GENOME GRAPHS

Presented By - Payal Banerjee

Genome graphs and the evolution of genome inference - Genome Res. 2017 May; 27(5): 665–676. doi: 10.1101/gr.214155.116

Human Reference Genome

- The first draft genome was published in 2001
- The ‘complete’ genome, meaning 99% of the euchromatic sequence with multiple gaps in the assembly, was announced in 2003
- Currently, the human genome is at version 38 (GRCh38)
  - *fewer than 1000 reported gaps (Genome Research Consortium [GRC])*

Pitfalls:

- Of the 20 donors the reference was meant to sample from, 70% of the sequence was obtained from a single sample, ‘RPC-11’, from an individual who had a high risk for diabetes
- The remaining 30% is split 23% from 10 samples and 7% from over 50 sources
- After the sequencing of the first personal genomes in 2007, the emerging differences between genomes suggested that the reference could not easily serve as a universal or ‘gold-standard’ genome
- The HapMap project and the subsequent 1000 Genomes Project were a partial consequence of the need to sample broader population variability

In short, there is lack of diversity in reference genome

Thus there is Reference Bias

Variations that are not present in the current reference will not be detected anyways if just resequencing

Potential Solutions

- **Gold-standard reference cohorts** each specific for a particular population

- Development of **pan-genomes**, comprising a collection of multiple genomes from the same species
  - *Pangenomes – these genomes could vary by insertions, deletions, structural rearrangements , large and small mutations*

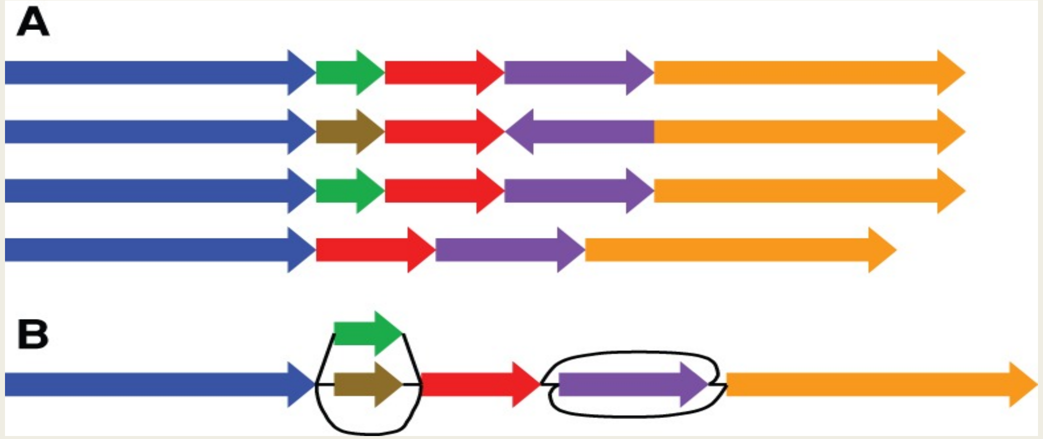
- De novo assembly to make inferences on individual samples

How to encapsulate all that information ?

Genome Graphs

Existing Co-ordinate systems

- Existing primary reference genome is a **poor coordinate space**
- **Doesn’t capture hundreds of alternative locus scaffolds** overlapping each genomic location, across the entire primary reference genome
- Much **large structural variation is not adequately described** by the coordinates provided by the primary reference.



(*A*) A reference cohort, in which there is no attempt to identify homologies between the genome sequences. (*B*) A genome graph, in which homologies are collapsed and included as alternate paths in the graph.

Existing Co-ordinate systems

- Existing reference coordinate systems:
  - ***Variant databases*** *use the reference coordinate systems*
  - *Gene and transcript* ***annotations***
  - ***Genome browsers*** *use linear tracks of genomic data, and graph visualizations*

Change can be tricky !

