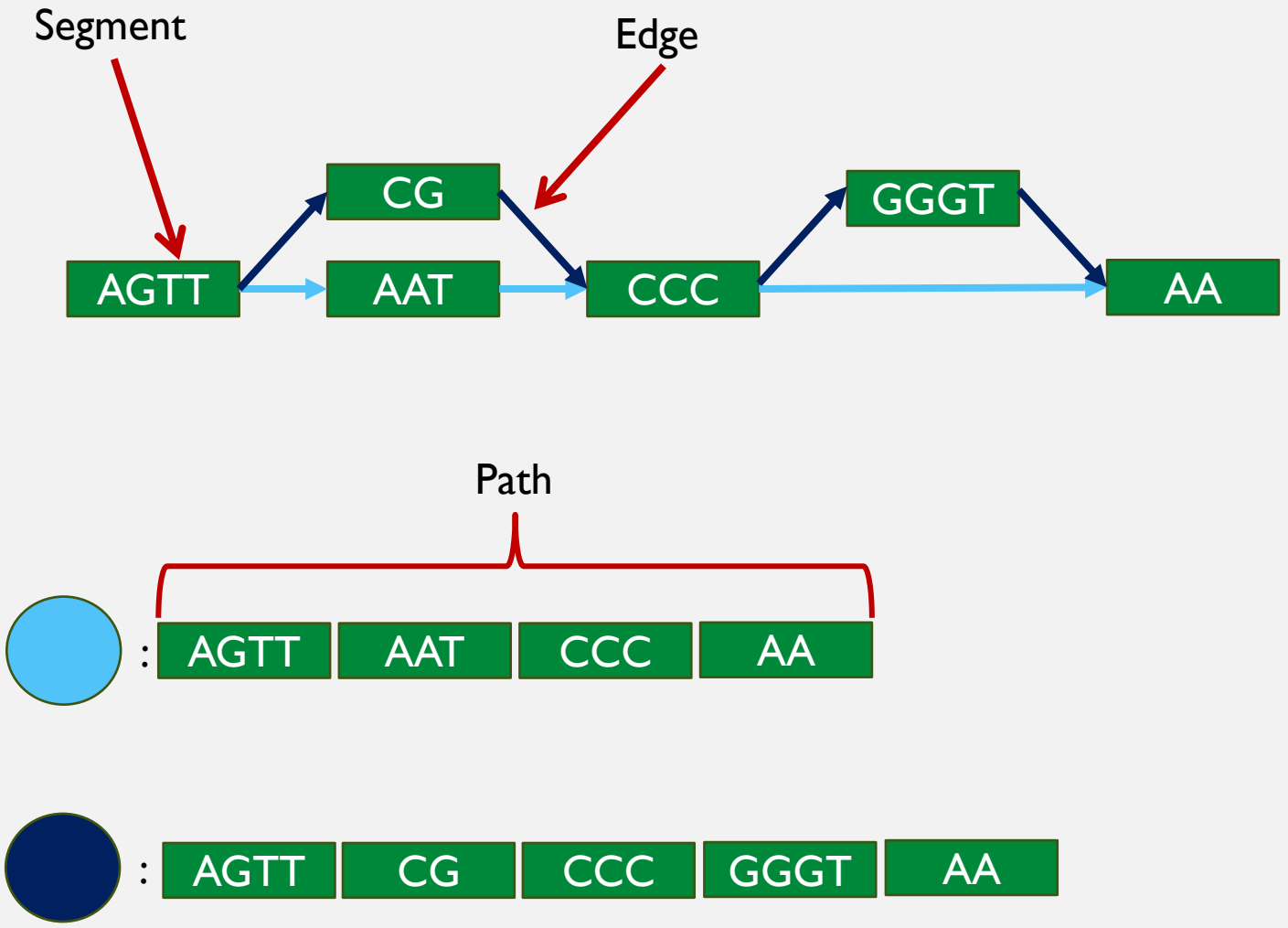
- We need comprehensive representation of human genetic diversity
- Enables better understanding of genetic variation across populations

Pangenome

- Consist of several fully sequenced haplotypes aligned in a graph
- Has many use cases and can often serve as basis of data analysis instead of linear reference genome