Sequence graphs

- Sequences represented as walks in graph
- **De Bruijn graph**s – **directed graphs**
    - Nodes are kmers
    - Edges are bases between “to” and “from” node
    - Alternative paths stand in for both structural and single variants

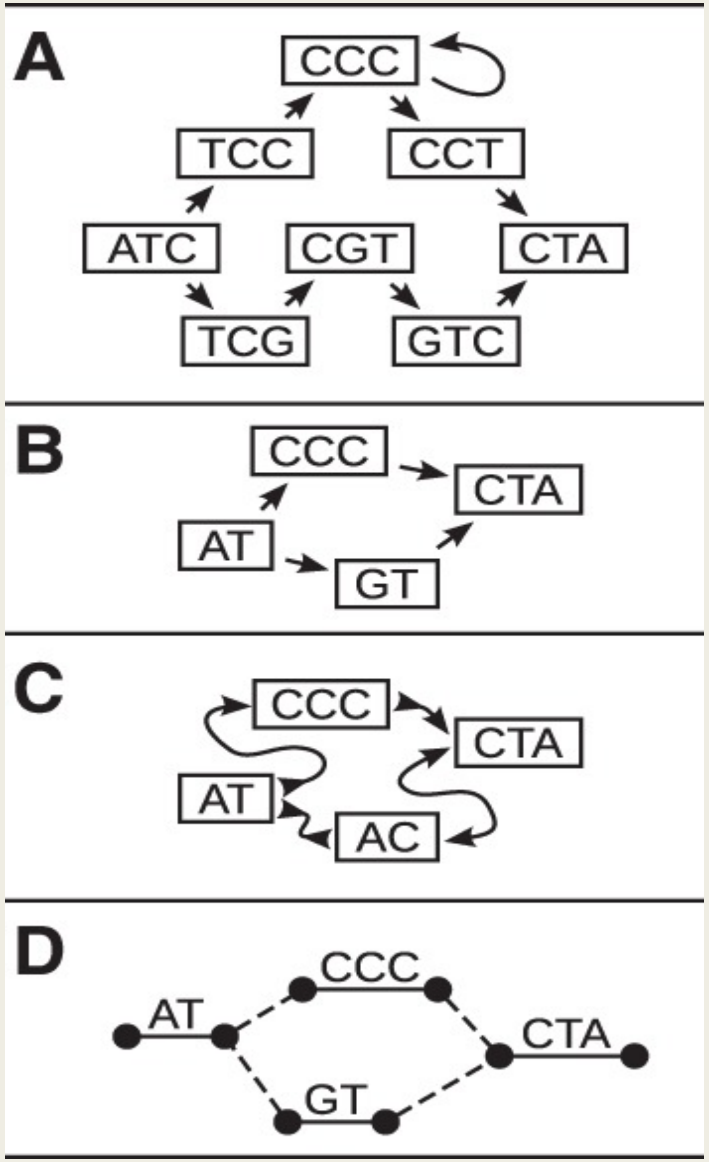
- But just directed graphs are not enough, as they *donot express strandness* – *forward or reverse*

- **Hence bi-directed graphs or Sequence Graphs !**

- But again all inversions, reverse tandem duplications, and arbitrarily complex rearrangements cannot be represented in just bi-directed graph

- **Thus Bi-edged graphs !!**
    - *Edge labeled version of bidirected graphs*

Sequence Graphs



Four types of genome graphs, all constructed from the pair of sequences ATCCCCTA and ATGTCTA. (*A*) De Bruijn graph. (*B*) Directed acyclic graph. (*C*) Bidirected graph (a.k.a., sequence graph). (*D*) Biedged graph (a.k.a., biedged sequence graph).

Things to consider:

The moment we move to sequence graphs **defining a locus** becomes a problem

- ***Graphs have multiple paths, which have complex relationships***

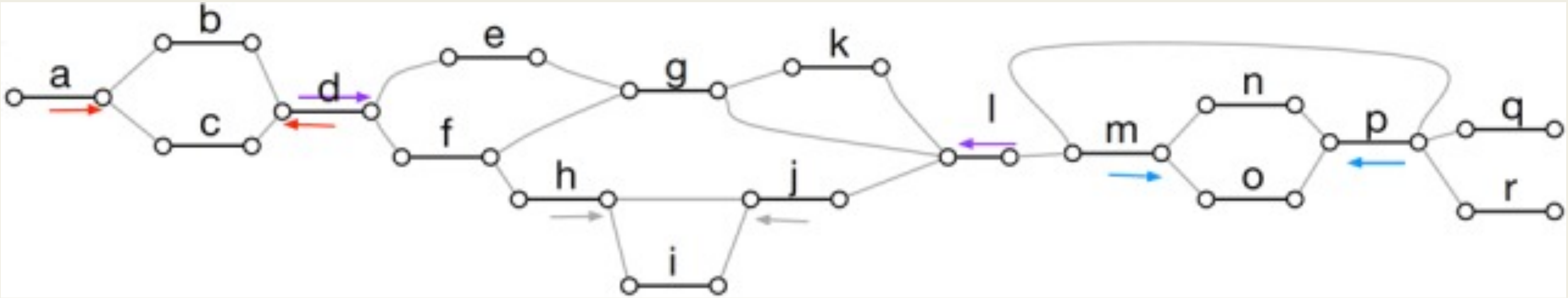
- ❖ *Allelism in graphs*
- ❖ *Repeatome*
- ❖ *Hierarchy*
- ❖ *Pangenome ordering*

Allelism in graphs

Sites described as motifs called **superbubble or ultrabubble**

- *Directed acyclic subgraphs - connects through 1 source node and 1 sink node*
  - *New variants - add Motifs*
  - *Nested and overlapping subgraphs in structural variants*

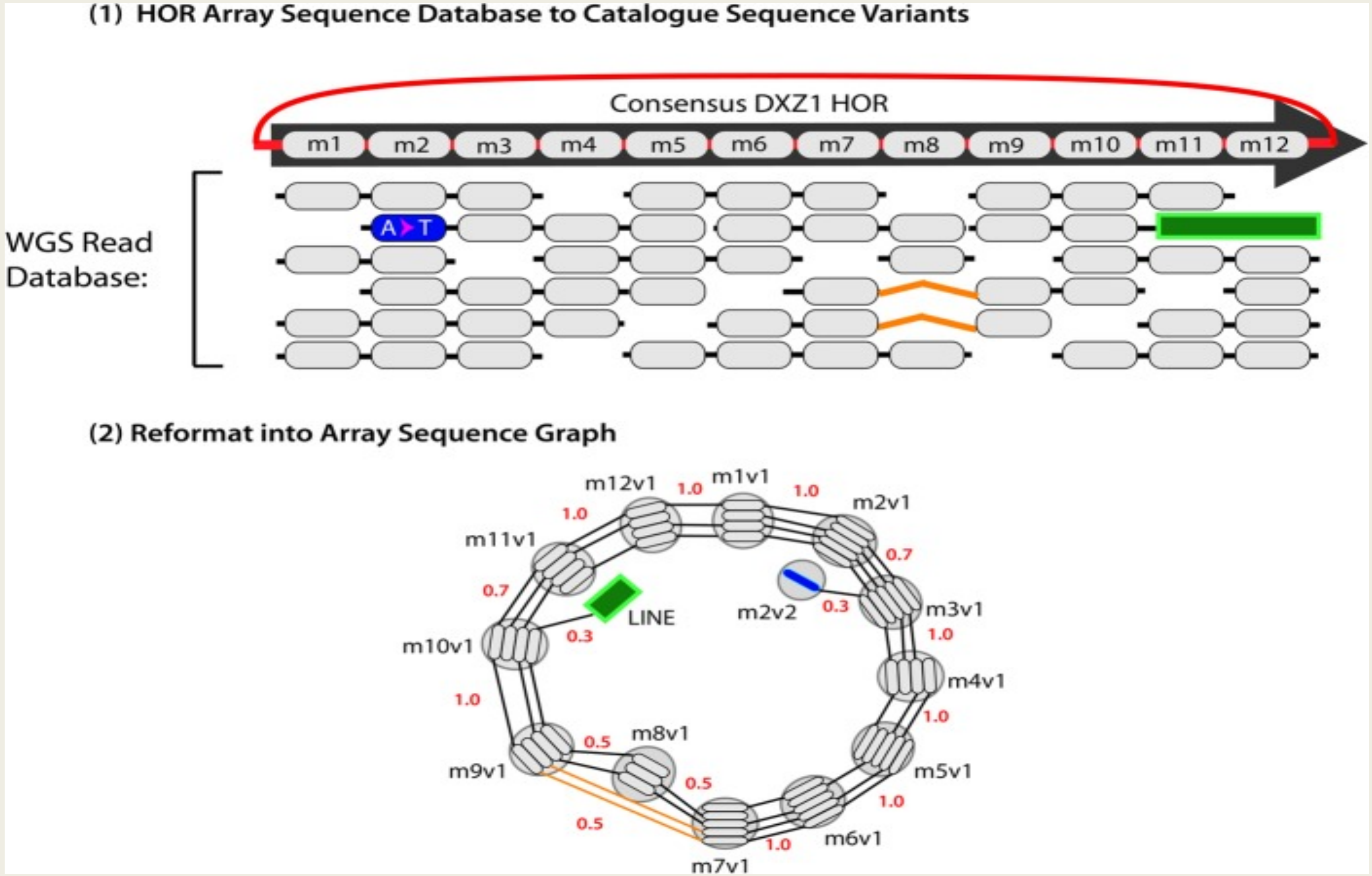
- However - Not all variation nicely partitioned