Pangenome Visualisation

- Nodes: DNA segments
- Edges: Connect segments
- Paths: Represent haplotypes

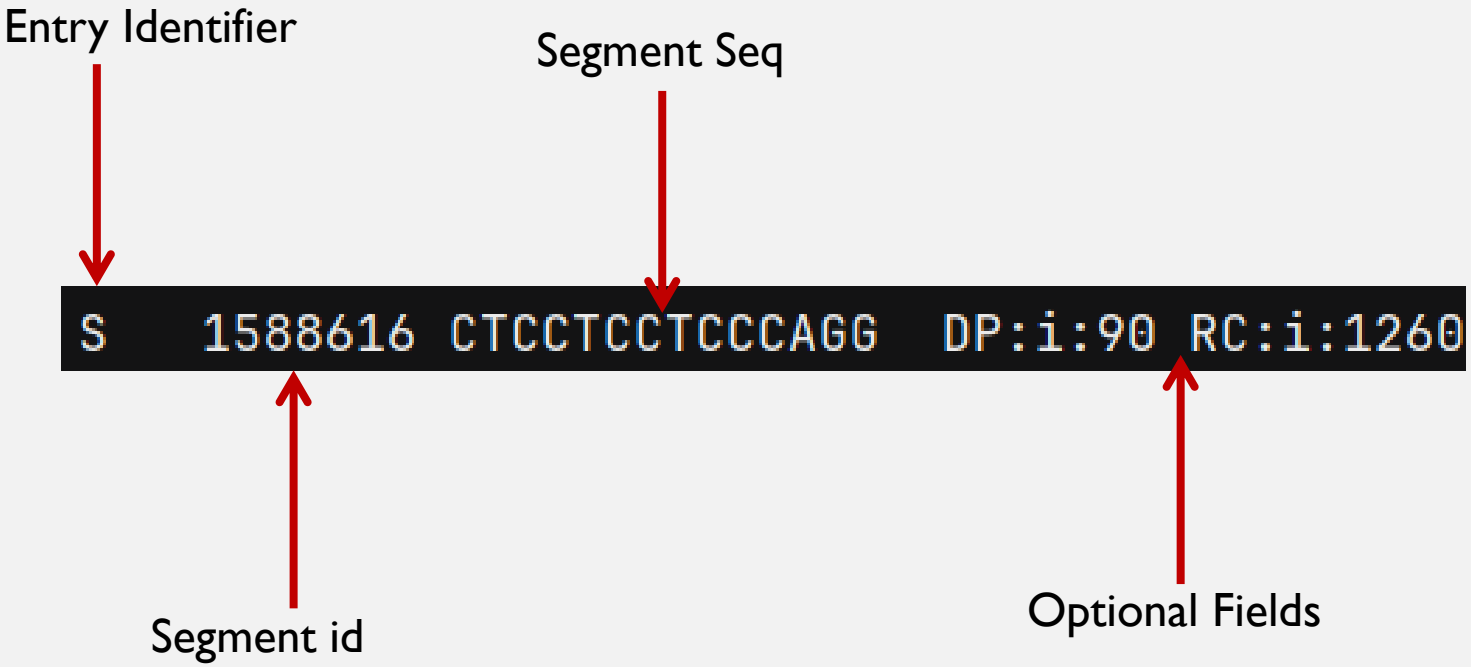


GFA File

- File format to represent Pangenome
- Tab separated, entries for segments, links and paths
- Focusing on version that is used for the original pangenome paper [A draft human pangenome reference]