Ultrabubble sites in a biedged sequence graph. Each arrow shows the terminal node of a site. The color of the arrows indicates the node pairing. Note that the ultrabubble denoted by the gray pair of arrows is nested within the ultrabubble denoted by the purple arrows.

Repeatome as Array Sequence Graphs

- Highly repetitive satellite arrays
  - *Centromeres*
  - *Ribosome*
- Subgraphs of instances of type of repeat
- Similar instances of repeat were combined within graph node

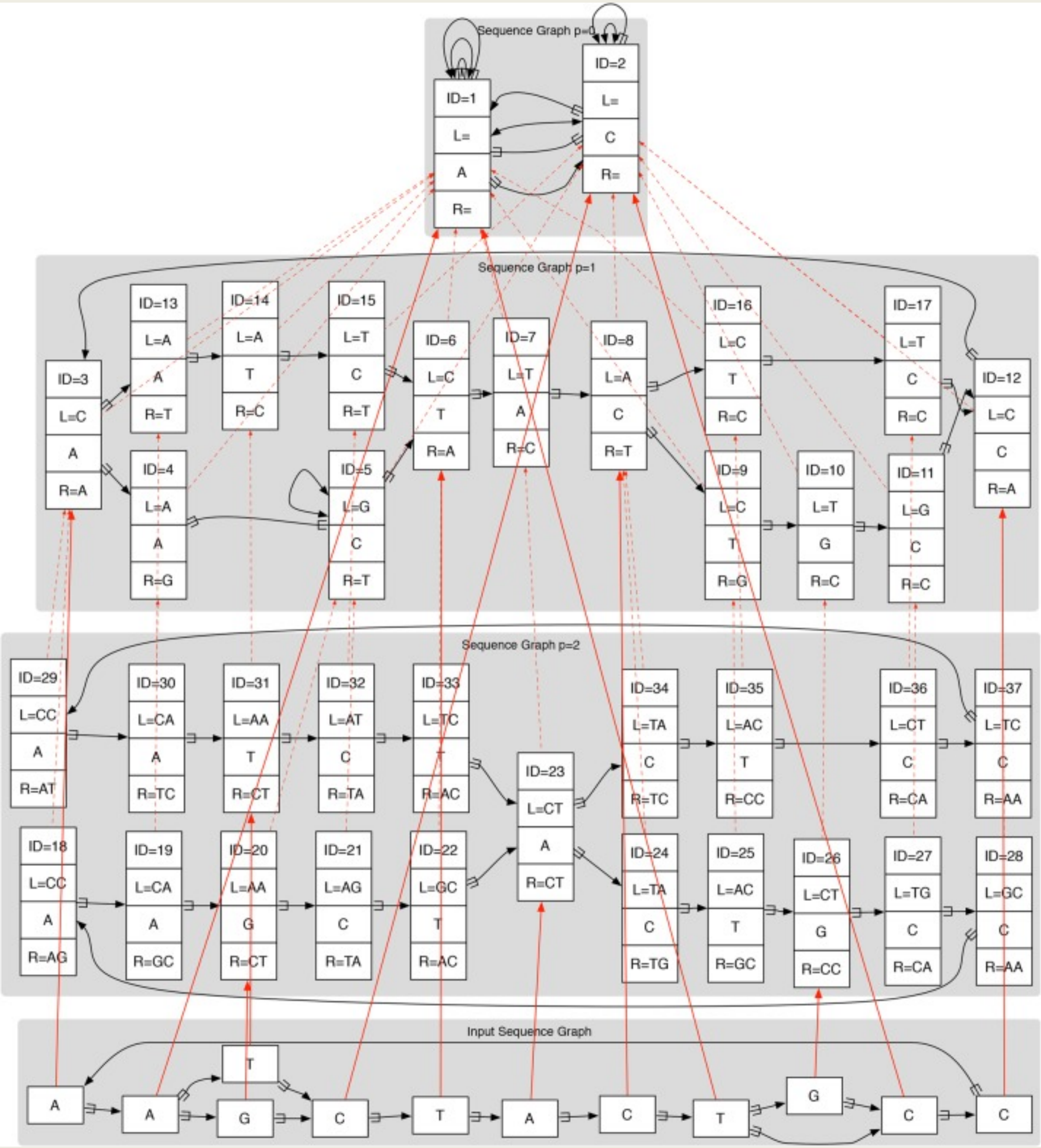


A schematic example of an "Array Sequence Graph" of the type used to construct a linearization of the DXZ1 repeat array in the X Chromosome centromere. A collection of reads (*top*) shown in the context of a consensus higher-order repeat are converted into a graph representation (*bottom*). A cycle around the graph represents a higher-order repeat, and the individual repeat units (oblongs) are represented within each node (circles). Edges between individual repeat units represent phasing information from input reads. Transitions between nodes are annotated with probabilities. (Adapted from Miga et al. 2014)

Hierarchy

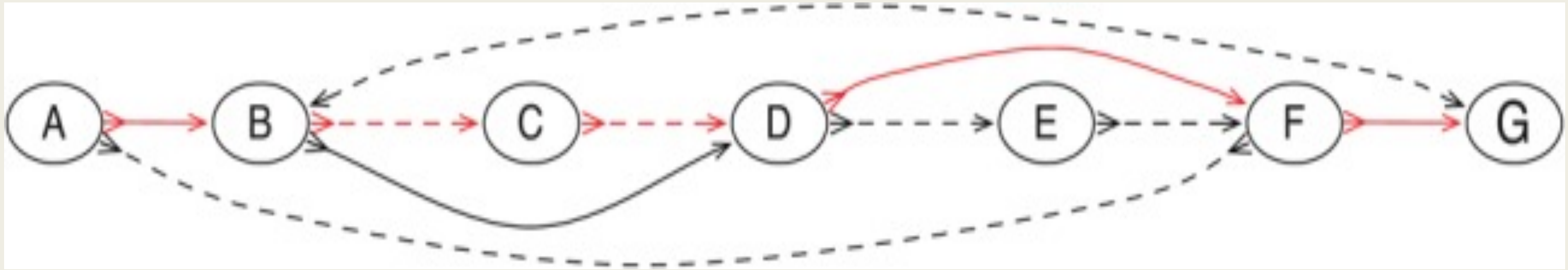
- Hierarchy of graphs
  - *In most collapsed graphs*
    - Repetitive elements are **completely collapsed**
    - Input sequences are **least collapsed**
    - Monoallelic representation of genome - **Intermediates**

- Mapping a subsequence to an element in more collapsed graphs in hierarchy automatically implies the mapping of the subsequence to all the more collapsed version of the element.
  - *Eg: mapping of repeat instance would imply mapping to the canonical copy, classifying it as an instance of the repeat type.*


Sequence Graph p=0
ID=1
L=
A
R=
ID=2
L=
C
R=
Sequence Graph p=1
ID=3
L=C
A
R=A
ID=13
L=A
A
R=T
ID=14
L=A
T
R=C
ID=15
L=T
C
R=T
ID=6
L=C
T
R=A
ID=7
L=T
A
R=C
ID=8
L=A
C
R=T
ID=16
L=C
T
R=C
ID=17
L=T
C
R=C
ID=12
L=C
C
R=A
ID=4
L=A
A
R=G
ID=5
L=G
C
R=T
ID=9
L=C
T
R=G
ID=10
L=T
G
R=C
ID=11
L=G
C
R=C
Sequence Graph p=2
ID=29
L=CC
A
R=AT
ID=30
L=CA
A
R=TC
ID=31
L=AA
T
R=CT
ID=32
L=AT
C
R=TA
ID=33
L=TC
T
R=AC
ID=23
L=CT
A
R=CT
ID=34
L=TA
C
R=TC
ID=35
L=AC
T
R=CC
ID=36
L=CT
C
R=CA
ID=37
L=TC
C
R=AA
ID=18
L=CC
A
R=AG
ID=19
L=CA
A
R=GC
ID=20
L=AA
G
R=CT
ID=21
L=AG
C
R=TA
ID=22
L=GC
T
R=AC
ID=24
L=TA
C
R=TG
ID=25
L=AC
T
R=GC
ID=26
L=CT
G
R=CC
ID=27
L=TG
C
R=CA
ID=28
L=GC
C
R=AA
Input Sequence Graph
A
A
T
G
C
T
A
C
T
G
C
C