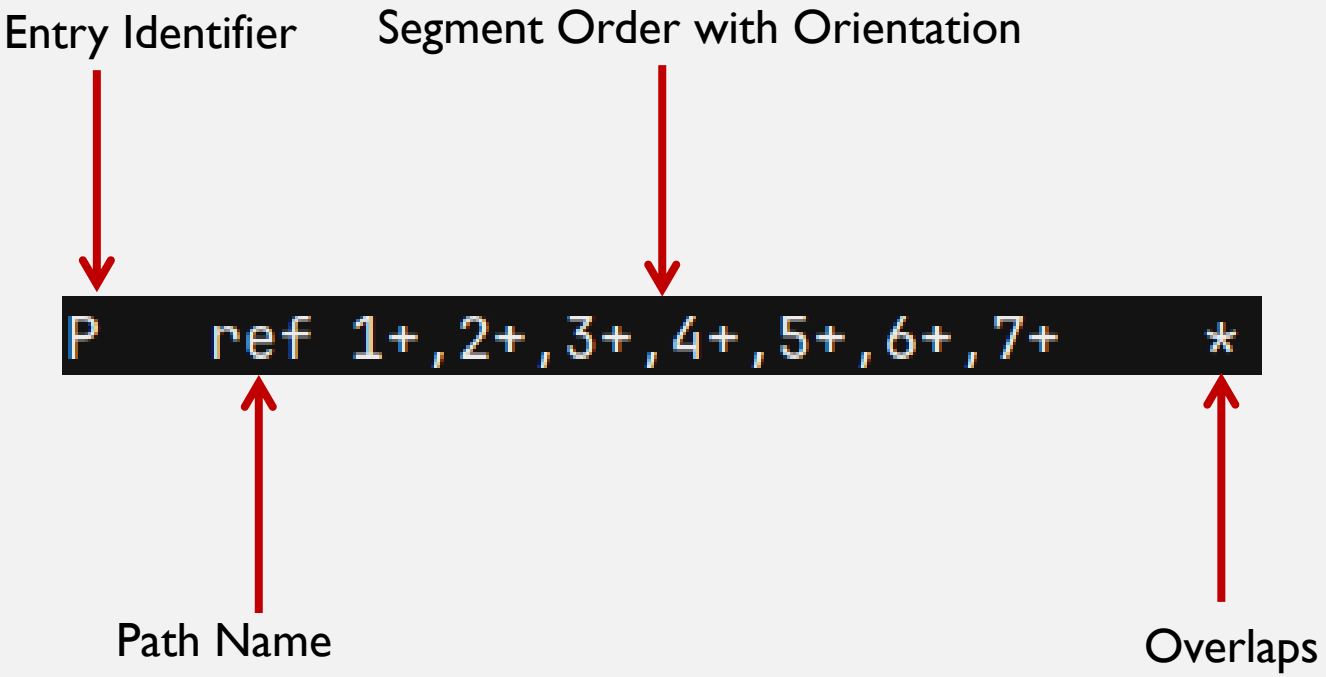
GFA Segment Entries

- Segment represent nodes in the graph
- Always associated with sequence



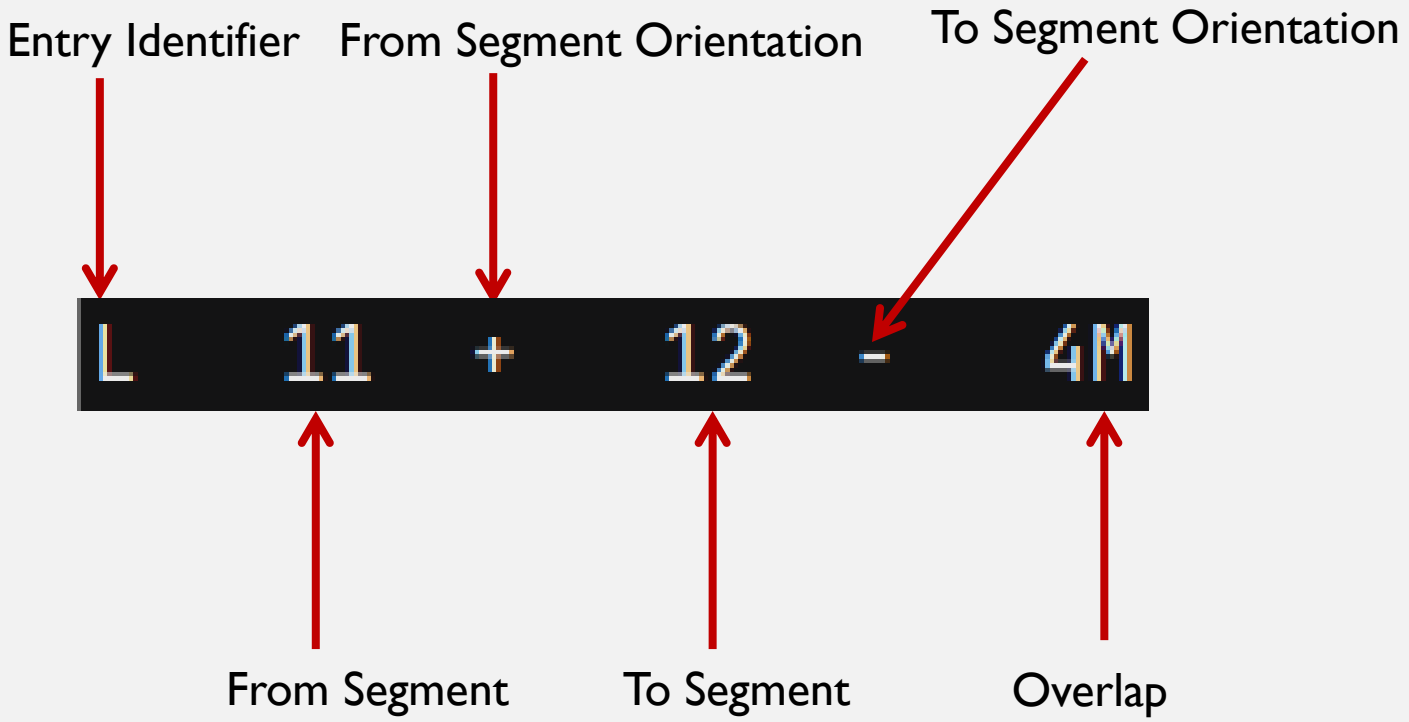
GFA Segment Entries

- Path represent haplotype in the graph
- Ordered sequence of segment ids
- Overlaps are usually left empty



GFA Link Entries

- Links two segments
- Overlaps between segments as cigar string
- Stores orientation of both segments



Note on Edges

- Overlaps don't make sense
  - Also 0 for every file you get
- Orientation is also stored in path

➡ **Obsolete!!!**

Variance

- Pangenome allows to analyse variance across several haploytpes

- Reference genome is included in pangenome
  - Can look at annotated regions

The Project

- Your task will be to analyze variance in pangenome
- Genes / annotated region should be the focus

Your Tools

- Input:
  - Original pangenomes
  - Mini-Gfas (recommended)
- Gfa reader
  - By me ☺
  - Optional (I will know though)
  - https://github.com/BetelgeuseBugFixer/GfaReader


Star my GitHub

Challenges

- Transfer annotation
- Define a measure of similarity for paths
- Sort and visualize results in meaningful way