Pangenome ordering

- Linear ordering of node

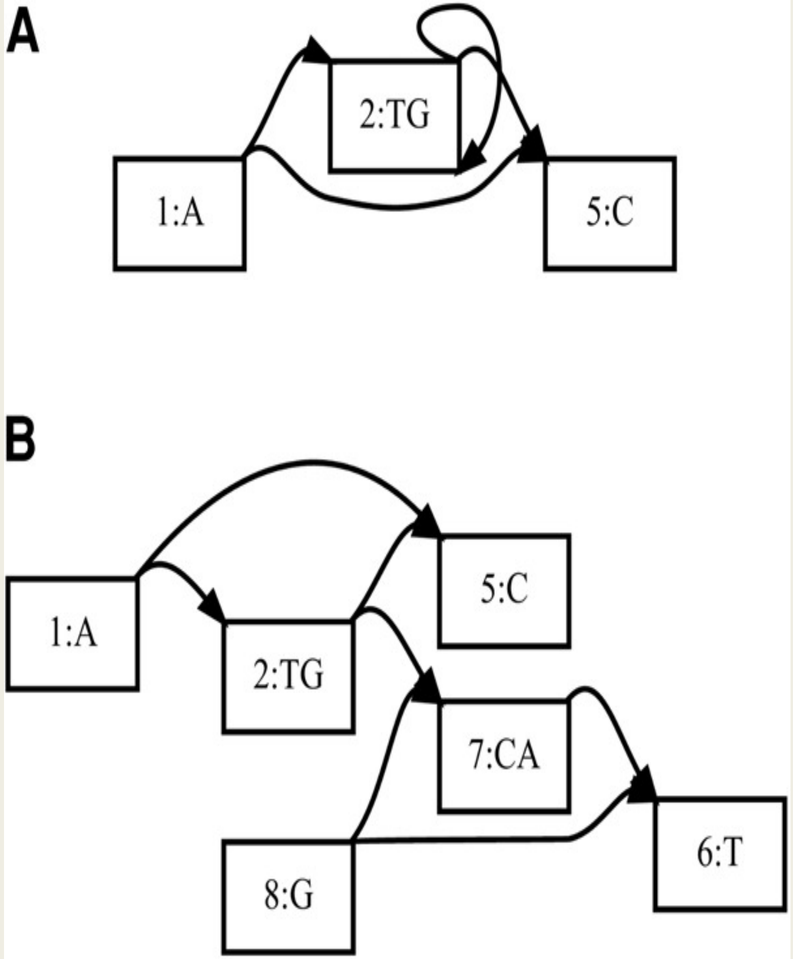


A pangenome ordering on a graph constructed from two genomes. The red edges indicate the path of the pangenome through the graph. The solid and dotted edges indicate the adjacencies between nodes in the two source genomes. (Adapted from Nguyen et al. 2015)

How to decipher Genome Graphs

- ”unrolls” and “unfolds”

bidirected cyclic graphs > directed acyclic graphs



Extending mapping to genome graphs

- Indexing
  - *Self-index*
    - Graph Positional Burrows Wheeler Transform - Graph-PBWT(gPBWT) [vg]
  - *K-mer lookup*
- Distance Measurement
  - *To account for the distance between paired end reads*

  Its difficult to calculate distance between mappings and the relative orientations in graphs
- Context mapping
  - *Flanking sequences*

Graphs should include

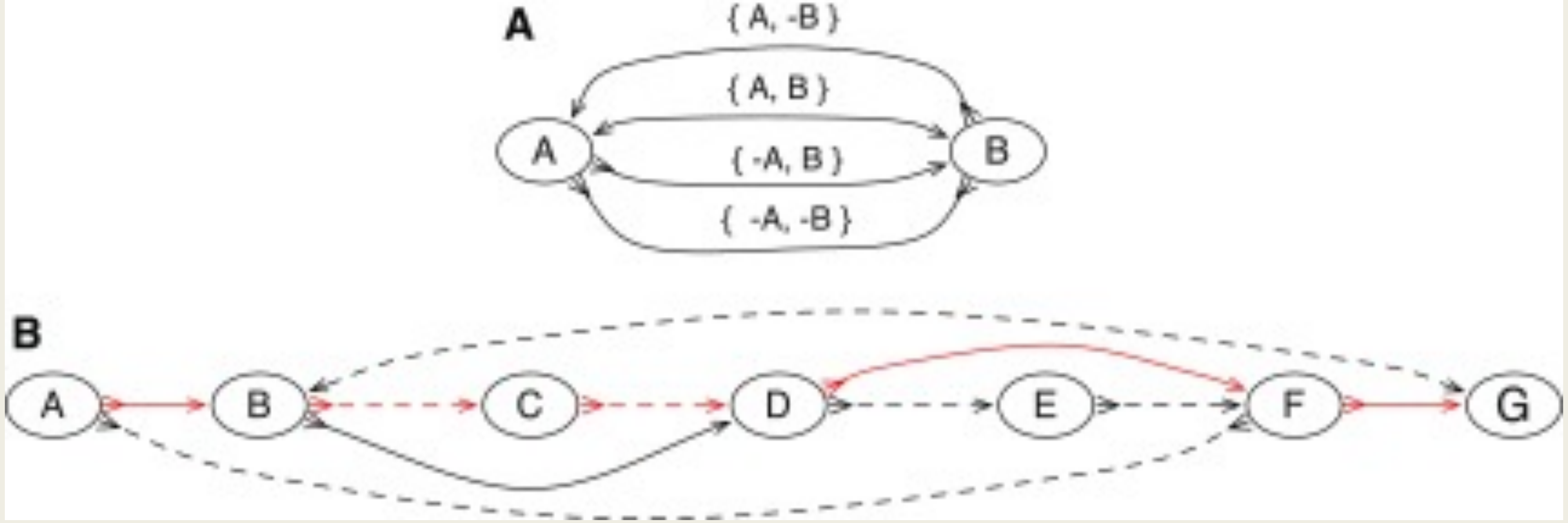
- *Population cohorts*
- *Representation of :*
  - Alternate allele
  - Repeats
  - Large structural variations
- *Potential to accommodate new novel variations*

Challenges to graph genomes

- *Making the cohort information compact enough to store in memory is difficult*
    - Compact data is hard to perform computation on
- *New tools to read and interpret graph based genomes*
- *Indexing in graphs is complicated*
- *New co-ordinate system*
- *New data formats*
- *Updating all databases, annotations and data formats*

Basically reinventing the wheel !

Discussion



An illustration of a pan-genome reference on a sequence graph. **(A)** A bidirected graph representing the four ways two blocks can be connected. The arrowheads on the edges indicate their endpoints: the sides of the vertices. **(B)** An example pan-genome reference on a sequence graph. There are two sequences, indicated by the color of the edges. The red sequence, represented by the thread *A*, *B*, *C*, *D*, *F*, *G*, and the black sequence, represented by the thread *A*, −*F*, −*E*, −*D*, −*B*, *G*. The red thread visits the edges {−*A*, *B*}, {−*B*, *C*}, {−*C*, *D*}, {−*D*, *F*}, and {−*F*, *G*}, and the black thread visits the edges {−*A*, −*F*}, {*F*, −*E*}, {*E*, −*D*}, {*D*, −*B*}, and {*B*, *G*}. Neither thread includes all the blocks. A pan-genome reference, indicated by the dotted edges, is *A*, −*F*, −*E*, −*D*, −*C*, −*B*, *G*. The dotted edges and the edges {−*B*, *D*} and {−*D*, *F*} are the edges consistent with the given pan-genome reference.